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

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

Abstract

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

1. Introduction

This proposal is about fish analysis - rapid determination of bulk composition and quality of marine biomass in Mass Spectrometry. Specifically, we aim to identify the type of fish and assess its suitability for use in fish products. In this section, we introduce the global fishing industry, fish processing in New Zealand, the potential for automation, and a review of the current state-of-the-art in this field.

This research focuses on improving waste utilization in the global fishing industry. According to [1], approximately 100 million tonnes of wild fish are captured each year, and only about 40% of these fish are processed into edible parts. The remaining portions are often processed into fish oil and fish meal, or discarded as non-fillet material. In addition, many fisheries are in decline and global fishing has not significantly increased in the past 30 years, making waste utilisation an important focus globally. We must maximize the utilization and value of every kilogram of marine biomass to preserve our fish stocks and ensure there are plenty of fish in the sea for future generations to reel in.

The many steps in the supply chain from ocean to plate, are prone to human error and criminal activity. Consider the 2013 European Horse Meat Scandal. Adulteration watered down high-value beef mince products with low-value horse meat, and sold them to an unaware public, as a criminal enterprise to increase profits. The beef with adulteration applies to the global fishing industry. According to [2] a meta-analysis comprised of 51 studies of the global fishing industry, there was an average mislabelling rate of 30%. We want to be confident we know what are eating, we must ensure the labels on seafood products are accurate. We need tools for quality assurance that can determine the composition and quality of fish products.

The New Zealand fishing industry prides itself on sustainability. New Zealand fisheries are well-regulated with strict quotas for over 100 marine species [3]. The NZ fishing industry does not have many 'high volume' fisheries, e.g. Hoki our largest fishery, as approximately 110,010 tonnes of quota each year [4]. On a global scale, this is minuscule, Norway alone have an aquaculture production of salmon of 4,000,000 tonnes a year [5]. This makes it difficult for fish processing, due to the variability in the catches, different boatloads of fish require different processing to maximize their value. The MBIE CyberMarine programme 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.

We aim to employ machine learning techniques to detect spoilage indicators, Quality Control, and contamination (ideally) on fresh marine biomass. We need tools for quality control in fish processing. 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. We are 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, we seek to find 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). We seek to develop tools that can identify contamination in marine biomass. We need these techniques to work on fresh (uncooked) marine biomass, as cooking the fish can destroy valuable proteins, collagen and active enzymes. Cooking is also energy-intensive and time-

consuming, it adds time and cost to fish processing, so processing fresh marine biomass is preferred.

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

2. Literature Review

This project aims to implement a real-time fish contamination detection and identification algorithm. This is a supervised machine learning task operating on Rapid Evaporative Ionisation Mass Spectrometry (REIMS) [6] fish oil data. Types of contamination include cross-species and mineral oil.

2.1 State-of-the-art Chemistry

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

1. **Rapid Evaporative Ionisation Mass Spectrometry (REIMS)** [6]
2. **Direct Infusion Mass Spectrometry (DIMS)**

These are two of the most powerful analytical tools for Mass-Spectrometry. These tools are very expensive, but as prices decrease they may be affordable for deployment in a marine biomass processing facility. REIMS [6] has shown promise in beef processing, where it was able to detect horse meat contamination in beef [7]. Most impressively, horse meat contamination was detected at ¡INSERT STATISTICS FROM PAPER HERE¿ very low levels. This demonstrates the REIMS technique is incredibly sensitive to contamination. REIMS has been applied to fish fraud detection to identify fish species and identify catch methods for fish products. The method was so accurate it was able to identify incorrectly labelled instances in the training data. However, it has not been applied to Adulteration detection and identification in marine biomass. In this work, we apply REIMS to speciation, cross-species / mineral oil contamination, identify QC parameters, and individual identification. We 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 [8, 9].
- **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 [2]. 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) [10].

- **Gas-Chromatography Mass-Spectrometry** - Previous work [11] demonstrated that Gas-Chromatography Mass-Spectrometry (GC-MS) can identify fish species with high accuracy. However, GC-MS techniques significant time and domain expertise is required to prepare and analyze samples. This is not applicable for real-time fish contamination detection.

