VICTORIA UNIVERSITY OF WELLINGTON
TE HERENGA WAKA

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

Navigating the analysis of mass spectrometry data for marine biomass and fish demands a technologically adept approach to derive accurate and actionable insights. This research will introduce a novel AI methodology to interpret a substantial repository of mass spectrometry datasets, utilizing pre-training strategies like Next Spectra Prediction and Masked Spectra Modeling, targeting enhanced interpretability and correlation of spectral patterns with chemical attributes. Three core research objectives are explored: 1) precise fish identification via multi-class classification; 2) Quantitative contaminant Analysis employing multi-label classification and multi-output regression; and 3) traceability through pair-wise comparison and instance recognition. By validating against traditional baselines and various downstream tasks, this work aims to enhance chemical analytical processes and offer fresh insights into the chemical and traceability aspects of marine biology and fisheries through advanced AI applications.

Chapter 1

Introduction

The section provides a **problem statement**, **motivations**, **limitations**, **research goals**, **summary**, and **organization of the proposal**. Each of those sections will be explored in greater detail in the remainder of this chapter.

Waste utilization in the global fishing industry needs improvement. As of 2020, approximately 100 million tonnes of wild fish are caught each year, but only about 40% of these fish are processed into edible parts [1]. The remaining portions are often processed into fish oil and fish meal, or discarded. In addition, many fisheries are in decline, despite global fishing not significantly increasing in the past 30 years, making waste utilization an important focus worldwide. The fishing industry must maximize the utilization and value of every kilogram of marine biomass to preserve fish stocks and ensure there are plenty of fish in the sea for future generations to reel in.

1.1 Problem Statement

Utilizing marine biomass is key to efficiency and sustainability of fish processing. To do so, tools are needed that can identify valuable products. Fish processing gets messy, once processed, cooked, or rendered into oil, it becomes impossible to identify fish species and body parts visually. More sophisticated tools that perform chemical analysis are required. Mass Spectrometry is one tool used to profile the chemical composition of marine biomass for further analysis. Interpreting the results of this analysis is time-consuming, manually laborious, and requires domain expertise in bio-chemistry. *This work aims to automate the analysis of Mass Spectrometry on marine biomass using machine learning.* The analysis is interested in identifying fish species and body parts, detecting contamination - cross-species and mineral oil, and labelling individual fish for troubleshooting. Increasing the speed of chemical analysis of fish will improve the throughput of fish processing. Tools for contamination detection and individual fish recognition offer new methods for quality assurance.

The New Zealand fishing industry prides itself on sustainability. New Zealand fisheries are well-regulated with strict quotas for over 100 marine species [2]. The NZ fishing industry does not have many 'high volume' fisheries, e.g. Hoki is the largest fishery with approximately 110,010 tonnes of quota each year [3]. On a global scale, this is minuscule, Norway alone has an aquaculture production of salmon of 4,000,000 tonnes a year [4]. Due to low-volume but high-variability in New Zealand's seafood industry, fish processing is a difficult task. The variability comes from boatloads of different species as input to a fish processing plant. Each species needs to be identified, separated, and checked for contaminants, then processed to maximize its value as output. Cyber-Marinecitepfr2020cybermarine 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.

The Cyber-marine Research Programme's goal is to *maximize waste utilization in fish processing, and maximize value for harvested and aquacultured marine biomass*. This research offers tools that can identify marine biomass, no longer recognisable by the human-eye, a homogeneous mince/oil of rendered raw/cooked fish, and determine whether it can be repurposed, to maximize its value. To identify fish species, body part, contamination, and as unique individuals. Real-time methods for rapid determination of fish can maximize the utilization of marine biomass in fish processing, quickly identify and eliminate contamination, and improve the quality, speed and efficiency of the factory overall. It is undertaken in collaboration with Plant & Food Research [5] and Callaghan Innovation [6]. This research serves as a proof-of-concept as part of a larger joint endeavour, the Cyber-marine Research Programme [7], which aims to achieve utilization and value for all harvested wild and aquacultured seafood. The proposed methods offer diagnosis and analysis tools for quality assurance in fish processing.

1.1.1 Quality Assurance

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 [8]. 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. A meta-analysis [9] of 51 studies of the global fishing industry found 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. Similar tools to [10] are needed for quality assurance that can determine the composition and quality of fish products.

1.1.2 Potential for Automation

This work will employ machine learning techniques to detect spoilage indicators, quality control, and contamination (ideally) on fresh marine biomass. Machine learning can be used to develop tools with the potential to be more effective, efficient, and less expensive than the equivalent human labour and domain expertise employed currently to analysis 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). Tools that identify contamination in marine biomass are needed. Techniques that work on fresh (uncooked) marine biomass are needed, as cooking the fish can destroy valuable proteins, collagen and active enzymes. Cooking is energy-intensive and time-consuming, adding time and cost to fish processing, so processing fresh marine biomass is preferred.

Automation of fish processing reduces laborious manual labour, expensive domain expertise, and speeds 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 without mechanistic interpretability, and focuses this work on explainable AI, whose models can be understood 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.

> Any sufficiently advanced technology is indistinguishable from magic [11].
>
> *Arthur C. Clarke*

1.2 Dataset

This research focuses on the marine biomass analysis for two mass spectrometry datasets taken at different speeds therefore providing tabular datasets of different resolutions or dimensionality. Those mass spectrometry techniques are:

1. rapid evaporative ionisation mass spectrometry (REIMS) [12]
2. direct infusion mass spectrometry (DIMS)

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:

Rapid evaporative ionisation mass spectrometry (REIMS) is one of the newest forms of AMS and, as is the case with many analytical innovations was created for medical research purposes. It operates using an electro- surgical knife, bipolar forceps or laser which creates an aerosol (smoke) when cutting into a tissue sample. The aerosol is evacuated from the sample through a transfer line into the ionisation source of a mass spectrometer where a heated collision surface is situated and the ionisation process occurs.Excerpt from (Black 2017) [13]

Direct mass spectrometry (DIMS) involves creating ions at atmospheric pressure from solid samples before ions are sucked into an MS detector for analysis. DIMS is the most common fingerprinting tool in metabolomics is direct infusion mass spectrometry (DIMS), based on the direct introduction of sample extracts containing whole metabolites into the mass spectrometer, which avoids the conventional time-resolved introduction of metabolites into the MS after chromatographic separation, improving analysis rapidity and reproducibility, nontargeted metabolite coverage, and, consequently, high-throughput screening capability Excerpt from (Gonzalez 2014) [14].

REIMS is rapid and low-resolution, DIMS is slower and high-resolution. The performance comparison of machine learning methods on both datasets provides insight as to whether rapid mass spectrometry techniques are sufficient for quality control in fish processing. Should they offer competitive results within an acceptable margin of error, but a fraction of the time, rapid mass spectrometry would prove capable of real-world use in the application of marine biomass analysis in fish processing. Please see section 3.1 for a comprehensive and thorough description of the rapid mass spectrometry dataset.

1.3 Machine Learning

The research will employ the AI techniques of binary/multi-class classification, multi-label/multi-output regression, pair-wise comparison and instance recognition. These techniques are applied to low sample volume dataset of chemical analysis of fish. Binary classification is used for identifying fish species, and contamination detection. Binary classification is used because there are two species of fish in the dataset for the classification task. Binary classification is used for contamination detection because a sample is either contaminated or not, a binary model with two class outputs. In both cases, the model predicts positive or negative for belonging to one class. Multi-class classification is used to detect fish body parts, as there are more than two classes for this task. Multi-label classification is used for contamination analysis, where a sample may belong to multiple classes. Multi-label classification is used because an individual sample may belong to one or more classes, and the model should predict the set of classes a sample includes. If a sample has been adulterated with cross-species, the model predicts all the species are present. If a sample is contaminated with mineral oil, the model predicts the species of the fish, and the presence of mineral oil, as both class labels are important for informing human decision-making in quality assurance. Pair-wise comparison is used for traceability, in identifying if two samples came from the same fish. Instance recognition is used to sample attribution, to provide a unique marker for each individual fish, and to detect if a new sample belongs to an existing individual. For pair-wise comparison and instance recognition, it becomes a few/one-shot learning task.

The remainder of this introduction introduces the three research objectives, 1) identification, 2) contamination, and 3) traceability. The motivations and limitations for each objective are given.

1.4 Identification

Identification provides relevant information to profile a sample of marine biomass. These profiles include, but are not limited to, the species of the fish, and the body part from which the sample was taken. In fish processing, a fish once rendered into a minced product, paste or oil, is completely unrecognizable from when it freely swam the oceans. Therefore, chemistry techniques can be used to retrieve this lost information and identify the contents of rendered marine biomass. Useful characteristics, such as species, and body parts; help decide how best to use that marine biomass. For example, due to variations in chemical composition between species and parts, some contain larger quantities of fatty oils, they can be repurposed into Omega-3 supplements. The Cyber-Marine flex-factory [7] aims to maximize waste utilization of marine biomass. Therefore, identifying characteristics of marine biomass waste, such as its species and body part, is useful. This knowledge informs decisions on how best to reduce, reuse and recycle that waste, to maximize the value of that marine biomass.

1.4.1 Motivations

Biomass analysis - Existing works into identification of biomass, let alone marine biomass, using rapid spectrometry are limited [13, 15]. Due to rapid evaporative ionisation mass spectrometry (REIMS) [16] being a recent technological development in chemistry, and the diffusion of innovation [17], access to the REIMS mass spectrometer, and subsequent research, and real-world applications of said technology, is sparse. The tools are cost-prohibitive for widespread adoption and use in industry. However, as part of the greater Cyber-Marine research project [7] serves as a proof-of-concept, for the adoption of REIMS for rapid analysis of marine biomass in the factory of the future - the flex-factory. This research aims will show the viability of rapid mass spectrometry in real-world applications of fish processing. Rapid spectrometry has been shown effective in detecting adulteration in biomass, [13] found beef mince that was contaminated with horse meat. Adulteration is the (often criminal) process of debasing the quality of food products, by intentionally mixing them with products of lower value, to maximize profits, and dishonestly selling them labelled as ONLY the higher value product [16]. The study, [13], showed that REIMS can detect adulteration of beef samples with cross-species contamination at levels as low as 1%, for certain horse-meat offal. Rapid spectrometry has demonstrated a use-case in marine biomass when identifying species of marine biomass for the real-world application of fish fraud detection. Previous works demonstrate that REIMS can be used to combat fraud and adulteration in food processing. This research aims to apply this method of analysis for determining the bulk composition and quality of marine biomass.

Variable marine biomass - Firstly, to apply rapid mass spectrometry methods to fish processing in New Zealand, the proposal tackles the unique market of New Zealand's seafood industry. Unlike other countries, for example, Canada or the United States, New Zealand has a high variability in marine biomass. In layman's terms, when a catch comes in from a fishing vessel, there is a diverse range of species, in that catch. The catches coming from trawling vessels in Canada or the United States consist mostly of one species - a homogeneous composition of marine biomass. However, the catches coming in from New Zealand vessels, consist of a diverse range of species - a heterogeneous composition of marine biomass. This translates to a multi-class problem with many classes in machine learning. This work focuses on a binary classification problem with a dataset containing two fish species. But leaves open the possibility of multi-class classification, as more classes of species could be introduced to the problem when new training classes of training instances become available in future datasets. The real-world problem of heterogenous fish species classification in New Zealand is a multi-class classification problem. It is more complex than the research problem and example dataset this work solves. Future work would look to extend the binary classification algorithms developed here, to handle multi-class classification of fish species, for real-world application in NZ fish processing.

Data scarcity - Secondly, a factor unique to New Zealand's seafood industry, and due to a much smaller fishing fleet and population, is a low sample size. Large trawling vessels in international waters, or the United States or Canada, have a large volume of homogeneous marine biomass, to collect and analyze with chemistry methods. Due to New Zealand's smaller size, and isolated geographical location, there is a much smaller volume of fish to create datasets from for analysis via chemistry techniques. As demonstrated in previous works [18], fish analysis for New Zealand marine biomass, is performed on high-dimensional data with data scarcity. Mutli-class problems with data scarcity shelf the stock-standard toolkit of deep learning methods. With data scarcity, DL methods risk overfitting by memorizing the training data, and not generalizing well on unseen data. The data scarcity requires an algorithm that is sample efficient, DL methods are often not sample efficient at all, as they (often) require thousands of samples to achieve reasonable performance at a given task. Furthermore, they don't produce interpretable models, that can be understood easily, by domain experts in chemistry and fish processing, nor can their results be verified and troubleshot in their real-world application, the Cyber-Marine flex-factory [7], for which they would be deployed.

Seasonal variation - Thirdly, a problem not unique to New Zealand marine biomass, and widely applicable to other marine biomass - is seasonal variation. The chemical composition of marine biomass changes dramatically, in periodic and reoccurring patterns, related to the behaviour of those fish. Take Hoki for example, a fish found in the Pacific Ocean in Australia and New Zealand, a fish species class for classification in this research. When spawning, the female Hoki extracts (almost) all her lipids to give as nutrients to her eggs [3]. This dramatically changes the chemical composition, specifically the lipid profile.

1.4.2 Limitations

Multi-class multi-label classification - The limitations of existing work in biomass analysis using rapid spectrometry are in their application. Thus far, existing literature has only applied biomass analysis for simple statistical applications, more complex and insightful downstream machine learning tasks have not been considered. [15] performs contamination detection by identifying outliers. However, it is not capable of analysis or quantification of the contamination it detects. [13] performs multi-class classification on a dataset with five species of white fish. It predicts one species, therefore is a single-label prediction multi-class classification task. The research is also only limited to white fish species. The model cannot handle cross-species contamination, where a sample contains biomass from one (or more) species of fish, or other types of marine biomass, such as oily fish, salmonid, shellfish, flatfish, tropical fish, freshwater, ... etc. Existing work [13] is geared towards homogeneous marine biomass typical of incoming trawling vessels in international waters, or countries like Canada, United States, or Norway [4]. Thus far, no work is focused on classification of diverse fish species, i.e. heterogeneous species with both white fish and oily fish, simultaneously.

Data scarcity - In rapid spectrometry, there exist few datasets for New Zealand's unique marine biomass. NZ has highly variable biomass. Therefore existing artificial intelligence (AI) models, are not well suited for the niche NZ seafood industry [13]. This research uses datasets of chemical analysis of marine biomass, for a niche seafood market. There are limited (next to no) existing datasets for this specific task, to draw from to increase the sample sufficiency. That was the motivation for the Cyber-Marine project [7] to produce these rapid spectrometry datasets in the first place. Due to the availability and cost of the REIMS measurement, the domain expertise required to operate it, and the niche of its application to the seafood industry, the dataset contains few samples. Each sample is high-dimensional due to the fine-grain resolution of spectrometry methods. High-dimensional data scarcity data is typical of fish oil analysis using chemistry techniques [18]. There exists one paper, [13], that performs REIMS for marine biomass, however, this was not in New Zealand. The marine biomass of each country is unique, both in its variability, and chemical composition of those species. New Zealand has a niche seafood ecosystem, different to Canada, Norway or the United States. A model trained for salmon processing in Norway will not work out-of-the-box on Hoki in New Zealand. Therefore, New Zealand's seafood industry requires AI models trained for its unique fish species and the low volume of data available for them.

Interpretability crisis - in safety-critical applications, e.g. human-grade seafood products, the biggest limitation is industry adoption. Industry is cautious to adopt new methods and techniques, as they introduce risk and uncertainty. Black-box AI techniques add to that uncertainty, by producing accurate models that work most of the time, but when they fail, there is no explanation or cause to diagnose. The explainability crisis nearly ground research on large language model (LLM)s to a halt. There was an open letter and petition to pause giant AI experiments for 6 months [19], with notable signatories in that letter [19], such as Yoshua Bengio [20], Stuart Russell [21], Elon Musk (SpaceX, Tesla) and Steven Wozniak (Microsoft co-founder). More recently, Geoffrey Hinton, the ”Godfather of AI” [22] left Google AI, with fears the generative artificial intelligence arms race will cause real-world harm. The claims of real-world harm and danger surrounding LLMs derive from the black-box nature of these models. Humans cannot understand their process, from input to output, as the weight matrices of a 32+ billion parameter deep neural network. Recall Arthur C. Clarke, and note that a technology that is not understood don’t is equivalent to magic. If industries are cautious about new technologies that are well understood, the black magic of neural networks [23], is going to be a hard sell. Existing works, [13, 15] perform dimensionality reduction that obfuscates the meaning of its features. The results are accurate, but humans, more specifically the domain experts in biochemistry and fish processing, cannot understand and interpret the model.

Concept drift - AI techniques require robustness to concept drift. A naive machine learning model, that is not robust to seasonal variation, may misclassify this adult female Hoki after spawning, not recognizing the mother as a Hoki, absent of her lipids. The phenomena of seasonal variation draw parallels to concept drift from data mining [24, 25]. Concept drift describes the shift in the distribution of data over time, in (possibly) reoccurring or periodic nature. To apply AI techniques for fish processing in New Zealand, robust models that can handle concept drift such as seasonal variation, are needed, for real-world applications to deliver (any) commercial value. A robust model impervious to concept drift would be seasonal invariant.

> Everything should be made as simple as possible, but not simpler [26].
>
> *Albert Einstein (attributed)*
> *Physicist, Nobel Prize (1921)*

1.5 Quantitative Contaminant Analysis

Contamination detection can detect contaminants such as cross-species - where two species of fish are mixed together in one sample, or mineral Ool - where oil from the fishing vessels engine or factory machinery has spoilt the marine biomass. This method can identify potential hazards and/or quality control issues in fish processing. It can help ensure the Cyber-Marine flex-factory [7] is running smoothly, and verify it can produce food grade products that are safe for human consumption [27]. Contamination detection can be implemented in three stages, *detection* - binary classification, *analysis* - multi-label classification, *quantification* - multi-output regression.

1.5.1 Motivations

Existing works address the problem in part, however, none address it in its totality. The end goal of this research is to create a novel and effective method for contamination detection of marine biomass in the Cyber-Marine flex-factory [7], via rapid spectrometry. REIMS have proven an effective tool for identifying spoilage in livestock and seafood products [15, 15]. In response to scandals like the 2013 horse-meat scandal, [15] showed rapid spectrometry can identify adulteration of beef products with horse-meat offal, in concentrations as small as 1% (for certain offal). This work uses supervised learning techniques of PCA-LDA [28, 29] with thresholding to detect outliers. This tackles the problem of cross-species contamination detection, an area addressed in this proposal. Previous work of the same author, [13], uses the same technique for fish fraud detection, to find fish products that have been mislabelled. [13] is a binary classification task of fish species prediction, that employs the same supervised learning techniques of PCA-LDA, on REIMS data for fish. More advanced artificial intelligence techniques, such as generative adversarial networks (GAN)s in the work [30], have been applied to anomaly detection in factory settings. These also rely on thresholding techniques to identify outliers, which indicate factory equipment that has likely malfunctioned.

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

Analysis - The second task extends this to multi-label classification, each instance may belong to multiple classes. For example, a contaminated sample may contain, 30% Hoki and 70% Mackerel, this sample label would be [Hoki, Mackerel]. In contrast to the binary classification problem, the model would also have to be able to distinguish between [Hoki], [Mackerel], [Hoki, Mackerel]. For binary classification, both [Hoki], [Mackerel] would simply be negative, not contaminated. But for multi-label classification, both the individual species and their combination, are combined to give the label annotation. It is a multi-label classification because an instance can belong to one or more classes. In the extreme, an instance may contain a mixture of two fish species and mineral oil, which would belong to three separate classes. The machine learning model would output a prediction that identifies all of those classes, e.g. [Hoki],[Mackeral],[Mineral Oil]. Furthermore, the outlier thresholding may detect an unknown fish species or foreign contaminant, giving the *reductio ad absurdum* example of a fish instance whose output label would be [Hoki], [Mackeral], [Unknown species], [Mineral Oil], [Unknown contaminant]. The possible mixture of multiple known, and unknown fish species and contaminants, and the combinatorially explosive nature of their class label powerset, makes multi-label classification, a feasible solution for training a classification model.

Quantification - The third task extends the previous two tasks, to associate a percentage of contamination associated with each contaminant. Not only does it perform the previous two tasks, contamination detection, and contamination analysis, it then provides quantification to those contaminants that it has identified. Recall the example from earlier, with 30% Hoki and 70% Mackerel. Quantification would identify two classes present, and their relative contribution to the same, i.e. [Hoki - 30%, Mackerel - 70%]. This is a multi-output multi-label regression. The annotated label is similar to the softmax [31] function on an output layer of the neural network, which normalizes the output layer of that neural network to sum to 1, to fit a probability distribution. Contamination detection has to be real-time, analysis and quantification can be done ”offline”, once a samples has been isolated and marked for follow up examination. It would be more efficient to only perform analysis and quantification on samples identified as contaminated. This speeds up the factory throughput by removing redundancy. The model’s uncertainty as to which class an instance belongs to can be captured by the softmax operator's probability distribution for each class label. However, this uncertainty for each class, can't be conflated with its confidence in the relative contamination percentages for each class. To clarify, the model confusion in classification output, is not the same as the presence of multi-class contamination, be it cross-species or mineral oil. If the model is uncertain about which class an instance belongs to, it does not imply the model is certain that the instance is contaminated. These are separate tasks and need to be treated as such. Otherwise, the model is left doing a lousy job at both tasks simultaneously. Instead, the model should recognize the distinction, and perform both tasks separately with high accuracy.

In summary, relative probabilities for each class do not necessarily map to relative proportions of contamination. The model being confused over which classes a sample belongs to, should not be mistaken for contamination detection. Contamination detection is agnostic of the model's certainty for multi-class predictions. The distinction is important.

1.5.2 Limitations

No qualitative analysis - The existing work in fish fraud detection [13], is limited as it is for multi-class classification only, it does not provide a model for cross-species contamination, where a single sample may contain fish from two (or more) species. Black [15] provides adulteration detection with thresholding techniques to identify outliers. The existing literature, [13, 15, 10]on rapid mass spectrometry data has only applied PCA-LDA to adulteration detection in biomass. GC-MS analysis has used CNNs for high-accuracy black-box predictions of downstream classification tasks, such as flavour profiling [32]. Similar work, [30], has shown the promise of GANs for outlier thresholding in anomaly detection. However, more recent innovations in deep learning should be considered also. However, no qualitative profiling of those outliers is given. Their technique does not say what the adulterant is, and to what concentration it is present. This is due to the nature of the thresholding technique for outlier detection.

Domain Expertise - The GANs used in [30], are advantageous over the PCA-LDA methods proposed in [13, 15], as they require less manual parameter tuning and domain expertise in the application, however, they produce black-box models, which can't be trusted or understood when deployed in fish processing.

Interpretability - Similar to GANs [30], the PCA-LDA [28, 29] used in [13, 15] produce feature embedding that are not interpretable either. principal component analysis (PCA) is a dimensionality reduction technique that projects features from a high-dimensional space, into a lower-dimensional space. By projecting along the top k eigenvectors of the covariance matrix [20] The principal components are linear combinations of the original features, and their interpretation concerning the original features is not straightforward. The PCA dimensionality reduction technique seeks to preserve the variance of the data, but the original semantic meaning of the features is lost.

GANs / deep learning - There are limitations of GANs [33, 34] from [30]. As they belong to the family of deep learning methods [20], they require computing resources, high sample complexity, and intricate hyperparameter tuning.

Specific, to GANs they are susceptible to training instability, mode collapse and outputting the mode of the data distribution. Mode collapse occurs when the generator fails to produce samples with sufficient diversity or variation. Resulting in a network that generates a few similar or identical outputs. It collapses multiple modes or variations of the target distribution into a single output mode - hence the name. Mode collapse is a specific type of limited diversity. But other scenarios, such as imbalanced training data, too simple generator network architecture, or, not enough training, can lead to limited diversity. This is where their generator fails to capture the complexity of the target distribution. Training instability is a common problem for GANs. They don't have a simple objective function. Instead, they aim to approximate the Nash equilibrium between a generator and discriminator network. This is a min-max game between adversarial networks, hence the name, i.e. generative adversarial networks (GAN). The generator attempts to produce samples indistinguishable from the real data. The discriminator attempts to spot the difference between the real and fake samples. The Nash equilibrium occurs when the generator can produce samples indistinguishable from the real data, i.e. the discriminator can no longer tell the difference. This Nash equilibrium is a more difficult objective function than say Kullback-Leibler (KL) divergence [35], or categorical cross-entropy (which can be derived from KL divergence [20]) for classification tasks.

1.6 Traceability

Traceability deals with isolating a contaminated sample, to track the origin, and identify potential causes for that contamination. The previous subsection addressed contamination, once a contaminated sample is found, a factory seeks to find ALL other marine biomass that originated from the same origin, that same sample, to isolate and discard the potentially hazardous samples, that are likely not safe for human consumption [27]. To do so, traceability must identify the unique characters of an individual sample, not just the species or body part, but unique characters that distinguish ONE particular fish, from the others. In computer vision, this is referred to as instance recognition [36]. Existing works related to traceability can be found in computer vision, those include instance segmentation [37], instance identification [38], and contrastive learning [39, 40, 41].

1.6.1 Motivations

Detection - Contrastive learning is an effective technique for few-shot learning pair-wise comparison, e.g. same-fish detection for traceability. Siamese neural networks, proposed in 1993 by LeCun [39], and prominent today in deep learning for object detection and segmentation [41], and ransomware classification [40]. Siamese neural networks are a type of contrastive learning. Contrastive learning is a type of unsupervised learning where the goal is to learn a similarity metric between two inputs, by contrasting them with other inputs. A similar method concept to the thresholding method for detecting outliers was previously mentioned in [13, 15]. However, contrastive learning is useful for few-shot learning for three reasons: (1) it allows efficient use of small amounts of labelled data, (2) it can leverage labelled and unlabeled data to learn robust and discriminative representations of the data, improving the model's ability to generalize to new classes with only a few labelled examples, (3) by learning to contrast similar and dissimilar examples, the model can develop a rich understanding of the underlying structure of the data, which can further improve its ability to generalize to new classes with few labelled examples. The REIMS dataset contains very few training instances. For the instance recognition task, there are fewer. As individual fish may only have a few (or single) labelled examples per instance. Therefore, instance recognition can be considered a few-shot, or in extreme cases a one-shot, learning problem. Traditional machine learning models often require large amounts of labelled data to achieve high performance, whereas few-shot learning specializes in training models to learn quickly from only a few examples. Due to the low sample sufficiency, and few-shot nature of the instance recognition ask, this research aims to amortize the training data, to allow for few-shot or even one-shot inference on unseen data. This amortization can be achieved through self-supervised contrastive learning, such as Siamese networks, or transfer learning, where models share information between related tasks, to improve performance on new related tasks. This research is going to semi-supervised contrastive learning on publicly available mass spectrometry datasets, to improve the few-shot similarity learning performance on novel mass spectrometry datasets with limited training instances. Semi-supervised learning implies a mix of self-supervised similarity-based learning on unlabelled datasets from different tasks, in combination with supervised few-shot learning on the target domain mass spectrometry dataset with annotated class labels. In short, similarity-based learning on a large corpus of mass spectrometry data, to improve few-short similarity-based learning on a specific mass spectrometry individual detection and instance recognition task.

Instance recognition - It is commonly used in various fields such as wildlife monitoring, security surveillance, and biometrics. To avoid confusion, the term instance recognition from computer vision, is not to be confused with instance identification [38], or instance segmentation [37]. However, in the domain of fish processing, and chemical analysis via rapid mass spectrometry, the term sample attribution is fitting for the real-world application. Thus, for a Chemist and AI Researcher, the terms sample attribution and instance recognition can be considered equivalent and used interchangeably.

Existing work for instance recognition task can be found in computer vision [36, 42, 43]. In [42], the authors propose HotSpotter a model to recognize instances based on their unique spots. This is a species invariant model, that differentiates between dissimilar species, e.g. zebras, giraffes, leopards, and lionfish. Fish and mammals are dissimilar but share spots. [42] was trained on a database with 1,000 images with five different species of animals, approximately 200 images per class. While images are far from rapid mass spectrometry data, this research aims to perform a similar task, by providing a species-invariant model that differentiates between dissimilar species of fish, e.g. whitefish and oily fish, based on their unique chemical compositions. The work of [43] proposed a multi-modal instance recognition that employs dense feature extraction on multi-modal features. This model is benchmarked on the Willow dataset from [44], which contains 37 views of 37 different objects (37 x 35 = 1295 images total) to be detected in a variety of tabletop scenes. Similar to [42], the classes are dissimilar objects, that must be uniquely identified from multiple observations of that same object. This work [43] was trained on a dataset of similar size to [42], with 1,295 and 1,000 images, respectively. Datasets with sample sufficiency, i.e. more than 1,000 instances, are naturally suited towards deep learning methods that require large volumes of data. Traceability must perform the same task. To uniquely identify marine biomass from multiple observations. In [43, 42] the observations are images - a computer vision task. However, in this research, the observations are rapid evaporative ionisation mass spectrometry (REIMS) measurements - a chemical analysis task. Although different fields, a mass spectrograph and pixel image are similar, they share local connectivity in their multidimensional representation, where values close to each other are related, and their proximity to each other is information in itself. In [36], they take instance recognition one step further than [43], with single-view instance recognition. They employ a general-to-specific training procedure, that pretrains the neural network on problems, of increasing granularity. The network is pretrained on a large multi-view dataset and then fine-tuned on a smaller single-view dataset. The neural network takes a feature embedding representation learnt from a general task, that can be transferred and applied to a more specific task. The largest multi-view dataset has 100 images of 124 objects (100 x 124 = 124,000 images total). The smallest single-view dataset has 1 view for 300 objects. In contrast to other works [43, 42], whose datasets contain more than 1,000 instances, this is a one-shot classification task with very few training instances. The REIMS dataset has 306 samples of marine biomass. Training on generalized tasks that are related, but where greater volumes of data are available, can improve the few or single-shot classification performance on datasets with data scarcity (few training instances).