2.2 State-of-the-art Machine Learning

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

In [10], 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 [10], each sample belonged to one species, however, for this problem, each sample can belong to multiple classes, e.g. a mix-species contaminated sample contains a mixture of two species. [7] performed detection and identification beef adulteration. It can identify samples that are adultered with offal, and specify which offal was present.

2.3 Limitations

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

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

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

2.4 Domain Knowledge

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

- Manual hyperparameter tuning (e.g. # principal components, threshold for outliers, mass range) can be automatically selected, or replaced by models that don't need them at all!

2.5 State-of-the-art ML

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

- Basic dimensionality reduction techniques (e.g. PCA [12]) were used.
  - PCA [12] Project data along the principal components, the axis of maximum variance in descending order.
  - The first principal component is the axis of maximum variance, the second principal component is orthogonal to the first and has the second largest variance, and so on.
  - This method does not take into consideration feature interactions, interactions with the class labels, and feature redundancy/relevance.
  - Future work should consider t-SNE [13], UMAP [14]
    - t-SNE [13]
      1. it creates a probability distribution of the similarity between points in the high-dimensional space.
      2. it defines a similar probability distribution over points in the low dimensional space.
      3. Then minimizes the Kullback-Leibler (KL) divergence [15] between the two distributions.
- Basic supervised statistical models (e.g. LDA, OPLS-DA) was used for classification. Future work should consider CNNs [16, 17], GANs [18], Diffusion [19, 20]
  - Denoising Diffusion Probabilistic Models (DDPM) [19], the original diffusion paper, behind diffusion-based image generation models.
  - Denoising Diffusion Implicit Models (DDIM) [20], a generalized DDPM that is faster and deterministic.
  - Genetic Programming for classification [21], feature construction [22, 23], feature selection

2.6 Transfer Learning

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

- Potential for transfer learning (incorporate previously existing data) to improve performance for few-shot classification tasks.
  - Due to manual labour, cost of machinery, domain expertise and high-resolution datasets, REIMS datasets have low sample complexity and high dimensionality.
  - Unsupervised learning techniques have utilized unlabelled data from the same distribution to improve classification accuracy. The REIMS dataset contains Quality Control (QC) samples. These don't belong to any class (?) and are used to calibrate/tune the machine, unlabelled instances drawn from the same distribution. Zemma et al. [24] incorporated unlabelled instances to draw more accurate support vectors and improve the classification accuracy for breast cancer diagnosis with SVM.

  - METLIN metabolites database, and LIPID MAPS can provide annotated labels for spectra [7].
  - This looks like that (R-CNN) [25], give annotated labels for lipids used to make a classification/regression decision (significant markers $\approx$ important features).

2.7 Taxonomy

The terminology used to describe their methodology with chemistry/statistics jargon. A clear explanation of the equivalent terms between chemistry/statistics/Machine Learning terminology would open the field to further multi-disciplinary input from ML researchers.

- Significant markers [10, 7]
- Outliers [10, 7]
- Standard deviation threshold [10, 7]

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 section, we focus on that preliminary work. We will discuss classification and feature selection techniques that were applied to other fish chemistry datasets; these include support vector machines, Particle Swarm Optimisation, Convolutional Neural Networks, and Genetic Programming. At the end of this section, we provide exploratory data analysis on a new fish chemistry dataset, Rapid Evaporative Ionisation Mass Spectrometry (REIMS), and discuss how the preliminary work can and cannot, be applied to the new dataset.

In particular, the preliminary works presented in this proposal are:

- Automated Fish Classification on GC-MS data.
- CNN for Fish classification on GC-MS data.
- Genetic Programming (GP) for GC-MS data
  - Single-Tree Genetic Programming (ST-GP)
  - Multi-Tree Genetic Programming (MT-GP)
  - Multiple Class-independent Feature Construction Method (MCIFC)
- REIMS Exploratory Data Analysis (EDA)

3.1 Automated Fish Classification on GC-MS data

In the preliminary work section, we first introduce my previous research [11], which 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 speciation and body partsidentification. This proposal also looks to find machine learning techniques for fish speciation, but now instead on state-of-the-art Mass-Spectrometry techniques. Should you be interested in Gas-Chromatography Mass-Spectrometry (GC-MS), speciation or body partsidentification, I would recommend this paper, [11], as supplementary reading material, to avoid repetition, I will not repeat the contents of that paper here.