1.6.2 Limitations

Few-shot learning - Traceability, or instance recognition, is the most difficult task proposed in this research. Few-shot and similarity-based learning methods perform better on out-of-distribution data when compared to traditional learning methods. This property is helpful for finding outliers, whose contamination or species is unknown, i.e. not in the training set. It is a few-shot (or in the extreme one-shot) learning task on a dataset with data scarcity. Therefore, with limited training instances, this work aims to maximize the knowledge that can be extracted from each instance - the amortization of data. Another way to achieve amortization is through transfer learning. To learn on a more general task, and transfer that knowledge to a more specific task through fine-tuning, as seen in [36] for single-view instance recognition.

Siamese networks - Alternatively to pretraining, Siamese networks [39, 41, 40] is another effective technique for few-shot contrastive learning. Contrastive learning is a useful technique for data scarcity datasets, where sample efficiency is critical. Deep learning techniques, like Siamese networks, are very computationally expensive to train and often require dedicated hardware, such as GPU clusters. Deep neural networks are very sensitive to their parameterization and require extensive hyper-parameter tuning that lacks theory and is comparable to black magic. In fact [23], coined the term ”Grad Student Descent” to describe the brute-force trial and error based-process of tuning parameters for deep neural nets. Those were limitations of deep learning methods in general. However, a limitation unique and specific to this research is the representation. Siamese networks [39, 41, 40] typically require a fixed-length input. However, as shown in [13, 15], there is flexibility in mass spectrometry for variable-length inputs. Chemists can increase the resolution to get more features, i.e. longer-length input. Conversely, they can decrease the resolution to get shorter-length input. A model that cannot handle variable-length inputs is not robust, it could only be trained and tested on datasets with a fixed resolution, that fixed resolution being the original fixed-length input of the measurements it was trained on.

Only computer vision - Existing work on traceability is limited to the related task of instance recognition from computer vision. Works [42, 43, 36] show applications of instance recognition for dissimilar classes, [36] extends this for one-shot instance recognition with data scarcity. Both [42] [43] require large sample complexity - a high volume of training data to work. Many deep learning and traditional machine learning methods require many training instances to achieve high-quality performance. The REIMS dataset in this research is scarce on data, with only 306 training instances. These deep learning and traditional machine learning techniques will not work out of the box on a dataset with data scarcity. In extreme cases, there may likely not be sufficient data for these models to fit data. It is more likely that these models will overfit the training data, and fail to generalize on unseen data. However, it is possible to use pretraining (or transfer learning) to allow for few or one-shot learning on a dataset with data scarcity, as shown in [36]. The limitations of [42], are being a relatively dated paper, 2012 paper [45] that proposed local naive bayes nearest neighbours (LNBNN), an extension of naive bayes nearest neighbours (NBNN) [46], where ”only the classes represented in the local neighbourhood of a descriptor contribute significantly and reliably to their posterior probability estimates”. The authors admit LNBNN, did not beat state-of-the-art methods such as feature pyramid networks [47], which rely on local soft assignment and max pooling operators. Convolutions and max-pooling are utilized in convolutional neural networks (CNN)s [48], a powerful model for computer vision-related tasks. With advancements in hardware and the lifting of the AI winter, are efficient to train at scale using GPUs. Since then, a plethora of convolutional neural networks (CNN) architectures dominate computer-vision tasks, such as LeNet [48, 49, 50, 51], AlexNet [52], VGG-16 [53], GoogLeNet [54], ResNet [55].

1.7 Research goals

This application-driven research aims to implement real-time (online) fish contamination detection and identification. This is a supervised machine learning task trained on scarce rapid evaporative ionisation mass spectrometry (REIMS) [12] fish oil dataset. Specifically, this proposal outlines the need for algorithms to perform the following tasks summarized in figure 1.1.

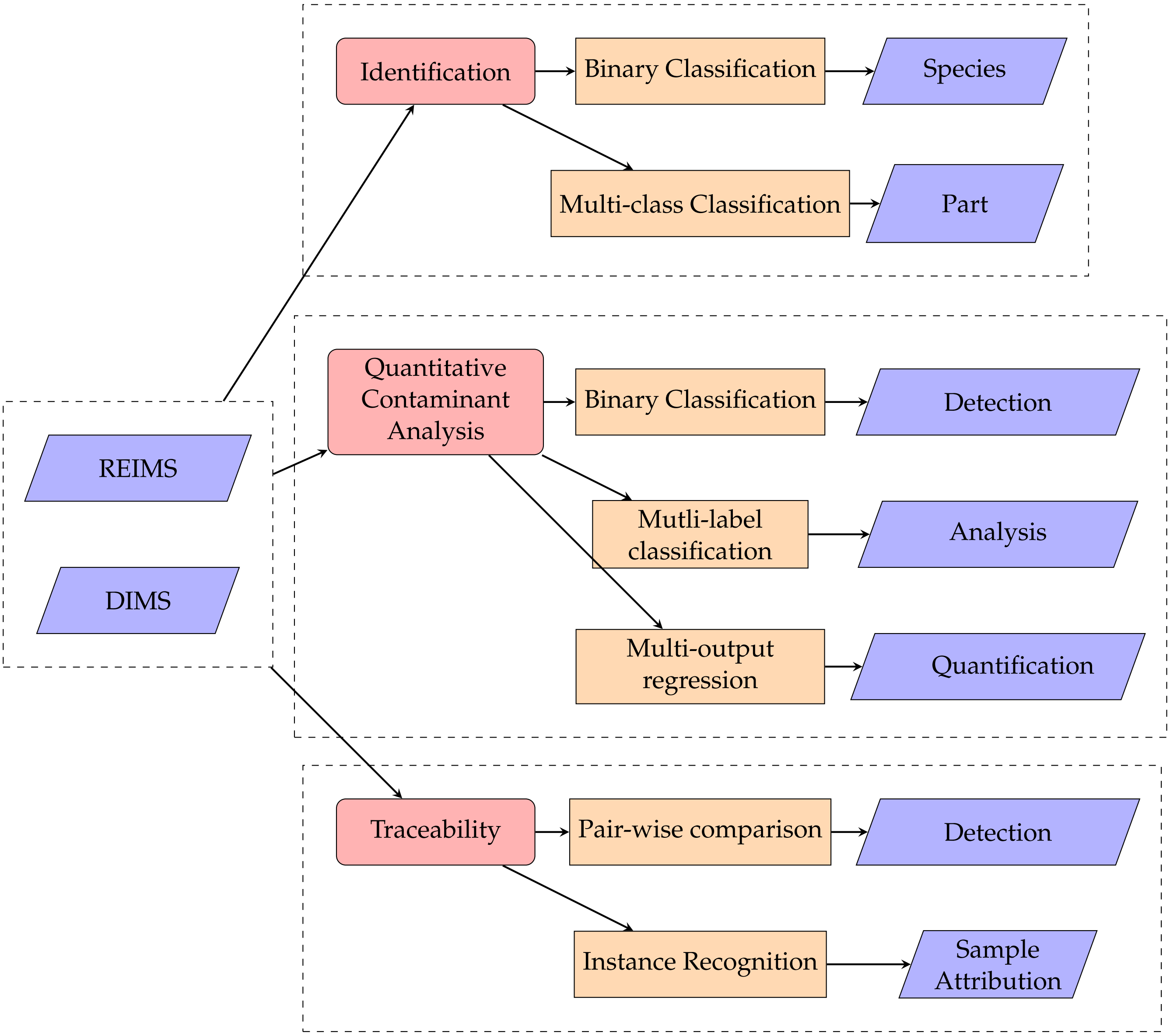


Figure 1.1: Research goals

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

The remainder of this section elaborates on concepts given both above and in figure 1.1. Each task is clearly defined concerning chemistry, fish processing and machine learning. This section is excellent reference material for downstream applications of machine learning in fish processing presented in this proposal.

1.7.1 Identification

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

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

Figure 1.2: Hoki *Macruronus novaezelandiae*

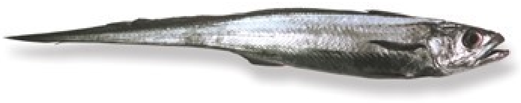
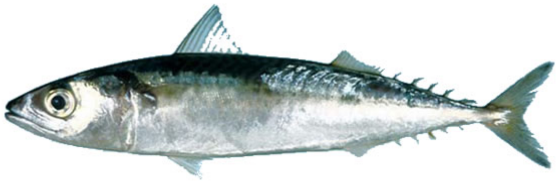


Figure 1.3: Mackerel *Trachurus symmetricus*



Part - the second sub-task is to predict the part (or tissue) where the measurement was taken. For these mass spectrometry datasets, there are six different classes of fish parts, e.g. fillet, liver, skin, guts, frame, and heads. This is trickier than the species identification because there are more than two classes. This constructs a multi-class classification problem, where a model predicts the fish part given many classes. Previous work [18] on chemistry datasets has also demonstrated that classifying fish parts proves more difficult than species. This is perhaps because there are more differences between two different species than there are similarities between the same part, but from different species. This work however was on Gas Chromatography datasets, which differ from the Mass Spectrometry dataset here.

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

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

Other types of fish, such as salmonid, shellfish and freshwater are excluded, as salmonid and shellfish typically belong to aquaculture [1]. Freshwater fish are caught in lakes and rivers, not the salt-water trawling vessels that are the scope of the Cyber-Marine flex-factory [7]. These types of fish can be excluded from the scope of the research, as they come from different sources entirely, and don’t require sorting via rapid mass spectrometry analysis.

1.7.2 Quantitative Contaminant Analysis

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

Detection - the first sub-task identifies if samples are contaminated. Contamination detection constructs a binary classification task, that predicts true or false if a sample is contaminated. Detection is unaware of which contaminants are present, and be thought of as a red flag, that warrants further investigation. In a factory setting, the detection task may provide higher recall than other models. An accurate contamination detection model can be used to identify areas of concern where future investigation may be warranted.

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

Quantification - the third and hardest sub-task is quantification. This performs the first two sub-tasks implicitly, with the added difficulty of providing a percentage ratio for that contamination. Contamination quantification constructs a multi-output regression problem, that predicts the contaminants present, and their respective percentage that contaminant contributes to the composition of the sample. Take for example quantification for cross-species contamination. Consider a contaminated sample with a mixture of both Hoki and Mackerel fish species. Quantification would tell you what percentage of that contaminated sample is Hoki, and what percentage is Mackerel. Alternatively, for mineral Ool contamination, quantification would predict what percentage of the sample is mineral Ool, and the rest fish. This task is similar to a softmax activation [31] as the final layer of a neural network. For a multi-class problem with each class indexed in a vector, the value at each index would correspond to the probability of that class, and the vector represents a probability distribution (which sums to 1). The quantification problem assumes the composition of known classes makes up the entire sample. For example, for cross-species contamination, if quantification predicts 33% Hoki, the remaining $(100\% - 33\%) = 67\%$ would be Mackerel.

A model with too many false positives would result in many fish that have not been spoiled being thrown away, which is not practical in a commercial setting of fish processing, where profit margins drive decisions. Conversely, a model with too many false negatives would let too many contaminated fish slip through into fish products and has the potential to cause real-world harm to consumers. Both of these models would have no commercial value if they fail to identify contamination with accuracy metrics suited to the task.

Furthermore, the black-box generative adversarial networks (GAN)s [30], or obfuscated PCA-LDA [13, 15], aren't ideal for qualitative and quantitative analysis. If the model consistently and correctly identifies contamination, workers at the fish processing factory would want to know what that is, and seek to remove the source of the problem. Neither, inscrutable matrices of floating points [30], nor obfuscated principal components, help those workers to identify the root cause of the problems. This work seeks to address limitations of interpretability for contamination detection, by providing accurate models for contamination detection, that can also provide qualitative and quantitative analysis of what those contaminants were. Therefore the representation of the model needs to be interpretable to domain experts in chemistry, and the application of fish processing. Therefore, the feature embedding must be mass spectrographically [sic] meaningful, and relevant for marine biomass. This work seeks models that don't just give the right answers but also can show their work, and provide insight. This work seeks to address the limitations of GANs anomaly detection [30]. As mentioned before, GANs, require expensive computational resources, are sensitive to hyperparameters, and a large volume of data; and, suffer from mode collapse, training instability and limited diversity. This work is hardware-constrained, as it must perform inference in real-time, on commodity hardware in a factory setting.

1.7.3 Traceability

In a fish processing factory, samples close to the detected source of contamination can be tested to see if they originate from the sample. Instead, there are two samples and the model detects if they come from the same fish. This constructs a pair-wise binary classification problem, given two samples, to predict if they came from the same fish. In layman's terms ”same fish detection”. Sample detection is agnostic to individual samples, it merely predicts whether two fish are the same, but does not append a unique identifier to that fish. However, sample attribution does this.

Sample attribution - The second sub-task is an extension of detection. Sample attribution is not just interested in if two samples are from the same fish. It is concerned with keeping track of those individual fish too. The goal of sample attribution is to allow for comprehensive troubleshooting for contamination detection on the assembly line. Should a contaminant, known or unknown, be identified, staff at the fish processing factory are interested in knowing the scale of that impact. Traceability offers new diagnostic tools that can track an individual marine biomass sample path throughout an assembly line, relying on machine learning analysis of rapid mass spectrometry alone.

Sample attribution would take a batch of fish samples and could identify and isolate the individual fish present in that batch. This is a similar concept to semantic instance segmentation from computer vision. Both tasks are not only interested in distinguishing between classes but look to uniquely identify individual instances. Sample attribution is likely the most difficult task proposed in this research, as it requires a very sensitive model to learn individual fish from very-few shot learning. It must account for seasonal variation and distinguish between different species of fish, where each species has a different variance. Initial findings from PFR [5] showed there was more variation in individual Hoki samples taken from the same fish than there was from different Mackerel.

(Optional) *Multi-sample attribution* - A more complex form of this problem is multi-instance recognition, where a sample can contain multiple unique instances, for which a classifier model must identify each unique individual present. Multi-instance recognition identifies unique markers for one or an individual fish that may be present in a sample. This is a species case of cross-species contamination, where an instance contains two species of fish, and the model provides unique identifiers for each individual fish. This would prove useful in isolating and containing sources of contamination in fish processing when factory workers want to assess how widespread systemic contamination is within the assembly line. This task could be formed with the cross-species or mineral Ool contamination data, from the previous research objective. This research objective is marked as (Optional), should time allow, this is an interesting direction to explore. Alternatively, this could make an interesting field for future research.

This work seeks to address the explainable limitations of existing work [13, 15] by focusing on transparent, explainable and semantically meaningful models [56, 57, 58]. For industry adoption and application of real-world AI, models are needed that enhance the existing knowledge of domain experts and assist them in their roles. Rather than existing works [13, 15], which get the same answers (most of the time), but refuse to show their working in a way that can be understood by humans. Transparency and explainability are required to be trusted and used in real-world applications. To address the general limitations of deep learning methods, i.e. compute, sample complexity, and grad student descent; this research proposes models that can perform real-time inference deployed on commodity hardware in a factory. This does not rule out deep learning methods entirely, it simply suggests that once effective methods are found, they would have to be optimized for deployment on commodity hardware. These models would also have to handle data scarcity in a few-shot learning task. And remove domain expertise in chemistry and machine learning, by automating the parameterization of network architectures. Neural architecture search (NAS) [59, 60, 61, 62, 63] can be used to automate the selection of neural network architectures, or, evolutionary computation [64] methods such as genetic programming [65] methods provide an alternative to neural networks altogether. Both approaches remove the need for domain expertise in deep learning and the application domain (i.e. chemistry). Variable-length input allows for pretraining or transfer learning on other mass spectrometry datasets measured at different resolutions. This allows the possiblity for semi-supervised / unsupervised learning methods to be applied to a vast array of publicly available mass spectrometry datasets. To address the limitation of fixed-length input in existing work on contrastive learning [39, 41, 40], this research proposes a flexible model that can handle variable-length input. Compared to each other, the REIMS and DIMS datasets are low-resolution and high-resolution, respectively. A variable-length input model would be robust, and applicable to both datasets. An additional bonus of variable-length input is that the model is naturally suited to pretraining or transfer learning on other mass-spectrometry datasets each taken at different resolutions. This is particularly important for the few-shot learning task of instance recognition, where pretraining and transfer learning, are needed to balance data scarcity. This research proposes a robust model that is naturally suited to a representation with variable-length input.

In summary, The proposed machine learning techniques allow for a particular individual fish species or body part to be tracked throughout a complex assembly line, using only rapid mass spectrometry data and rapid analysis with machine learning techniques proposed in this research.

1.8 Summary

	Identification	**Contamination**	**Traceability**
Motivations	biomass analysis variable biomass data scarcity seasonal variation	contaminated fish multi-class/label profile contaminants	same fish mark individuals instance recognition few/one shot
Limitations	multi-class/label data scarcity interpretability concept drift	no quantification domain expertise GANs deep learning	few-shot learning siamese networks only computer vision
Tasks	species classification part classification	detection analysis quantification	detection sample attribution multi-sample attribution
Existing works	beef adulteration [15] fish species [13] fish species & part [18] REIMS [13, 15]	cross-species [13] anomaly detection [30] GANs [33, 34, 30] PCA-LDA [13, 15]	one-shot [36] few-shot contrastive [39, 41, 40] siamese networks [39, 40, 41] instance recognition [42, 43]

Table 1.1: Summary: limitations, motivations, research goals, and existing works

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

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

1.9 Organization of the proposal

This proposal is divided into four chapters, **introduction** (this), **literature review**,**preliminary work**, and **contributions and project plan**. Each chapter and section provides a brief description of its contents (like this one here) for clarity.

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

Please see the table of contents for a more detailed breakdown of the contents of this proposal. This document is structured with the suggestions in [66], and with inspiration from the layout of the very usable textbook and guide to user experience [67].

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 research 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**, **mass spectrometry**, **machine learning**, **evolutionary computation**, **automated fish classification** on GC-MS, and their respective **limitations**. Finally, the chapter concludes with a **summary**, which positions this proposal 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° tilt [68], 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 the 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 [3], with a noticeable lack of lipids. Objective 3, sub-objective 2, involves uniquely identifying different individual instances of Hoki. Not just Hoki versus Mackerel. Important for the robustness of AI models, as they understand the chemical composition of fish within a species, can change dramatically with seasonal variation. Intra-class variation is introduced as a challenge by seasonal variables. An AI model for species prediction of Hoki would need to account for this. Robust models would be able to identify all Hoki species, regardless of seasonal variation, what is called seasonal invariant. A more complex model for instance recognition, would perform tasks two fold, identify the species as Hoki, and use the seasonal variation as a potential marker for an individual. Seasonal variation is closely related to conceptual drift from data stream mining [69, 24]. 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 [70, 25] 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,

- **rapid evaporative ionisation mass spectrometry (REIMS)** [12]
- **direct infusion mass spectrometry (DIMS)** [14]

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 [12] has shown promise in beef processing, where it was able to detect horse meat contamination in beef. Most impressively, horse meat contamination was detected at 1-5% - very low levels [15]. 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 algorithms 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 [71, 72].

- **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 [9]. 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) [13].

- **gas-chromatography mass-spectrometry** - Previous work [18] demonstrated that gas-chromatography mass-spectrometry (GC-MS) can identify fish species with high accuracy. However, GC-MS techniques require significant time and domain expertise 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. This discussion all fits under the umbrella of ”machine learning”. The section explores biomass analysis on rapid mass spectrometry data, state-of-the-art deep learning methods for mass spectrometry analysis, and existing works on marine biomass analysis using machine learning. This application research area is a very niche highly specialized field of expertise, this results in next to no existing literature on the exact problem this proposal seeks to solve. Hence, the novelty of this research area. As it is not possible to draw on non-existent solutions to this research problem, the literature review draws an array of related work from tangential existing works on closely related (but not identical) tasks. Thus this section discusses; 1) existing machine learning techniques for biomass analysis, 2) state-of-the-art deep-learning methods for mass spectrometry analysis, and 3) existing works machine learning methods for marine biomass analysis; under the umbrella term ”machine learning”

In [13], 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 [13], each sample belonged to one species, however, for this problem, each sample can belong to multiple classes, e.g. a cross-species contaminated sample contains a mixture of two species.

[15] performed detection and identification beef adulteration. It can identify samples that are adulterated with offal, and specify which offal was present. This is not marine biomass, but instead machine learning analysis of rapid mass spectrometry applied to animal agriculture. Everything in their work [15] translates to this proposal's research, with the exception of the animal agriculture domain. This research focuses on marine biomass, whereas their research focuses on animal agriculture. The insights from their analysis are likely universal and can be applied to marine biomass analysis. Rapid mass spectrometry measurement techniques are a relatively new and niche innovation [12], so work from different application domains, should be considered.

The detection task of traceability is very similar to signature verification, a pair-wise comparison that predicts if two instances have the same origin, i.e. they both belong to the same fish, or the signatures match. Siamese networks [39] were originally developed for the task of signature verification. Given two signatures, an authentic signature known to belong to an individual, and the ”query signature” whose veracity is being tested, determine if the query and reference signature were written by the same person. The model would predict if a signature is genuine or forged. The *detection* task in traceability is a simplification, a pair-wise comparison between two rapid spectrometry samples, to see if they originate from the individual fish. Although signature verification and sample attribution are from different domains, the task is identical. Given two inputs, predict if they are the same. Siamese networks consist of two identification neural networks, sharing the same weights and architecture. Given a pair of signatures, a reference, and a query, one network takes the reference, the other network takes the query. The output of both networks is combined using a distance metric, to produce a similarity score. In the paper [39], the Euclidean distance was used to compute the distance between the two outputs. The score would indicate the similarity between the two signatures, the closer the Euclidean distance, the more likely the query was genuine. The greater the distance, the more likely the query was forged.

Works on few-shot instance recognition [36, 42, 43] show applications in the computer vision domain. However, mass-spectrometry is not computer vision, as the mass spectrograph is not a pixel image. Despite the different domains, deep learning methods work on MS data [73]. In [73], a CNN achieves 93% accuracy fingerprinting GC-MS data. In 1998, Turing award winner LeCun et al. developed LeNet [48, 49, 50, 51], the first convolutional neural network, it was used for handwritten digit recognition. This CNN architecture has revolutionized deep learning, with automatic feature extraction via filters. Through a combination of convolutional, pooling and fully connected layers, CNNs are trained to learn filters that can detect specific features in the dataset, features important for downstream applications, such as classification. These networks are specialized for datasets with local connectivity, such as images [50], audio, or chemistry datasets [73, 32]. In 2012, Alex Krizhevsky in collaboration with Ilya Sutskever, and another Turing Award winner Geoffrey Hinton, developed AlexNet [52], which won the 2010 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) - a classification task with 1000 different classes. AlexNet consists of eight layers, including five convolutional layers and three fully connected layers, a standard deep CNN architecture.

Its success inspired the development of deeper CNN architectures, such as VGG-16 and VGG-19 [53], and GoogleNet [54] In the 2014 ILSVRC, GoogLeNet [54] won, and VGG-16 [53] came second. GoogLeNet [54] was the winner of the ILSVRC 2014 challenge. This deep CNN model was introduced by Christian Szegedy et al. in 2014. It consists of 22 layers and uses a novel architecture called the Inception module, which allows for efficient use of computing resources by using multiple filter sizes at each layer. VGG-16 and VGG-19 [53] came second in the ILSVRC 2014 challenge. These are deep CNN models introduced by Karen Simonyan and Andrew Zisserman in 2014. They consist of 16 and 19 layers, respectively, and have a uniform architecture that only uses 3x3 convolutional filters. VGG-16 and VGG-19 achieved excellent performance on the ImageNet dataset and helped establish the importance of deeper CNN architectures. VGG-16 and VGG-19 [53] are deep neural networks, which introduces the problem of vanishing gradients. Deeper neural network architectures - such as VGG-16, VGG-19 [53], GoogLeNet [54] - introduce the problem of vanishing gradients. Backpropagation computes gradients with reverse mode automatic differentiation, that computes gradients by chain rule [74]. It requires one forward pass to calculate the errors, and one backward pass to propagate the proportional error of each weight and adjust those weights accordingly. The chain rule takes the product of nearly n small numbers for a n-layer network, when computing early layers, resulting in vanishingly small gradients (error signal) for early those early layer.

ResNet [55] addresses the limitation of vanishing and exploding gradients, by adding skip connections, which act as gradient superhighways between layers, these allow the gradient to flow freely between non-adjacent layers. This is a family of deep CNN models introduced by Kaiming He et al. in 2015. ResNet models use a residual learning framework that allows for training of very deep networks by adding shortcut connections between layers. The original ResNet model (ResNet-50) consists of 50 layers, and the deeper versions (ResNet-101, ResNet-152) have up to 152 layers. ResNet achieved state-of-the-art performance on several computer vision tasks, including ImageNet classification and COCO object detection.

2.4 Evolutionary Computation

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 [75], 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 [76]. In later work from 1996, Dawkins proposed the evolved imagination, where every organism is a microcosm of its environment.

> ”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.” [77]

This is the bread and butter of EC, where each individual is a candidate solution, an approximation of a 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.” [77]

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 [78] and LeCun [79] have argued strong AI require an explicit world model [80].

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 [18], the EC technique of particle swarm optimisation (PSO) [81] 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) [65] and multi-tree genetic programming (MT-GP) [82, 83] are used for feature construction and classification.

2.5 Automated Fish Classification on GC-MS Data

The preliminary work starts by introducing previous research [18], 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 the reader be interested in gas-chromatography mass-spectrometry (GC-MS), species and part identification, this paper would suffice [18], as supplementary reading material, to avoid repetition, the contents of that paper are ommited.

2.6 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 or Online learning, 4) no taxonomy (lost in translation).

2.6.1 Domain Knowledge

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) [84]. 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 [23] 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, [13, 15] in the REIMS literature suffers from this same critique. That critique is nuisance variables, e.g. hyperparameters, for setting up statistical models chosen by chemists without theoretical or data-driven justifications. Hyperparameters seem to be chosen rather arbitrarily by humans, for example, the number of principal components, relative standard deviation (RSD) threshold for outliers, and mass range for mass-spectrometry in [13, 15]. An automated model that programmatically searches the hyperparameter space for ideal configurations for these variables. Or models could be chosen that don’t need those hyperparameters at all! Instead of handcrafted rules of thumb discovered via trial and error, this research aims to automate exploration of the hyper-parameter space through intelligent heuristics. This reduces the need for domain expertise in chemistry to design models and avoids falling into the same pitfalls of previous work.

2.6.2 State-of-the-art Biomass Analysis

Mature statistical techniques are used for dimensionality reduction and classification, not state-of-the-art machine learning. In the existing literature, principal component analysis

(PCA) [28] is the only dimensionality reduction technique used. PCA is limited as it does not take into account feature interactions, interactions with class labels, and feature redundancy/relevance.

Other dimensionality reduction methods should be considered. T-distributed stochastic neighbour embedding (t-SNE) creates a probability distribution of the similarity between points in the high-dimensional space. It defines a similar probability distribution over points in the low dimensional space. Then minimizes the Kullback-Leibler (KL) divergence [35] between the two distributions. uniform manifold approximation and projection for dimension Reduction (UMAP) utilizes a theoretical framework based on Riemannian geometry and algebraic topology. The method makes three assumptions, the data is distributed on a Riemannian manifold, the Riemannian metric is locally constant (or approximately so), and the manifold is locally connected. Given these assumptions, UMAP can model the manifold with a fuzzy topological structure. The data is projected into a lower dimension with the nearest approximate fuzzy topological structure.

Basic supervised statistical models (e.g. LDA, OPLS-DA) have been used for classification. Future work should consider CNNs [32, 73], GANs [30] and Diffusion [85, 86], and genetic programming for feature construction [82, 83] or feature selection. CNNs [48] are powerful on datasets that contain local connectivity, such as images, where a collection of neighbouring pixels may represent an edge. Adjacent features have a very close mass-to-charge ratio, the x-axis of the data of MS data, is locally connected. Interpretability Research suggests salience maps as CNN explainers are misleading [87]. They are independent both of the model and of the data generating process, and visual assessment of the salience map alone is misleading, as edge detectors produce eerily similar output. While a CNN peak detector would be effective, a model that can identify peaks, the y-axis of mass spectrometry data, salience maps are not suitable explainers for its interpretability. Previous work has shown CNNs are very effective on mass spectrometry [32] and other chemical datasets [73]. GANs have proven useful for detecting anomalies with outlier thresholding techniques [30]. GANs can be used to apply a latent bottleneck for high-dimensional datasets to make the data more amenable to downstream applications. Those downstream applications include the machine learning tasks proposed in this research. GANs can also be used to synthesize high-quality data, to artificially increase the low sample volume. Diffusion models offer an alternative approach to GANs. The recent breakthroughs in diffusion [85, 86, 88, 89] have shown diffusion models can often outperform GANs at the same task. denoising diffusion probabilistic models (DDPM) [85], the original diffusion paper, behind diffusion-based image generation models. denoising diffusion implicit models (DDIM) [86], a generalized DDPM that is faster and deterministic. [88] provides a clear explanation for the design of diffusion-based generative. [89] proposed ControlNet to add conditional control to diffusion-based models. Additionally, diffusion models are excellent at denoising, and could be used to denoise the dataset in pre-processing. The representation of deep learning methods are inscrutable matrices of floating points.