3.2 Genetic Programming for GC-MS data

In the Genetic Programming (GP) subsection of the preliminary work, we benchmark three GP methods, to my previous work, [11], that was addressed in the last subsection. In particular, the three GP methods proposed in this work are:

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

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

Algorithm 1 GP-based multiple feature construction

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

3.2.1 Representation

Multiple Class-independent Feature Construction Method (MCIFC) [23]. is a Multi-tree GP that constructs a smaller number of high-level features, proportional to the number of classes, from the original features. This method is based on the intuition that problems with more classes are likely to be more complex, and thus require more features to capture said complexity. The number of constructed features m, determined by $m = r \times c$, where r is the construction ratio (set to 2), and c is the number of classes. MCIFC constructs 8 features for the 4-class fish species problem and 12 features for the 6-class fish species problem.

3.2.2 Crossover and Mutation

MCIFC limits both the crossover and mutation operators to only one of the constructed features described in Algorithm 2. This approach favours exploitation over exploration, making small random changes to constructed features with monotonically increasing fitness due to elitism.

Algorithm 2 MCIFC Crossover and Mutation.

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

Table 1: 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%

3.2.3 Fitness

MCIFC takes the balanced classification accuracy of an SVM classifier as the fitness function. The SVM classifier is known to be effective for fish oil data [11]. Balanced accuracy avoids results bias towards the majority class, which is relevant for the fish species dataset, with the majority class 44% of samples belonging to fish species blue cod. The balanced accuracy is given by

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

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

3.2.4 Experimental Setup

Table 1 shows the datasets used in the experiments and their respective characteristics. Due to the high dimensionality of gas chromatography data, this paper employs a GP-based FC approach. The dataset is suited towards dimensionality reduction, as previous work [11] demonstrated FS can improve classification accuracy. The small number of instances is due to the expensive and time-consuming nature of performing Gas Chromatography on fish tissue.

The data is pre-processed to fix the instrumental drift by imputing missing timestamps with zero filling. Features are normalized in the range [0,1] based on the training set.

Table 2 describes the parameter settings of all GP-based methods used in the experiments. The function set has standard arithmetic operators $+, -, \times$, a protected division operator that prevents division by zero returning 0 instead, and the unary *neg* operator reverses the sign. The feature set, and randomly generated constant $r \in [-1, 1]$, are used in

Table 2: Paramter settings.

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

Table 3: Results

Dataset	Method	Train	Test
Species	KNN [26]	83.57	74.88
	RF [27]	100.0	85.65
	DT [28]	100.0	76.98
	NB [29]	79.54	75.27
	SVM [30]	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	–	–
	MCIFC	–	–

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.

3.2.5 Results

Table 3 compares the classification results from [11], to the ST-GP, MT-GP, and MCIFC methods proposed in this preliminary work. We 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. We average over 30 runs to ensure results are statistically significant due to the stochastic nature of population-based Genetic Programming.

3.2.6 Discussion

The balanced classification accuracy of 86.73% is a significant improvement on the previous Single-Tree and Multi-tree GP classification, whose best results both could not exceed 55% accuracy on the training set. Both these approaches fail to capture the complexity of the fish species dataset. For Single-tree GP, this is likely due to the difficulty of the GP finding a single expression that can fit the class boundaries for the classification map. For Mulit-tree GP using the one-vs-rest approach, we see improvement over Single-Tree GP, but still not as good as MCIFC.

The Multi-tree GP has to learn how to perform accurate classification, which is a difficult task. MCIFC is a wrapper-based method, which plays to the strengths of GP for feature construction, and SVM for classification tasks.