Alternatively, methods from evolutionary computation offer interpretable models that can be understood by domain experts. This differentiates them from deep learning methods, such as GANs or diffusion. GP uses tree-based representation, consisting of a terminal set and function set, that can be simplified into an arithmetic expression.

2.6.3 Transfer learning

Mass spectrometry produces high dimensionality datasets. However, due to the expensive equipment, manual labour and domain expertise, there is a low sample volume. Low sample volume means there are few instances for each class, which leaves a few-shot learning task. Transfer learning can apply knowledge learned from one task to another related task. There is a large body of existing mass-spectrometry data that can help compensate for the low sample volume of the REIMS dataset. There are many ways to utilize this knowledge including semi-supervised/unsupervised learning, pre-training, and synthetic datasets.

Semi-supervised / unsupervised learning can utilize partially labelled and unlabelled data. Unsupervised learning techniques don't require labelled data. Semi-supvervised learning uses both labelled and unlabelled data. Zemma et al. [90] incorporated unlabelled instances to draw more accurate support vectors and improve the classification accuracy for breast cancer diagnosis with SVM [91]. The key idea, unlabelled data can help a model fit the underling distribution of the data to generalize on unseen data. Existing MS datasets for other tasks, can be stripped of their class labels, and used in semi-supervised learning as unlabelled data. Unlabelled data from related tasks provides a free performance boost.

Chemists are interested in the lipid profile of marine biomass in fish processing. Lipids are a broad group of organic compounds, that include fats, waxes, sterols and fat-soluable vitamins. Databases exist for known lipid profiles, e.g. METLIN metabolites database [92] and LIPID MAPS [93]. These provide reference for annotated mass spectra labels, as demonstrated in [15]. Existing databases for mass spectra can provide label annotations for important features (e.g. significant markers). region-based convolutional neural networks (R-CNN) [94] extract regions of interest in object detection. Allowing for a black-box method to provide predictions that can be easily understood, e.g. this area of interest (bounding box) is like class X. Using a 1-D R-CNN for mass spectrometry would yield regions of interest for each prediction. Then label annotations for mass spectra in that region, would provide lipid profiles for significant markers.

mass-spectrometry (MS) is not a dataset fornautral language processing (NLP), but insights from NLP models such as transformers can help. Pre-training trains a model on a related task then fine-tunes that model on the desired task. Deep learning methods perform better with greater volumes of data, but human annotated label data is expensive. Unsupverised pre-training [95] rose to popularity for transformers [96, 97]. For large text corpa, simple and unsupervised tasks such as next sentence prediction (NSP) and masked language modelling (MLM), gained popularity for learning a semantically meaningful text embedding. Unsupvervised tasks, such as Next Spectra Prediction and Masked Spectra Modelling, inspired by NSP and MLM could be used for pre-training on other MS databases [93, 92]. The pre-training allows for features embedding to capture a larger distribution of mass spectra data, the rich feature embedding could improve few-shot learning on the desired task.

To address the limitation of low sample volume, synthetic datasets should be considered. Synthetic dataset can artificially inflate the sample volume. Statistical methods [98, 99, 84] GANs [33, 30] or Diffusion-based [85, 86, 88, 89] methods can synthesize realistic samples. Synthesized samples can be filtered for quality by similarity comparison to real world MS data [93, 92]. 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 [98], uses Synthetic dataset to test effectiveness of feature selection algorithms, in a simulated customer data problem. In the original paper for χ^2 feature selection algorithm [99], 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 Relief-F paper [84], 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 data scarcity, 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 sufficiency, to provide more training data.

2.6.4 Online learning

The Cyber-Marine flex-factory, [7], is a real-world application where the algorithms proposed in this research will be deployed. The flex-factory is a real-world application that provides a continuous stream of data, the marine biomass being processed by the factory, making it uniquely suited to online learning. There is an important distinction between online and offline learning [100]. Online learning means that you are doing it as the data comes in. The model is constantly trained on new data, as it becomes available. Offline learning means that you have a static dataset. The model is trained once, on the existing data available at the time, and deployed as is. A smart system would amortize daily test data coming from the factory, by re-training daily to dynamically adapt. Applications in identification, contamination and traceability must perform real-time inference on the factory floor. Deploying online learning in a factory naturally increases the volume of training data over time. As the datasets grow the quality of the model will likely improve, and yield models more robust than their offline counterparts. Online learning provides algorithms that can dynamically adapt to new patterns in the data.

AI models often fail to generalize to unseen data, especially in cases of out-of-distribution anomalies [30], or conceptual drift [24, 25] where the distribution of the data changes with time. An example of an out-of-distribution anomaly would be a new type of contamination that the flex-factory has not seen before. This anomaly would be a contaminant with no reference data to identify automatically. However, this sample should be flagged as anomalous, isolated/removed from the production line, and sent away for further testing offline to identify the unknown contaminant. Using this process, if unknown contaminants are repeatedly found, systems can identify new sources of contamination to the factory workers. Taking this one step further, chemists can provide supervised label annotations for unknown contaminants and append those newly labelled contaminants to the dataset, for automatic detection in the future. This outlines human-in-the-loop online learning process [100], where out-of-distribution contaminants are automatically detected, manually annotated, appended to future training data, and then automatically detected. This human-in-the-loop online learning is a powerful method to bootstrap algorithms for robustness.

An example of concept drift would distribution of Hoki caused by seasonable variation in their composition [3]. For a species classification model, the factory could provide supervised label annotation of fish species each season for a select few samples. These samples would serve as a quality control to measure concept drift for the data. Furthermore, these examples could be appended to the existing datasets, to ensure models are robust to concept drift caused by seasonal variation, or other unknown factors.

2.6.5 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 is the start of building that bridge between these disciplines. This research will create a *taxonomy to foster multi-disciplinary collaborative work in marine biomass analysis of rapid mass spectrometry data with machine learning*.

Jargon limits the dissemination and communication of interdisciplinary research. Jargon is the highly-specialized terminology used to describe methods in a particular field, e.g.

chemistry, biology, statistics, machine learning. For interdisciplinary research to flourish, each discipline needs to understand the equivalence and difference in their semantics. For example, what chemists refer to as a variable, an ML researcher would call a feature. The terms may be used interchangeably, 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. An AI researcher would refer to a feature as a column in the raw mass spectrometry data, e.g. mass-to-charge ratio. However, a chemist would use the term variable to refer to higher-level domain knowledge For example, if they are interested in measuring lipids, the lipid would be a variable of interest. The lipid is not explicit to raw mass spectrometry data, mass-to-charge ratio versus intensity. But its presence is implicit and obvious to those knowledgeable in the application domain, lipids can be explicitly found by matching to reference spectra label databases [93, 92]. When a chemist says variable it is inherently linked to domain-specific knowledge and means a very specific thing. An AI researcher may be confused by AI terminology without sufficient domain knowledge, and vice versa for the chemist. Here is some other domain-specific jargon that could cause some confusion are quality control (QC) [13, 15], outliers [13, 15], and Significant markers [13, 15].

It is important for researchers without the necessary background in chemistry to understand this jargon for the specific domain of rapid mass spectrometry. But without gatekeeping, this research must communicate the fundamentals of the necessary chemistry and ML for a specialist in one domain to understand the work. A clear explanation of the jargon will encourage further multi-disciplinary work in this field.

2.7 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 in [13, 15]. 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 [13, 15]) was used for classification. Future work should consider CNNs [32, 73], GANs [30], Diffusion [85, 86], or Evolutionary Computation [82, 83, 18].

- **Transfer learning** - There is a large body of existing mass-spectrometry data [93, 92]. 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 AI models completely collapse when presented with new data, whether that be out-of-distribution anomalies [30], or conceptual drift [24, 25] where the underlying probability distribution changes over time - for example seasonal variation in composition of Hoki [3]. 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.

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. This section discusses and **Exploratory data analysis** on the rapid mass spectrometry dataset, and **Genetic programming for GC-MS dataset**.

3.1 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.

This section 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.

3.1.1 Annotated Labels

Figure 3.1 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.

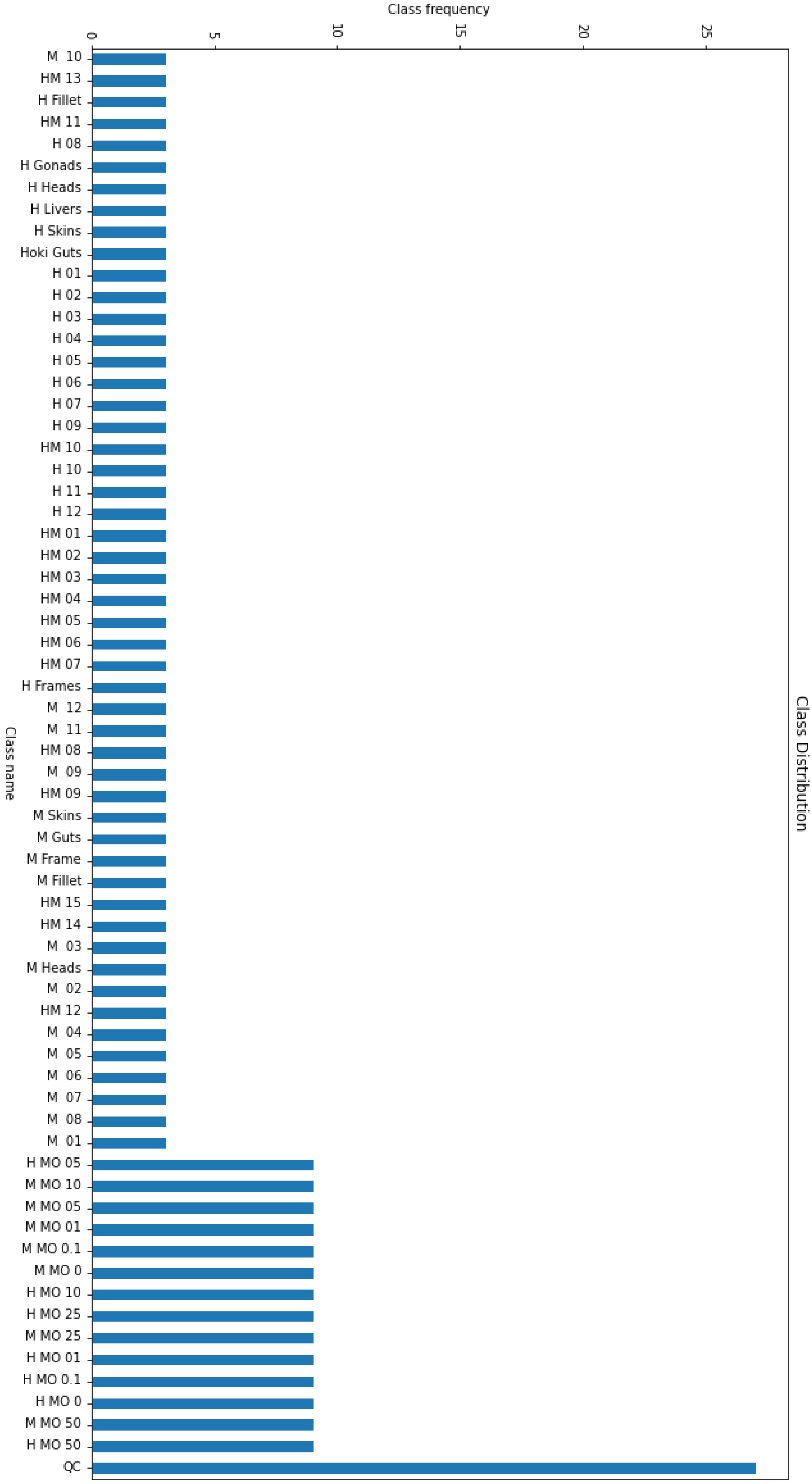


Figure 3.1: Class distribution

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 → Hoki - a species of fish.
- M → Mackeral - a different species of fish.
- HM → 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 Ool (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 Ool (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 Ool, at a contamination rate of 0.1%. The mineral Ool 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.

3.1.2 Species Identification

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

3.1.3 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 3.2 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.

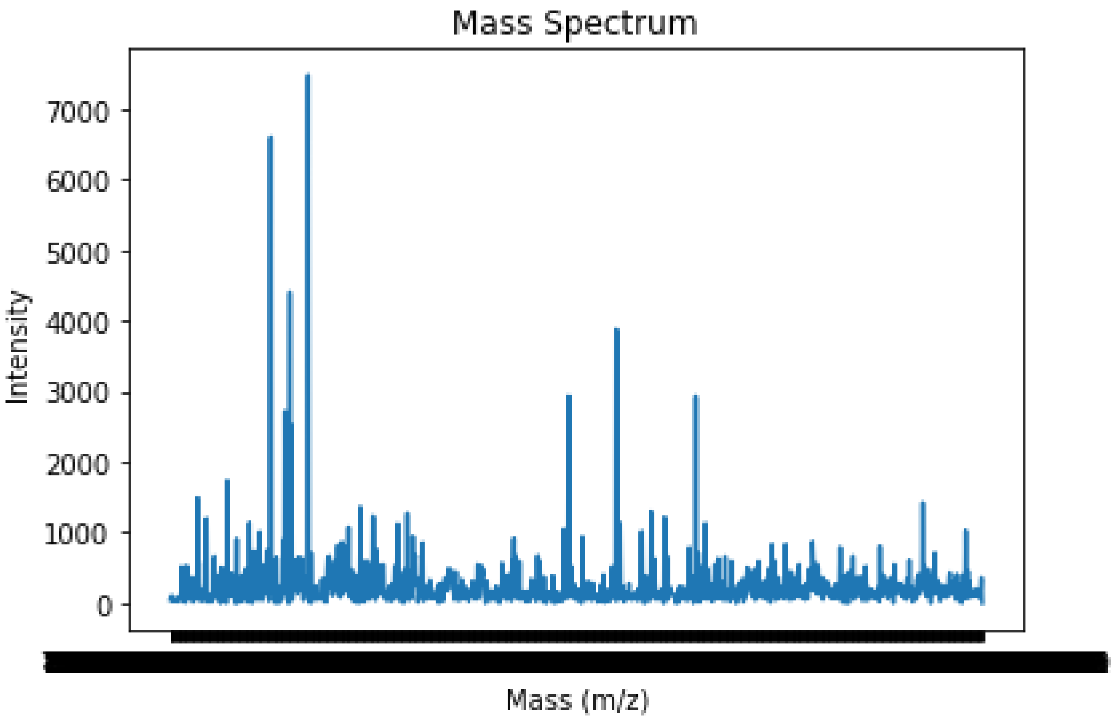


Figure 3.2: Mass spectrum for a Hoki fillet

Figure 3.3 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 3.1.