The Linear SVM classifier has a balanced classification accuracy of 92.4% on the test set, which is a significant improvement on the previous best result of 86.73% using MCIFC. However, after tweaking the parameters of MCIFC we expect to match (or exceed) the SVM classifier in the future.

(*Note:* The code needs to be adjusted to record train and test accuracy for each individual to get more comprehensive results, but these initial results were a proof of concept.)

3.3 REIMS Exploratory Data Analysis

3.3.1 Annotated Labels

Figure 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. We separate this dataset into four sub-datasets to address four sub-tasks: speciation, cross-species, engine oil, and individual.

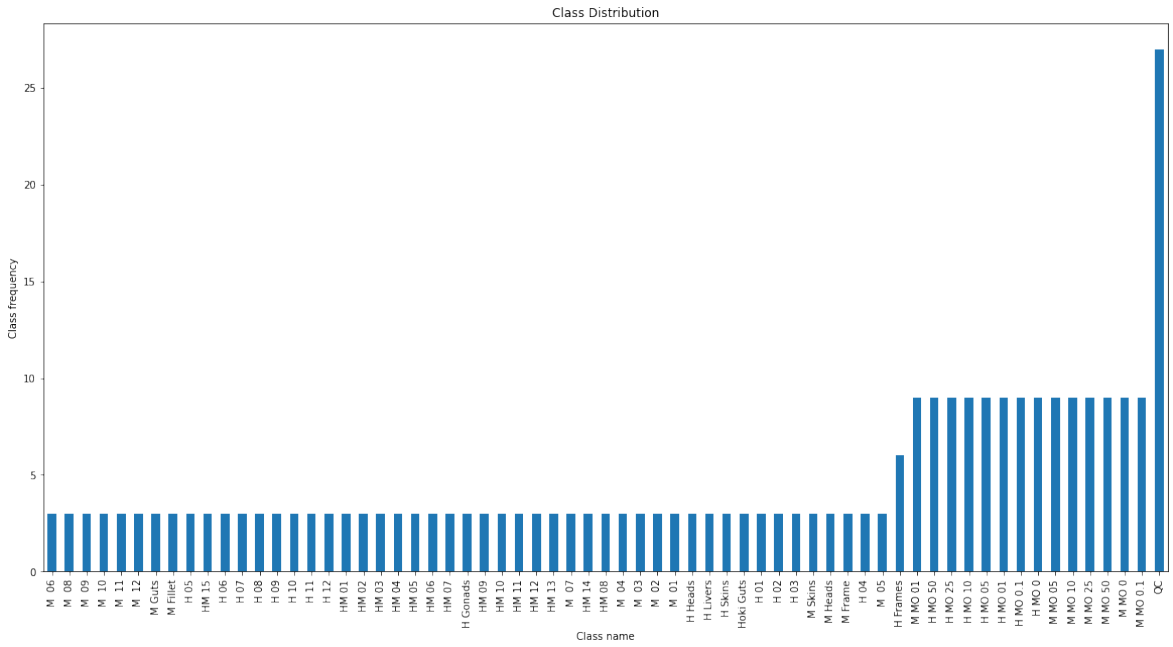


Figure 1: Class Distribution

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

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

Proceeding with the species tag, there is either a number - the individual, fish part - where the sample was taken from, or Machine Oil (MO).

Machine Oil (MO) - The first two are self-explanatory, but for MO, these annotations contain a decimal afterwards. Take, for example, "M MO 0.1", this represents a Mackeal species, contaminated with Machine Oil, at a contamination rate of 1%. The Machine Oil contamination rates $\in$ [0.1%, 1%, 10%, 25%, 50%]. Samples are contaminated at different rates because we 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 Mackeral, 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. 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.

3.3.2 Speciation Task

Table 4 gives the results for preliminary experiments, exploring the performance of different dimensionality reduction techniques and classification algorithms on the REIMS dataset. In these preliminary experiments, the classification task is speciation, please see subsection 4.3 Speciation, for more information. We give the mean and standard deviation classification accuracy over 10-fold cross-validation. The best-performing reduction method and classification, and respective classification accuracy, are in bold.