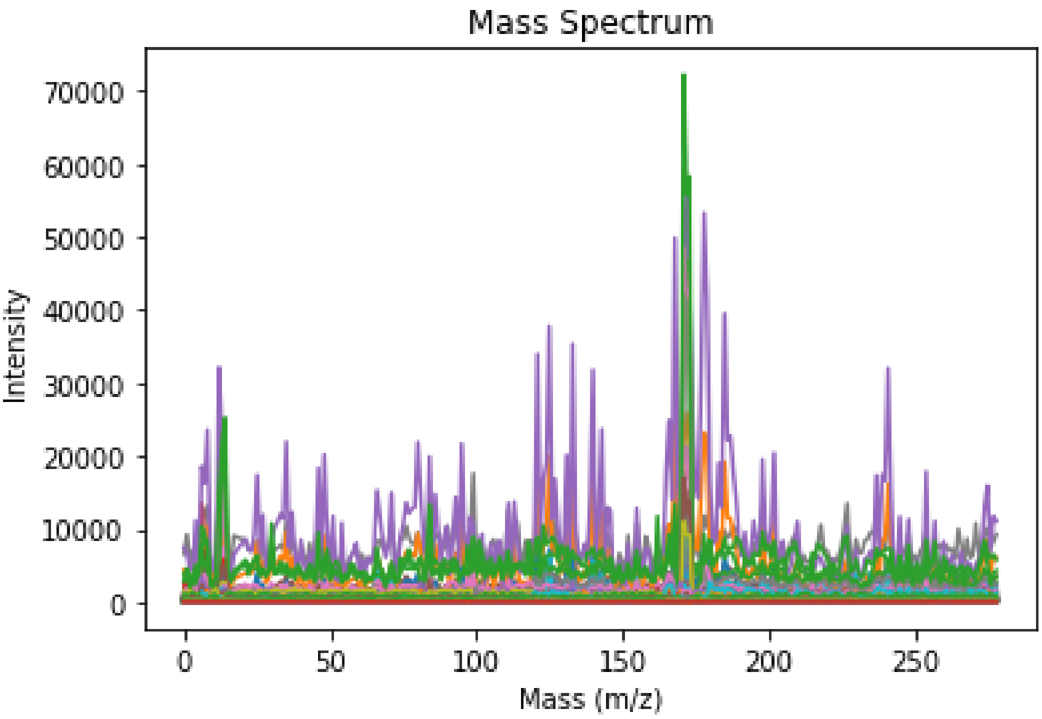


Figure 3.3: Mass spectrums for entire REIMS dataset

3.1.4 Results

Table 3.1 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 [91]	KNN [101]	DT [102]	RF [103]	XGBoost [104]	**LDA** [105]
PCA [28]	0.88 ± 0.17	0.85 ± 0.13	0.83 ± 0.15	0.87 ± 0.13	0.88 ± 0.14	**0.92** ± **0.13**
t-SNE [106]	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 [107]	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 3.1: Dimensionality reduction / classification Methods for Species Identification

The table shows PCA-LDA [28, 105] (**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, [13, 15] 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.

The EDA provides insight into suitable models for the dataset. PCA [28] 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. The poor performance of UMAP could be attributed to three assumptions of algorithm not holding for this dataset. The data may not be uniformly distributed on a Riemannian manifold, or the Riemannian metric may not be locally constant, or the manifold is not locally connected, or all of the above [107]. Denoising the dataset had a significant effect on the classification performance. This suggests it may be an important step in preprocessing, where PCA can be used in combination with classification models. Or, that a model with implicit denoising, such as a denoising auto-encoder [20] with a fully connected network for each sub-task, may yield noteworthy results. Furthermore, GANs have shown promise in anomaly detection [30], which is a closely related field to contamination detection and identification presented here.

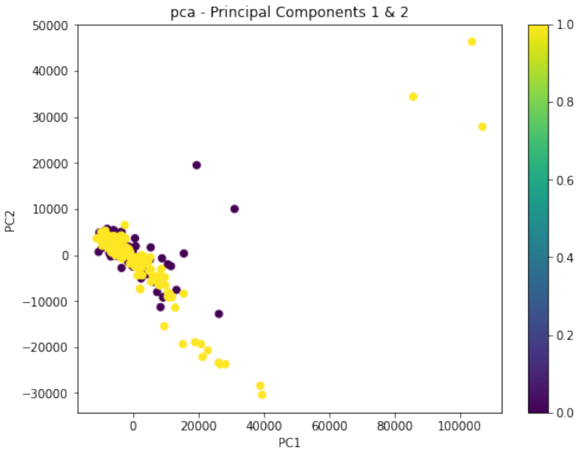
3.1.5 Ablation Studies

The ablation study 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.

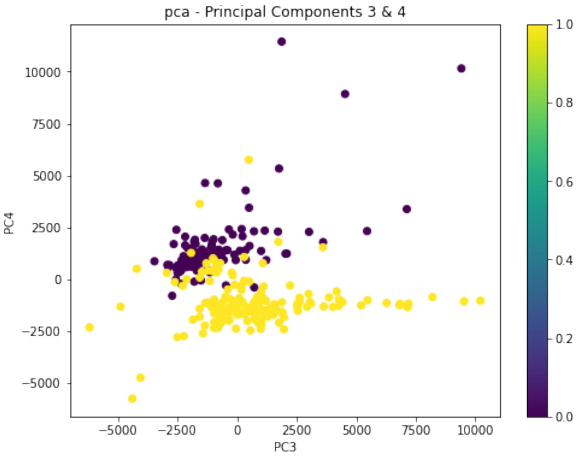
Table 3.2: Visual intuition for dimensionality reduction techniques and their respective feature subsets

PCA [28]

Features 1 & 2

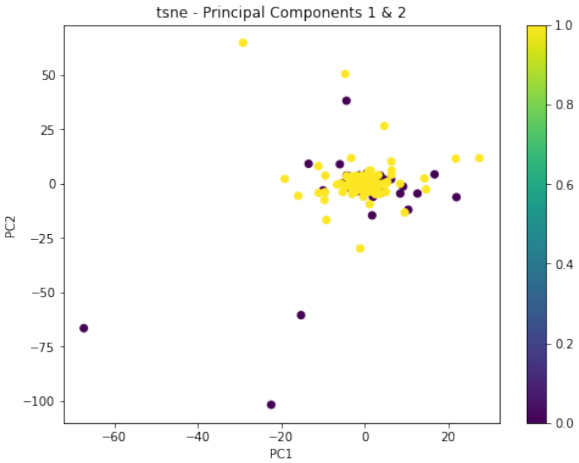


Features 3 & 4

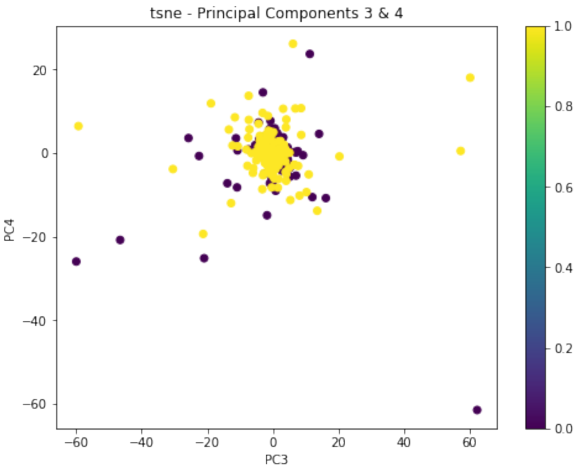


t-SNE [106]

Features 1 & 2

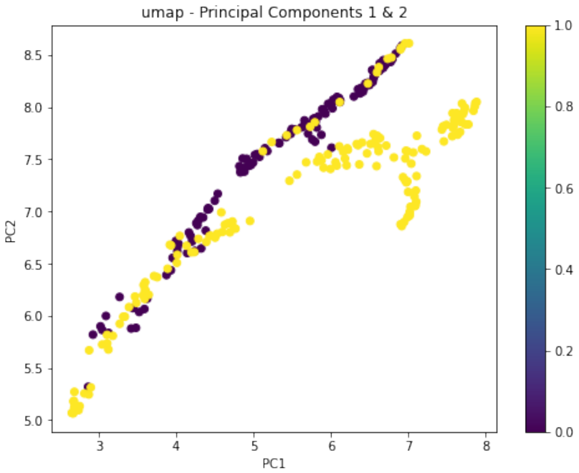


Features 3 & 4



UMAP [107]

Features 1 & 2



Features 3 & 4

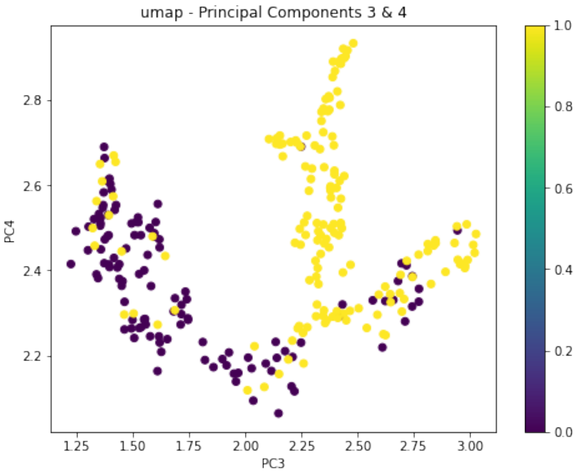


Table 3.2 gives the class distribution for features 1 & 2, versus features 3 & 4, for each dimensionality reduction method, PCA [28], t-SNE [106] and UMAP [107]. This gives intuitive and visual proof of the ability of each technique to tolerate noise in the dataset. The results agree with the conjecture 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 [106] and UMAP [107], struggle to extract and isolate this noise, as the class distribution remains muddles for both features 1 & 2, and features 3 & 4.

Table 3.3: Empirical evaluation of dimensionality reduction techniques and their respective feature subsets

Method	Features 1 & 2	**Features 3 & 4**
PCA [28]	55.47± 6.68	**86.40 ± 16.25**
t-SNE [106]	57.24 ± 2.03	55.80 ± 3.69
UMAP [107]	85.27 ± 15.17	81.23 ± 17.15

Table 3.3 gives the cross-validation score for each dimensionality reduction method, PCA [28], t-SNE [106] and UMAP [107], 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 [28] 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 [106] and UMAP [107], suggesting that these provide an excellent signal for the species identification task.

The results demonstrate 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 [28] 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.

3.2 Genetic Programming for GC-MS Data

This section describes preliminary work using genetic programming (GP) on gas-chromatography mass-spectrometry (GC-MS). 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. In the genetic programming (GP) subsection of the preliminary work, experiments benchmark three GP methods, to my previous work, [18], that was addressed in the last subsection. In particular, the three GP methods proposed in this work are 1). single-tree genetic programming (ST-GP), 2). multi-tree genetic programming (MT-GP), and 3). multiple class-independent feature construction method (MCIFC).

ST-GP - The first method, ST-GP, is a standard genetic programming (GP) [65]. MT-GP extends single-tree GP, now it returns multiple trees - a list of single-tree GPs [82, 83]. MT-GP can be thought of as multiple single-tree GPs working together to make a multi-class classification prediction. The trees combine to make the class prediction. MT-GP is more expressive than ST-GP, as each multiple tree constructs multiple features, giving a richer feature set to make predictions form. Whereas ST-GP needs to learn to predict multiple classes with only one output, a difficult task, that can be imperfectly solved with classification maps [108] for multi-class classification.

MT-GP - Algorithm 1 shows the pseudo-code of the multi-tree genetic programming (MT-GP). The multi-tree representation has *m* trees. The limitation of multi-tree genetic programming (MT-GP), is that the class information is entangled, i.e. randomly split between all the trees, without any particular preference or logic.

MCIFC - Disentangling the class prediction allows for the simultaneous training of multiple trees, each specialized in predicting only one class - this is multiple class-independent feature construction method (MCIFC). This method has a separation of concerns, where each tree has a dedicated class to predict, and nothing else. MCIFC, shown in [82, 83], is an extension of MT-GP that ensures each tree corresponds to a class, and its value, high or low, predicts the likelihood of that class. MCIFC only has to worry about learning to predict its respective class. Note: the pseudocode for simpler algorithms for ST-GP and MT-GP are trivial, compared to MCIFC, and whose derivation are left as an exercise for the reader.