Method	SVC [30]	KNN [26]	DT [29]	RF [27]	XGBoost [31]	**LDA**
PCA [12]	0.88 ± 0.17	0.85 ± 0.13	0.83 ± 0.15	0.87 ± 0.13	0.88 ± 0.14	$\mathbf{0.92 \pm 0.13}$
t-SNE [13]	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 [14]	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 4: Feature Reduction / Classification Methods for Speciation

The table shows PCA-LDA (**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, [10, 7] 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.

4. Contributions

This research aims to evaluate two state-of-the-art Mass-Spectrometry techniques on their ability to rapidly determine bulk composition and quality of marine biomass. 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. In this section, we define those techniques and sub-tasks, and then explore each in further detail.

4.1 Mass Spectrometry: State-of-the-art

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

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

There exists an age-old trade-off between speed and quality, told in the fable of the Tortoise and the Hare. These two datasets demonstrate this trade-off - REIMS is fast but low-resolution, DIMS is slow but high-resolution. Work from [10] 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 [32]. For deployment in a factory setting, speed is a must. We want rapid results that match the pace of the production line. However, we 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 [10], REIMS results were worse on cooked biomass. Studies [10, 7] show that Mass-Spectrometry works on raw biomass products.

4.2 Marine Biomass: Composition and Quality

We have two datasets that describe marine biomass, each with trade-offs - inherent strengths and weaknesses. Now we need sub-tasks related to fish processing, 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. Speciation
2. Cross-species contamination
3. Mineral Oil contamination
4. Individual identification

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

4.3 Speciation

Speciation [33] - can REIMS / DIMS data be used to classify different species tissues? What variables are responsible?

- Same task as [11], but instead of GC-MS, this is REIMS and DIMS
- Classification
- Feature Importance - Interpretable,
  - similar to significant markers from [10, 7]
  - and interpretability from [11, 25].

4.4 Cross-species Contamination

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

- Similar to [10], but instead of beef-horse, this is for fish contamination.
- few-shot learning (very few training instances)
  - transfer learning, active learning or zero-shot inference may be needed.
- Detection $\approx$ Multi-label classification
- Identification $\approx$ multi-output regression
  - find anomalous instances!
  - Identify the percentage of cross-species contamination.
  - Potentially, even those outside of annotated labels.
- Feature importance (again) - significant markers
  - profile - how much contamination? confidence?

4.5 Mineral oil contamination

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

- Marine biomass can be contaminated with several things, for example, plastics and mineral oil - which is carcinogenic (it kills). We seek to develop tools that can identify contamination in marine biomass.
- Detection $\approx$ classification
- Identification $\approx$ multi-output regression/classification, i.e. identify true/false oil contaminated, and what percentage is oil?
- Feature importance (again x2) - significant markers
  - profile - how much engine oil? dangerous? confidence?

4.6 Identification

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

- Identification
- Feature importance (again x3) - significant markers
  - profile - species? part? confidence?

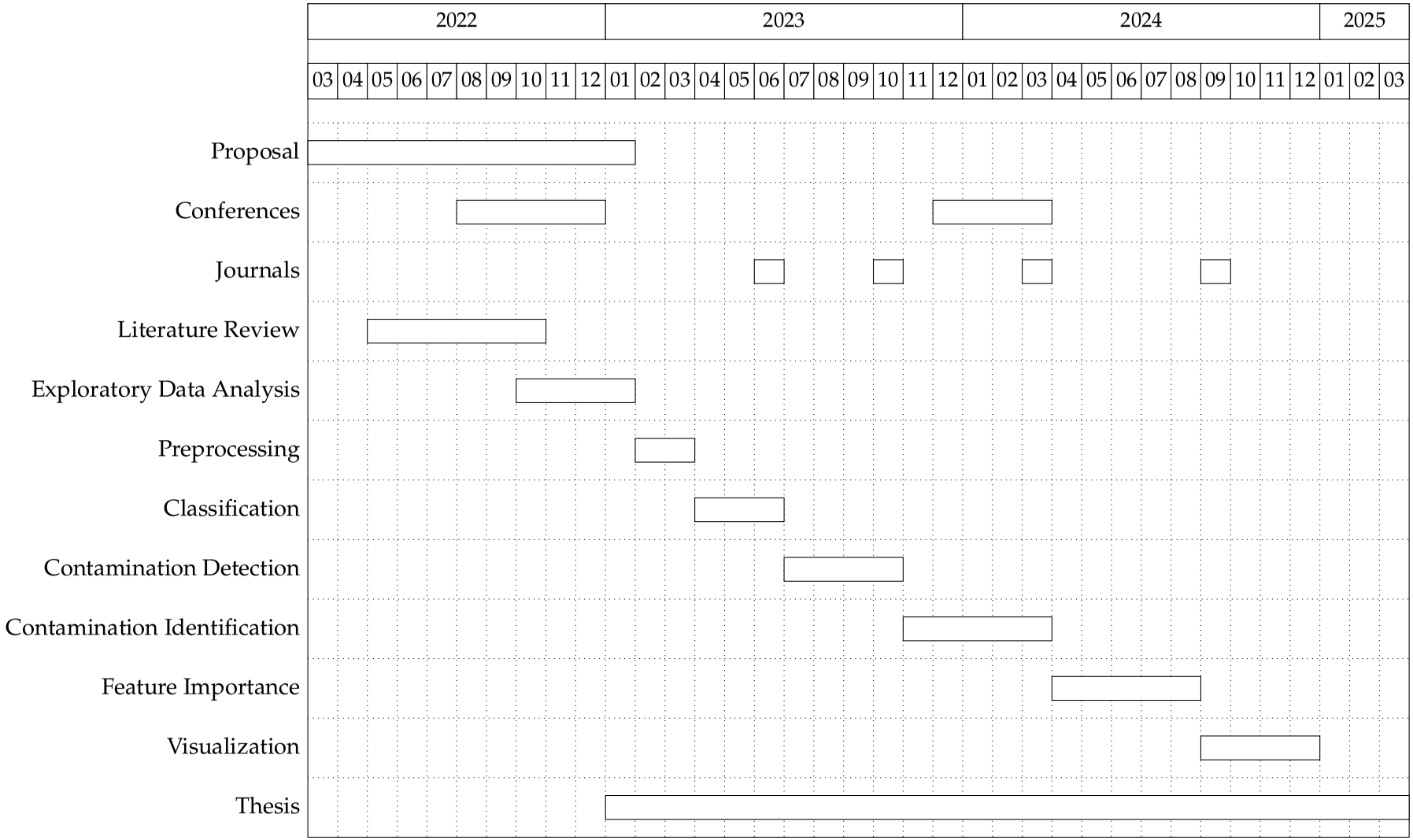
5. Milestones

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

1. Proposal
2. Conferences (x2)
3. Journals (x4)
4. Literature Review
5. Exploratory Data Analysis
6. Preprocessing
7. Classification
8. Contamination Detection
9. Contaminiaiton Indentification
10. Feature Importance
11. Visualization
12. Thesis

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

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


2022
2023
2024
2025
03 04 05 06 07 08 09 10 11 12 01 02 03 04 05 06 07 08 09 10 11 12 01 02 03 04 05 06 07 08 09 10 11 12 01 02 03
Proposal
Conferences
Journals
Literature Review
Exploratory Data Analysis
Preprocessing
Classification
Contamination Detection
Contamination Identification
Feature Importance
Visualization
Thesis

6. 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, we outline the following structure.

1. Introduction
2. Background
   (a) Mass-Spectrometry
   (b) REIMS / DIMS
   (c) Detection / identification
   (d) Interpretable ML
3. Preparations
   (a) Exploratory Data Analysis
   (b) Preprocessing
4. Applications
   (a) Speciation
   (b) Cross-species Contamination
   (c) Mineral oil Contamination
   (d) Individual identification
5. Discussion
6. Conclusion
7. Appendix
   (a) Taxonomy
   (b) Glossary

In this thesis, we outline the various steps and techniques that will be employed in this process, including the use of Mass-Spectrometry techniques such as REIMS/DIMS and the application of interpretable machine learning for detection and identification. We also describe the necessary preparations, including Exploratory Data Analysis and preprocessing, and the specific applications of this method, including fish speciation, cross-species contamination detection, and individual identification. Finally, we discuss the results of our research and provide a conclusion.