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;
```

3.2.1 Representation

Single-tree genetic programming (ST-GP) has a genotype of an arithmetic syntax tree [65]. To evaluate the output of the phenotype of single-tree representation, one feeds feature values as arguments into the leaf nodes of that arithmetic syntax tree, and calculates the output of the resulting arithmetic expression. Reverse polish notation can be a useful condensed shorthand for representing the evaluation of single-tree GP.

The genotype of multi-tree genetic programming (MT-GP) extends ST-GP to include as many trees as there are classes in the dataset. The evaluation of individual trees is identical to the above, however, a winner-takes-all approach is used to evaluate phenotype, which is the class whose tree corresponds to the largest input, corresponding to the MT-GP's prediction.

Multiple class-independent feature construction method (MCIFC) [83] is MT-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.

3.2.2 Crossover and Mutation

Crossover - In standard single-tree genetic programming (ST-GP) using the DEAP library for multi-class classification, the crossover operation is a fundamental genetic operator that combines two parent trees to create one or more offspring trees. Here's a brief description of how the crossover operation works:

1. Selection of Parent Trees: Two parent trees are selected from the population using a selection method like tournament selection or roulette wheel selection. These parent trees represent potential solutions to the problem.
2. Subtree Exchange: The crossover operation identifies one or more subtrees within each parent tree. These subtrees are typically chosen randomly, but there are various strategies for subtree selection, such as choosing subtrees of similar size to maintain balance in the offspring.
3. Exchange Subtrees: The selected subtrees from the two parent trees are swapped or exchanged to create two offspring trees. This swapping operation results in two new trees with genetic information from both parents.
4. Offspring Creation: The two offspring trees replace the parent trees in the population. These offspring trees inherit genetic characteristics from both parents, potentially leading to improved solutions over time through the process of evolution.

The crossover operation allows the genetic algorithm to explore the search space by combining the information from two different solutions. This process helps the algorithm discover new, potentially better solutions for the multi-class classification problem.

Mutation In single-tree Genetic Programming (GP), the mutation operator is a fundamental genetic operator that introduces small random changes to an individual tree within the population. This operator helps in exploring the search space by creating variations of existing solutions. Here's a description of the mutation operator in single-tree GP:

1. Selection of Parent Tree: First, one parent tree is randomly selected from the population. This tree represents a potential solution to the problem.
2. Node Selection: Within the selected parent tree, a random node (or subtree) is chosen. The node can be any part of the tree, including internal nodes (function nodes) or terminal nodes (leaf nodes).
3. Mutation: The selected node is replaced with a new randomly generated node. The replacement node can be of the same type (e.g., replacing an addition operation with another addition operation) or a different type (e.g., replacing an addition operation with a multiplication operation). The replacement node can also be a randomly selected terminal node if the selected node was an internal node or vice versa.
4. Offspring Creation: The resulting tree with the mutated node replaces the original parent tree in the population. This offspring tree is now a slightly modified version of the parent tree, introducing genetic diversity into the population.

The mutation operator in single-tree GP plays a crucial role in maintaining genetic diversity and exploring the solution space. By introducing random changes to individual trees, it allows the algorithm to escape local optima and potentially discover better solutions over time through the process of evolution.

MT-GP - Reproduction or the cross-over operator for multi-tree genetic programming (MT-GP), where trees are represented as a list, is where crossover happens to a subtree that is selected at random. Crossover operations are limited to parents from the same tree. Mutation happens to a tree selected at random when an individual is selected for crossover. The same crossover and mutation operators are used for MCIFC.

MCIFC - This method 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. Only updating one tree at a time for multi-tree methods ensures that an improvement to fitness for one tree by a crossover or mutation is not cancelled out by a decrease in fitness for another tree. Only changing one tree at a time guarantees monotonic improvement of fitness - that is the fitness can either remain the same or improve, it cannot get worse.

Algorithm 2 MCIFC crossover and mutation.

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

3.2.3 Fitness

Balanced classification accuracy is given by giving a stratified accuracy score, that measures the accuracy of the GP tree, scaled to the proportions of each class frequency, to prevent bias towards to majority class in the base of an imbalanced dataset. Balanced accuracy 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.

ST-GP - Single-tree genetic programming (ST-GP) uses classification maps for multi-class classification, as in [108]. This approach uses a number line where the user arbitrarily draws floating points that correspond to class boundaries, where there is an interval on that number line, for each respective class in the classification problem. If the phenotype, e.g the evaluation of GP tree produces an output within the interval for a given class, that class becomes the prediction given for the model.

MT-GP - Multi-tree genetic programming (MT-GP) has a phenotype of winner-takes-all. There is one tree for each class, and the class whose tree gives the largest output is the predicted class for multi-tree GP. The winner-takes-all is implemented as an

$$argmax$$

function of the evaluated output of all trees concatenated together.

MCIFC - MCIFC uses the same balanced classification accuracy as MT-GP, but also contains a regularization term that maximizes the distance between dissimilar classes, and minimizes the distance between similar classes.

3.2.4 Datasets

The dataset is the gas-chromatography dataset previously mentioned in previous work [18] that was discussed in the literature review. If needed please consult this paper for a comprehensive description of this dataset. 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 3.4 gives a gas chromatogram - the artefact of the gas chromatography - for tissue taken from the skin of a Snapper. Please see [18], for an example gas chromatogram and a more thorough description of the measurement technique.

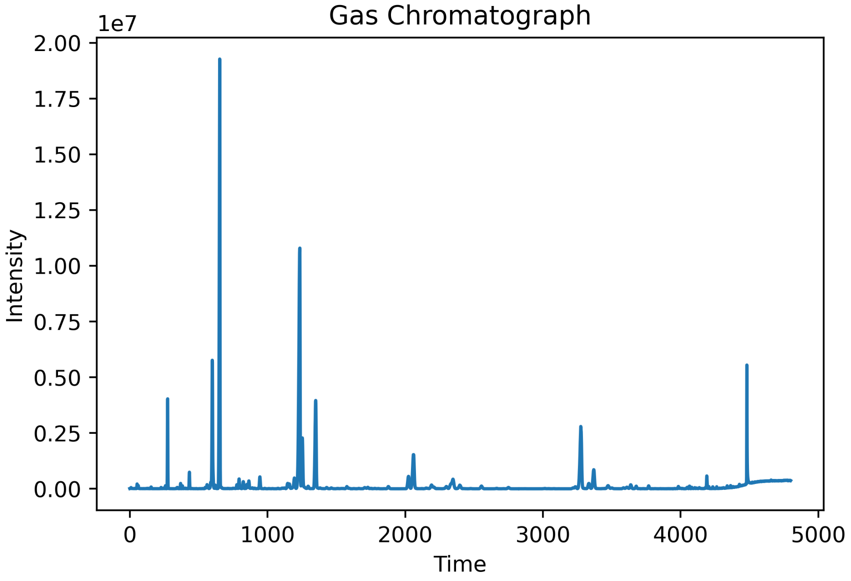


Figure 3.4: Gas chromatogram of fatty acid methyl esters from snapper skin.

Table 3.5: 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 3.4: 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 3.4 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 [18] 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.

3.2.5 Experimental Setup

Table 3.5 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.

3.2.6 Results

Table 3.6 compares the classification results from [18], 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 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.

Table 3.6: Results

Dataset	Method	Train	Test
Species	KNN [101]	83.57	74.88
	RF [103]	100.0	85.65
	DT [109]	100.0	76.98
	NB [102]	79.54	75.27
	SVM [91]	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
	MT- GP	**84.30**	**86.80**
	MCIFC	**97.81**	84.30

Evolutionary computation methods, MT-GP and MCIFC, both offer competitive performance to the traditional machine learning methods, KNN, RF, DT, NB, SVM, from [18], on chemical datasets for marine biomass. 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. These GP methods are compared to FS methods from [18]. Firstly, for fish species identification, MCIFC exceeds performance of all FS methods, [99, 81, 110, 111], with SVM [81] Secondly, for fish part identification, MCIFC is better than χ^2 [99] and the full dataset. MCIFC offers same performance as PSO [81] MCIFC is worse than ReliefF [110] and MRMR [111]. MT-GP offers competitive performance to MRMR [111], 86.80 % compared to 86.94 %, respectively.

3.2.7 Summary

This preliminary work explored evolutionary computation methods for fish species & fish body parts, binary & multi-class classification, respectively, on a gas chromatography dataset. Here is a summary of the key takeaways from this work:

- **Fish species** - MCIFC gives the best accuracy for training and testing on the fish species binary classification.
- **Fish parts** - The fish parts dataset, which is known to be harder from [18]. **MT-GP** - gives the best test accuracy for the harder fish parts multi-class classification.
- **MCIFC** - offers better training accuracy for the fish parts dataset but performs worse than MT-GP for the test dataset. **MCIFC** - unable to generalize as well as MT-GP on unseen data, likely overfitting the training data, which explains the stark differences in accuracy between train and test.
- **EC** - However, evolutionary computation methods MT-GP & MCIFC, both offer competitive results, that outperform traditional machine learning methods that were evaluated in [18]. Evolutionary computation methods are a strong competitor and alterna-

tive to traditional machine learning methods for chemical analysis of marine biomass, for downstream classification tasks in fish species and fish body parts.

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. AI 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 is explored in further detail.

4.1.1 Mass Spectrometry

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 [13] 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 [14]. 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 [13], REIMS results were worse on cooked biomass. Studies [13, 15] show that mass-spectrometry works on raw biomass products. A difference between the REIMS and GC dataset from [18], 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 - fish species and part, (2) qualitative contaminant analysis (QCA) - cross-species and mineral oil, (3) traceability - detection and sample attribution. For the remainder of this section, each sub-task is defined, concerning biology/chemistry/fish processing, and their relation to machine learning.

4.1.2 Identification

Species and part identification are a binary and multi-class classification problem. This research proposes classification models for both REIMS / DIMS datasets for these tasks. The species and part identification problem is the same as [18], but instead of GC-MS, it focuses on rapid mass spectrometry datasets. For each task, high classification accuracy and interpretable models are desired. Chemists are interested in significant markers [13, 15] - important features that can be reliable used to identify a class. Intrepretable models [18, 94] are ones whose outputs can easily be explained and understood by domain experts. These can be deployed in a factory settings, because their outputs can be analyzed and troubleshoot for diagnosis.

4.1.3 Quantitative Contaminant Analysis

qualitative contaminant analysis (QCA) is a few-shot multi-label multi-output classification/regression problem. It is supervised few-shot learning task because there are few training samples for each contaminant. Contaminated samples are either a mix of two species - e.g cross-species contamination, or a fish sample mixed with mineral oil. Contamination detection is a binary classification task, where the positive class indicates a sample is contaminated. Similar to the beef adulteration detection in [13], but instead of detecting horse adulteration, models detect cross-species and mineral oil contamination in fish. Contamination analysis is a multi-label classification task that identifies the class label(s) that are present in a sample. In multi-label classification task, a sample may contain one or more classes. Contamination quantification is a multi-output regression task, that associates what percentage of the sample is contaminated with that contaminant. That percentage can be used for quality control, to identify if contaminants are dangerous, and exceed thresholds safe for human consumption. QCA would yield models that predict the percentage of the sample that is contaminated, would add multi-output regression to the existing multi-label classification task. For real-world applications in a fish processing factory, the models need to be interpretable, and identify significant markers - or important features - that are used to identify contaminated samples.

A robust model would find not only existing contaminants that are expected, but new sources of contamination - black swans, classes that are possibly not even in the training data. Out-of-distribution test data could likely belong to a new class of contaminant. QCA should identify these black swans [112, 113], the unknown unknowns, or out-of-distribution classes of contaminants. Black swans, or unknown contaminants, cannot possibly be correctly classified in a supervised learning task, because the model is unaware the class exists. However, [30, 15] show that outlier thresholding techniques can identify these extreme 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. This is an example of the human-in-the-loop online learning process, outlined earlier, for robust models that can adapt to black swans.

4.1.4 Traceability

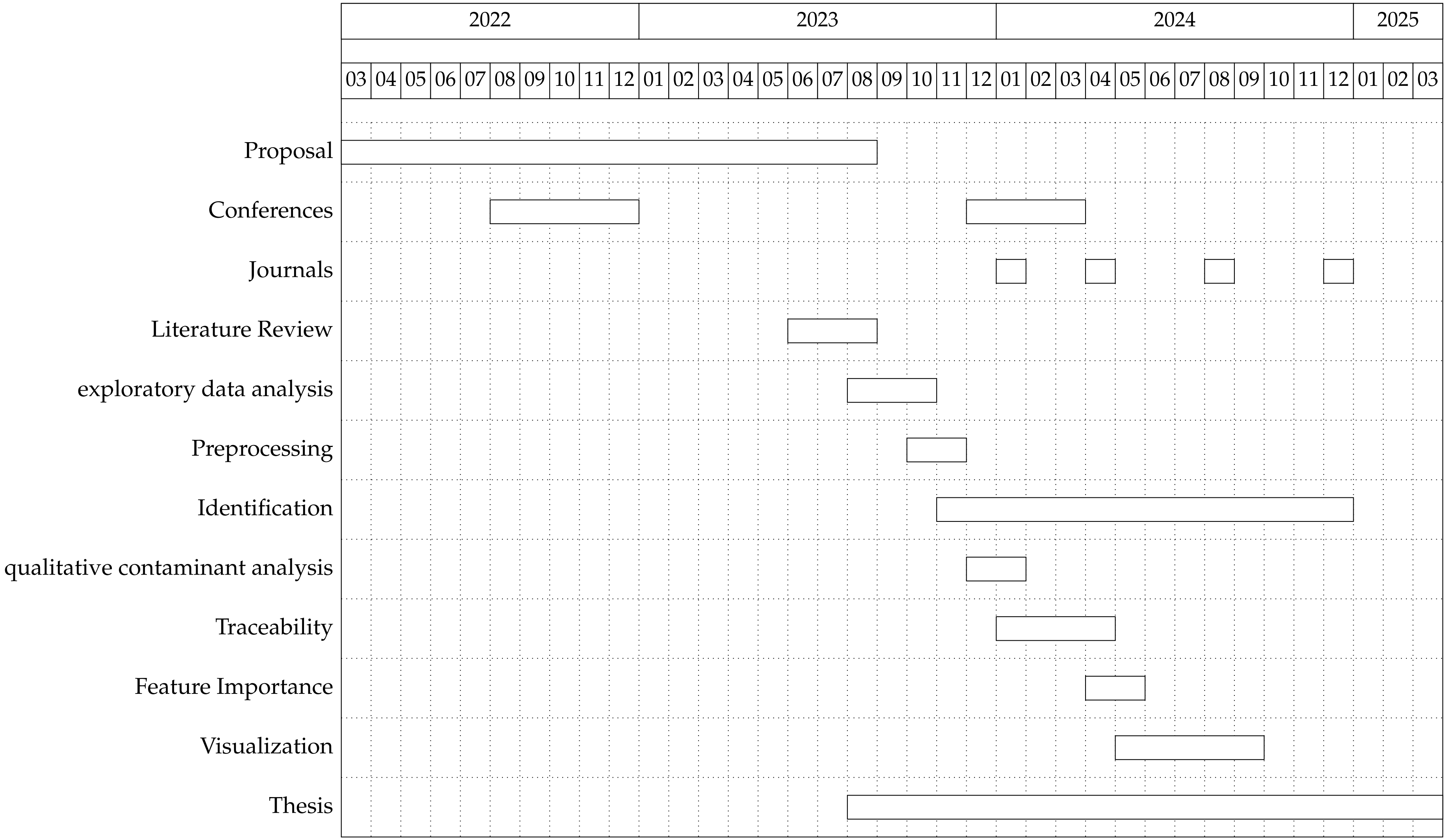
Firstly, detection is a pair-wise comparison is a contrastive learning task. The model learns between similar and dissimilar pairs of objects. Contrastive learning [39, 40, 41] learns representations that maximize the similarities between objects of the same class, and minimize the similarity between different classes. In this research, similar classes is two samples from the same fish, and dissimilar classes would be two samples from different fish. Secondly, sample attribution is the task of instance recognition. Instance recognition [42, 43, 36] learns unique markers for instances in the dataset, and can identify and differentiate individual instances. In this research, unique instances samples taken from the same fish. A real-world example of traceability would be the factory identifies a black-swan, an unknown contaminant, during QCA. Given it was not a known class of contaminant, the factory workers are concerned with how far that contamination has spread. For quality assurance they perform traceability detection on random samples throughout the factory, to see if that contamination is effecting other samples. Then afterwards, sample attribution would determine if that contamination is likely present in existing samples that have already been processed.

4.2 Milestones

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

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

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


2022
2023
2024
2025
03
04
05
06
07
08
09
10
11
12
01
02
03
04
05
06
07
08
09
10
11
12
01
02
03
04
05
06
07
08
09
10
11
12
01
02
03
Proposal
Conferences
Journals
Literature Review
exploratory data analysis
Preprocessing
Identification
qualitative contaminant analysis
Traceability
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:

Abstract

Glossary

Chapter 1 - Introduction

Chapter 2 - Literature Survey

Chapter 3 - Datasets and Processing

Chapter 4 - Fish Species and Part Identification

Chapter 5 - Fish Quantitative Contaminant Analysis

Chapter 6 - Fish Traceability analysis

Chapter 7 - A Case Study, Demonstrations and Discussions

Chapter 8 - Conclusions

Bibliography

Index

The thesis outline includes a glossary to bridge the multi-disciplinary gap in knowledge. Most readers will likely have expertise in one discipline. 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

4.4.1 Human resources

In addition to these resources, AI researchers have gained valuable experience through previous field trips to NZ Plant and Food Research, where AI researchers saw GC-MS first-hand for my previous publication [18]. 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, AI researchers 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.

4.4.2 Financial

Publications to conferences are to be expected, following on from [18], 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 [16] . 2, 5, 8, 9, 18, 21, 22, 47

AI artificial intelligence. 4, 6–8, 11, 13–15, 20, 24, 29, 30, 46

analysis Analysis is concerned with identifying which contaminants are present. Not to be confused with **detection**, which simply tells us if a sample is contaminated. Analysis takes this one step further and gives predictions for which contaminants are present in the fish tissue. Take for example, **cross-species contamination**, contaminant analysis predicts which species are present in a contaminated sample, e.g. detection: contaminated, analysis: Hoki and Mackerel both present . 2–6, 11, 15, 16, 18, 19, 22, 38, 47, 48, 51

anomalies Anomalies refer to out-of-distribution 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 [30]. In fish processing, an example of an anomaly would be a new species of fish, or marine biomass, that is not a labelled class or present in the training or validation data . 26, 28, 30, 47

black swans "Black Swans are events or pieces of knowledge that sit outside our regular expectations and therefore cannot be predicted." [114] Popularized by Nassim Taleb, a risk analyst, in his books [112, 113]. 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 [114] . 47, 48

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 . 33

CNN convolutional neural networks. 13, 22, 23, 26, 29

concept drift See **conceptual drift** . 7, 21, 28

conceptual drift A term from data stream mining, [69, 24], 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 . 21, 28, 30

contamination Food contamination is generally defined as foods that are spoiled or tainted because they either contain microorganisms, such as bacteria or parasites, or toxic substances that make them unfit for consumption. A food contaminant can be biological, chemical or physical, 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 [27] . 2, 4–8, 10, 13, 15–17, 21, 22, 28, 33, 36, 47, 48