The appendix includes a taxonomy and glossary to bridge the multi-disciplinary gap in knowledge. The majority of readers will only know one of those disciplines. The glossary provides a quick point of reference for jargon, to reduce the cognitive load for the read. A taxonomy - this will break down the terminology from a chemistry/biology, fish processing and machine learning perspective. This addresses an important gap in the existing literature, where current papers, [10, 7] rely heavily on jargon from chemistry and statistics, where synonyms or equivalent terms in machine learning exist. Removing the barrier of jargon between disciplines will make it easier for multi-disciplinary future work, making the field more accessible to machine learning researchers.

7. Resources

Table 5: Resources

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

To effectively conduct this research, we will be utilizing a variety of resources. In this section, we break those down into hardware, software, and human resources. Table 5 gives a high-level view of those resources. For the remainder of this section, we address each of those resources in further detail.

7.1 Software

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

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

7.2 Hardware

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

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

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

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

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

7.3 Human Resources

In addition to these resources, I have also gained valuable experience through previous field trips to NZ Plant and Food Research, where I saw GC-MS first-hand for my previous publication [11]. This trip gave insights into steps in the ocean-to-plate supply chain, as their research laboratory processed whole fish into fish oil tissue samples suitable for Mass-Spectrometry techniques. With another trip to the Nelson-based Plant and Food Research, I could see DIMS in person. Lastly, it would be invaluable to plan a trip to the Wellington-based Callaghan Innovation, to see the REIMS in person.

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 [34] . 1–3

CNN Convolutional Neural Networks. 4, 5

contamination Food contamination is generally defined as foods that are spoiled or tainted because they either contain microorganisms, such as bacteria or parasites, or toxic substances that make them unfit for consumption. A food contaminant can be biological, chemical or physical in nature, with the former being more common. These contaminants have several routes throughout the supply chain (farm to fork) to enter and make a food product unfit for consumption [35] . 1–3, 11, 12, 14, 15

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. . 8

DDIM Denoising Diffusion Implicit Models. 4

DDPM Denoising Diffusion Probabilistic Models. 4

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 . 2, 12, 15

DIMS Direct Infusion Mass Spectrometry. 2, 11–13, 15, 17

EDA Exploratory Data Analysis. 5, 13–15

FC Feature Construction. 7

GC-MS Gas-Chromatography Mass-Spectrometry. 3, 5, 6, 8, 12, 17

GP Genetic Programming. 2, 5–8

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 . 2, 5, 11–13, 15

KL Kullback-Leibler. 4

MCIFC Multiple Class-independent Feature Construction Method. 5–8

ML Machine Learning. 5, 15

MO Machine Oil. 10

MS Mass-Spectrometry. 2, 4, 5, 11, 15, 17

MT-GP Multi-Tree Genetic Programming. 5, 6, 8

PCA-LDA Principal Component Analysis - Linear Discriminant Analysis. 3, 10

PSO Particle Swarm Optimisation. 5

QC Quality Control. 1, 2, 4, 10

REIMS Rapid Evaporative Ionisation Mass Spectrometry. 2–5, 9–13, 15, 17

significant markers Significant Markers (or important variables) are ions that are unique to a specific offal cut, and present in all samples [7] . 12, 13

speciation Darwin [33] Differentiating between distinct species [10] . 2, 5, 9–12, 15

spoilage TODO . 1

ST-GP Single-Tree Genetic Programming. 5, 6, 8

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 . 8

SVM Support Vector Machine. 7

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 . 15

Bibliography

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

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

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

[4] "Hoki macruronus novazelandiae," Oct 2021.

[5] "Fisheries and aquaculture in norway," Jan 2021.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

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

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

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

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

[16] 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.

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

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

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

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

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

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

[29] 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.

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

[31] 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.

[32] 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.

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

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

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