cross-species Cross-species refers to a form of contamination, where two species are mixed together, e.g. a sample with both Hoki and Mackerel. In the mass spectrometry datasets, these species are mixed thoroughly in a blender to give a homogeneous sample with a maximum blend of the two species . 5–9, 16–18, 21, 22, 33, 47

cross-validation For *k*-fold cross-validation, the method divides the data into *k* folds such that the proportions of the classes in each fold are representative of the proportions in the whole dataset. Each fold plays the testing role, while the remaining (*k*-1) folds are combined to form a training set . 35, 38, 43

Cyber-marine Cyber Physical Seafood Systems (Cyber-Marine) is a new multi-million dollar research programme aimed at achieving 100% utilisation and maximised value for all harvested wild and aquacultured seafood. Making use of all raw material will allow the industry to achieve growth targets without increasing catch volume from wild-capture fisheries as well as maximise value from increasing aquaculture. Once established for the seafood industry, the technology could be adapted for any bio-industrial process [7] . 2

DDIM denoising diffusion implicit models. 26

DDPM denoising diffusion probabilistic models. 26

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 . 7, 10, 15, 21, 36, 47, 48

DIMS direct infusion mass spectrometry. 3, 14, 18, 21, 46, 47, 51

DL deep learning. 5

domain knowledge Knowledge related to the application domain. For example, biochemistry and fish processing . 29

EC evolutionary computation. 17, 23, 24

EDA exploratory data analysis. 31, 35, 48, 50

FC feature construction. 43

FS feature selection. 43, 44

GAN generative adversarial networks. 8–10, 16, 18, 26, 27, 29, 36

gas chromatogram Gas Chromatography for fatty acid analysis in [18]. 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 . 42

GC gas-chromatography. 46

GC-MS gas-chromatography mass-spectrometry. 21, 22, 24, 25, 31, 38, 43, 47, 51

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 . 24

GP genetic programming. 17, 21, 26, 38, 39, 43, 44

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

homogeneous This term is used heavily in chemistry. In the context of chemistry homogeneous means the same, or having a similar structure. In fish processing, the fish tissue samples are taken from a homogeneous blend of marine biomass. Also, in the Hoki season, the input to the flex-factory is predominantly one species, this may also be referred to as homogeneous. The marine biomass of Canada or the United States, the incoming catch from trawling vessels, is homogeneous, as it consists of mostly one (or few) species - a narrow range of marine biomass . 5, 6, 14

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

identification Different to detection, identification involves detecting the presence of phenomena in a sample and then specifying what the phenomena were. E.g., an identification system can find **cross-species** contamination and identify both species in the contamination . 4, 10, 14, 21, 22, 24, 31, 33, 35, 36, 38, 47, 48, 50

instance identification In computer vision, this is referred to as instance identification [38], not to be confused with **instance segmentation** [37]. Instance identification is the task of identifying unique instances in a photograph. Take for example a photograph with 5 sheep. Instance identification would identify each of the individual sheep, as a unique individual. In the context of fish processing, instance identification correctly assigns the origin of a sample, which unique individual fish it originated from. In chemistry, we refer to this as **sample attribution**, and for this proposal, treat the terms as interchangeable . 11, 54

instance recognition The machine learning term for recognizing individuals that may belong to the same class is ”instance recognition” [36], or ”individual recognition”. For fish processing, instance recognition would involve recognizing each individual fish in the samples and assigning a unique identifier or label to each of them. This would allow the model to differentiate between individual fish even if they belong to the same species. Instance recognition is a type of object recognition task that goes beyond simply recognizing object classes and aims to identify each individual instance of an object class. It is commonly used in various fields such as wildlife monitoring, security surveillance, and biometrics . 10–12, 18, 22

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 . 34

KL Kullback-Leibler. 10, 26

KL K-nearest neighbours. 25

LLM large language model. 7

LNBNN local naive bayes nearest neighbours. 12

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 . 1–7, 10, 11, 20, 21, 27, 28, 46, 47, 51

mass The amount of matter in an object . 33

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 . 33

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, and the intensity peak in a **mass spectrum** represents the number of ions with a particular mass-to-charge ratio that is detected by the mass spectrometer . 31, 34

MCIFC multiple class-independent feature construction method. 38–44

ML machine learning. 28, 29

MLM masked language modelling. 27

MO mineral Ool. 7, 16, 17, 33

MRMR maximum relevance - minimum redundancy. 27, 44

MS mass-spectrometry. 3, 14, 15, 18, 21, 22, 24–27, 29, 33, 34, 46, 51

MT-GP multi-tree genetic programming. 24, 38, 39, 41–44

NBNN naive bayes nearest neighbours. 12

NLP nautral language processing. 27

NSP next sentence prediction. 27

offline see **online** . 28, 46, 47

offline learning The dataset is static, the system is not constantly trained on new data. It is trained once, and then deployed as is. Offline learning s the opposite of **online learning**. Not to be confused with **offline** . 28

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** . 13, 46

online learning Online learning refers to a model that can be updated and adapt to new instances after its initial training. Take for example the Tesla FSD training programme. The FSD edge cases are referred to as the long tail of computer vision. These edge cases are where the car demonstrates undesirable behaviour, e.g. a crash, swerve, unsafe/irregular driving, are sent back to the DOJO computing facility, and the model is retrained via Monte-Carlo simulation of that edge case, to perform the desired behaviour. This human-in-the-loop online learning is a powerful method to bootstrap algorithms for robustness. Not to be confused with **online** Online learning is the opposite of **offline learning** . 25, 28, 30

part A fish part refers to which tissue of the fish body the sample was taken from. The fish parts considered in this research include fillet, frames, gonads, head, liver & skin . 14, 15, 21, 24, 25, 33

PCA principal component analysis. 9, 25, 35, 36, 38

PCA-LDA principal component analysis - linear discriminant analysis. 8, 9, 22, 35

PFR Plant and Food Research New Zealand Ltd.. 17, 36

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, or one-hot encoded vector . 24

PSO particle swarm optimisation. 24, 44

QA quality assurance. 15

QC quality control. 2, 21, 29, 33

QCA qualitative contaminant analysis. 15, 47, 48, 50

quantification Quantification assesses how much a sample is contaminated. Take for example **cross-species contamination**, **quantification** is interested in the percentage of contaminants from each species, e.g. 70% Hoki, and 30% Mackerel . 6, 8, 15, 16, 18, 47, 48

R-CNN region-based convolutional neural networks. 27

recall Recall is a metric for classification accuracy. It measures the proportion of actual positives that were correctly identified. It can be thought of as $\frac{TP}{TP+FN}$, where TP is true positives, and FN is false negatives. **contamination detection** requires a high precision. There is lenience for false positives, a flex-factory should catch all samples that are truly contaminated. Closely related, and not to be confused with **precision** . 15

REIMS rapid evaporative ionisation mass spectrometry. 3–8, 10–14, 18, 21, 22, 25, 27, 31, 33–36, 38, 46, 47, 51

RSD relative standard deviation. 25, 29, 33

sample attribution Sample attribution is a chemistry term, that refers to identifying which individual sample a measurement was taken from. In **mass spectrometry**, several measurements are taken from the same fish **tissue** sample. Being able to identify measurements from a common origin, i.e. the same sample, is important for **traceability**. This can be used to isolate contaminated samples or deduce a sample's path through the factory . 11, 16, 17, 47, 48

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 [3] . 6, 7, 17, 18, 20, 21, 28, 30

significant markers Significant Markers (or important variables) are ions that are unique to a specific offal cut, and present in all samples [15] . 29

SOTA state-of-the-art. 29

spawning Spawning is the reproductive process in which marine biomass releases their eggs and sperm into the water. This is important for producing new offspring. The spawning of [3] is of particular interest, as it causes **seasonal variation** . 6, 7, 20, 21

species This refers to the species of fish that the tissue sample belongs to. The fish species in this research are Hoki and Mackerel. The species considered in previous work [18] were Bluecod, Gurnard, Snapper & Tarakihi. For differentiating between distinct species in fish fraud detection see [13]. See [75] for the biological definition from Darwin . 14, 15, 21, 24, 25, 33, 35

spoilage Spoilage in a fish processing context refers to the decay or deterioration of fish or seafood products, resulting in a loss of quality and edibility. Fish spoilage can occur due to various factors such as bacterial growth, enzymatic activity, oxidation, and physical damage during handling, transportation, or storage . 2, 8, 15

ST-GP single-tree genetic programming. 24, 38–40, 42, 43

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 . 43

SVM Support Vector Machine. 27, 38, 44

t-SNE T-distributed stochastic neighbour embedding. 26, 35–38

taxonomy A taxonomy is a hierarchical classification system that organizes a set of concepts or subjects into categories and subcategories based on shared characteristics. Taxonomies are often used in fields such as biology, where they are used to classify and organize living organisms into a systematic hierarchy based on their characteristics and evolutionary relationships. They are also used in other fields, such as information science and library science, to classify and organize knowledge in a way that is easy to understand and navigate . 25, 28, 29

tissue See part . 3, 14, 33, 42, 43, 46, 51

traceability Traceability is a term from quality assurance, which is important in a factory setting. Should a problem arise, a factory needs to be able to isolate and determine the origin and potential causes for that problem. Take for example fish tissue contaminated by mineral oil. After detecting said **contamination**, traceability would be concerned with identifying other tissue samples from the same fish that are likely contaminated . 4, 10–12, 22, 28, 47, 48, 50, 51

transfer learning Transfer learning is a machine learning technique where shared knowledge is transferred between related tasks. Take for example, the source task of riding a bike, and the target task of riding a motorcycle. Although the tasks are different, there 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, 12, 18, 25, 27, 29

UMAP uniform manifold approximation and projection for dimension Reduction. 26, 35–38

Bibliography

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

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

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

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

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

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

[7] 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.

[8] J. Premanandh, "Horse meat scandal–a wake-up call for regulatory authorities," *Food control*, vol. 34, no. 2, pp. 568–569, 2013.

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

[10] Y.-H. P. Hsieh and J. A. Ofori, "Detection of horse meat contamination in raw and heat-processed meat products," *Journal of agricultural and food chemistry*, vol. 62, no. 52, pp. 12536–12544, 2014.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

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

[18] 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.

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

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

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

[22] C. Metz, “The godfather of a.i.’ leaves google and warns of danger ahead,” *The New York Times*, 2023.

[23] 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.

[24] 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.

[25] 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.

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

[27] M. A. Hussain, “Food contamination: major challenges of the future,” 2016.

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

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

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

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

[32] 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.

[33] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[34] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[52] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

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

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

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

[56] 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.

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

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

[59] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," *arXiv preprint arXiv:1611.01578*, 2016.

[60] C. White, M. Safari, R. Sukthanker, B. Ru, T. Elsken, A. Zela, D. Dey, and F. Hutter, "Neural architecture search: Insights from 1000 papers," *arXiv preprint arXiv:2301.08727*, 2023.

[61] J. Huang, B. Xue, Y. Sun, and M. Zhang, "Ede-nas: An eclectic differential evolution approach to single-path neural architecture search," in *AI 2022: Advances in Artificial Intelligence: 35th Australasian Joint Conference, AI 2022, Perth, WA, Australia, December 5–8, 2022, Proceedings*, pp. 116–130, Springer International Publishing Cham, 2022.

[62] G. Yuan, B. Xue, and M. Zhang, "A two-stage efficient evolutionary neural architecture search method for image classification," in *PRICAI 2021: Trends in Artificial Intelligence: 18th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2021, Hanoi, Vietnam, November 8–12, 2021, Proceedings, Part I 18*, pp. 469–484, Springer International Publishing, 2021.

[63] S. Li, Y. Sun, G. G. Yen, and M. Zhang, "Automatic design of convolutional neural network architectures under resource constraints," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[64] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A survey on evolutionary computation approaches to feature selection," *IEEE Transactions on evolutionary computation*, vol. 20, no. 4, pp. 606–626, 2015.

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

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

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

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

[69] 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.

[70] 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.

[71] 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.

[72] 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.

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

[74] S. Linnainmaa, *The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors*. PhD thesis, Master's Thesis (in Finnish), Univ. Helsinki, 1970.

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

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

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

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

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] 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.

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

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

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

[87] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," *Advances in neural information processing systems*, vol. 31, 2018.

[88] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 26565–26577, 2022.

[89] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," *arXiv preprint arXiv:2302.05543*, 2023.

[90] 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.

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

[92] C. A. Smith, G. O'Maille, E. J. Want, C. Qin, S. A. Trauger, T. R. Brandon, D. E. Custodio, R. Abagyan, and G. Siuzdak, "Metlin: a metabolite mass spectral database," *Therapeutic drug monitoring*, vol. 27, no. 6, pp. 747–751, 2005.

[93] V. B. O'Donnell, E. A. Dennis, M. J. Wakelam, and S. Subramaniam, "Lipid maps: Serving the next generation of lipid researchers with tools, resources, data, and training," *Science signaling*, vol. 12, no. 563, p. eaaw2964, 2019.

[94] 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.

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

[96] 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.

[97] 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.

[98] 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.

[99] 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.

[100] S. C. Hoi, D. Sahoo, J. Lu, and P. Zhao, "Online learning: A comprehensive survey," *Neurocomputing*, vol. 459, pp. 249–289, 2021.

[101] 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.

[102] 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.

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

[104] 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.

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

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

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

[108] W. R. Smart, "Genetic programming for multiclass object classification," *BSc (Honours) Research Project*, 2005.

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

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

[111] 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.

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

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

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