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 detection and identification of contamination within fish products using rapid ionisation evaporative 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

- Scope - place the problem in the world.
  - Approximately 100 million tonnes of wild fish are captured each year [1].
  - Edible parts of these fish account for only about 40%, with the remaining portion usually processed into fish oil and fish meal.
  - In some cases non-fillet material is used as fertilizer or even discarded.
  - Waste utilisation of marine biomass has become an important global focus. Many fisheries are in decline and global fishing has not increased significantly in 30 years (See Figure 1 in [1]).
  - It is critical that we maximise the utilization and value from every kilogram of marine biomass.
- Specifics to New Zealand, sustainability:
  - New Zealand has a well-regulated fishing industry, with strict quotas for over 100 marine species.
  - A major challenge for the NZ fishing industry is that we do not have many 'high volume' fisheries.
  - Hoki is our biggest, with approximately 150,000 t of quota each year.
  - This is miniscule on the global scale. For example, Norway aquaculture production of salmon is $>$1,000,000 t each year.
  - The other, smaller NZ fisheries mean that we have a highly variable biomass.
  - In [2] a 2016 meta-analysis of 51 studies with total (n=4,500) samples, found an average mislabelling rate of 30%. Showing that the many steps involved from ocean-to-plate in producing fish products is prone to error.
  - This makes things difficult for the fish processing industry, because the variability in raw material means that different 'boat laods' of fish require different processing to maximise 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.
- Fish processing - quality control, contaimination.
  - Marine biomass is highly prone to spoilage. This can be due to enzymatic spoilage as the proteases and lipases in the fish begin to digest the dead animal, microbial digestion, and also because of oxidation in air.
  - Marine lipids in particular are highly unsaturated, which makes them especially prone to oxidation. For these reasons it is important that marine biomass is carefully handled after being caught.
  - The list of quality control parameters for marine products can be extremely long, especially for marine oils.
  - Also, marine biomass can be contaminated with several things, including plastics and mineral oil, which can be carcinogenic.
  - While cooking stabilises the marine biomass, it destroys potentially valuable native proteins e.g. collagen and active enzymes.

  - Cooking the biomass is also energy intensive, so processing fresh biomass would be highly desirable – if possible.

- Fish processing - automation:

  - To automate optimal processing the composition of a biomass must be understood in real time.
  - This is very challenging.
  - To be understood, we require interpretable models that can be verified/troubleshoot by domain experts in biochemistry.

- Current state-of-the-art

  - 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).
  - Previous work [3] 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 too prepare and analyze samples. This is not applicable for real-time fish contamination detection.
  - 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 compute required to process it, rules out genomics data for real-time fish contamination detection.
  - However, arguably the most powerful analytical tool currently available is MS. These instruments are very expensive, but prices continue to decrease and it is possible that in the future affordable instruments could be deployed – for example – in a marine biomass processing facility.
  - Direct Infusion Mass Spectrometry (DIMS) involves creating ions at atmospheric pressure from solid samples, before ions are sucked into an MS detector for analysis.
  - Rapid Evaporative Ionisation Mass Spectrometry (REIMS) is one example of this, and has been used to detect horse meat contamination in beef [4]. This has even been done quantitatively to a very low level, which indicates how incredibly sensitive the technique is.
  - The technique has been applied fish fraud detection [5], to identify fish species, and identify catch method of fish products. This method was accurate enough to identify incorrectly labels for samples in the training data. It has not been applied to Adulteration detection and identification.
  - We're going to apply this to fish – to tell species apart, detect contamination and quantitative components of interest.
  - We're also going to compare data from this technique to lipidomics data generated using direct infusion of lipid extracts from the marine biomass samples. This approach is much less rapid because oils must first be extracted, but it will be interesting to explore the strengths and weaknesses of each technique.

2. Literature Review

The aim of this project is 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) fish oil data. Types of contamination include cross-species and mineral oil.

State-of-the-art

In [5], 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 [5], 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. [4] performed detection and identification beef adulteration. It can identify samples that are adultered with offal, and specify which offal was present.

Limitations

- 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!
- Basic dimensionality reduction techniques (e.g. PCA [6]) were used.
  - PCA [6] 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, feature redundancy/relevance.
  - Future work should consider t-SNE [7],
    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 [8] between the two distributions.
- Basic supervised statistical models (e.g. LDA, OPLS-DA) was used for classification. Future work should consider CNNs [9, 10], GANs [11], Diffusion [12, 13]
  - Denoising Diffusion Probabilistic Models (DDPM) [12], the original diffusion paper, behind diffusion based image generation model.
  - Denoising Diffusion Implicit Models (DDIM) [13], a generalized DDPM that is faster and deterministic.
- Potential for transfer learning (incorporate previously existing data) to improve performance for few-shot classification tasks.
  - METLIN metabolites database, and LIPID MAPS can provide annotated labels for spectra [4].

3. Prelimary Work

- Automated Fish Classification on GC-MS data.
- CNN for Fish classification on GC-MS data.
- Genetic Programming (GP) for GC-MS data
  - Single-Tree GP
  - Multi-tree GP
- REIMS exploratory data analysis

4. Contributions

- Each of these research questions will be applied to two state-of-the-art Mass-Spectrometry datasets:
  1. Rapid Evaporative Ionisation Mass Spectrometry (REIMS)
  2. Direct Infusion Mass Spectrometry (DIMS)

  These are the research questions from Plant and Food Research.

- **Speciation** - can REIMS / DIMS data be used to classify different species tissues? What variables are responsible?
  - Classification
  - Feature Importance - Interpretable
- **Cross-species contamination** - can REIMS / DIMS data detect mixed-species contamination in fish tissues? At what concentration? What variables are responsible?
  - Classification
  - Regression
  - Feature importance - Interpretable
- **Mineral oil contamination** Can REIMS / DIMS data detect mineral oil contamination in fish? At what concentration? What variables are responsible?
  - Classification
  - Regression
  - Feature importance - Interpretable
- **Individual identification** - can REIMS / DIMS data be used to distinguish between different fish individuals? What variables are responsible?
  - Identification
  - Feature importance - Interpretable

5. Milestones

1. Literature Review
2. EDA
3. Preprocessing
4. Classification
5. Contamination Detection
6. Contamination Identification
7. Significant markers / Feature importance
8. Visualization
9. Thesis

6. Thesis Outline

1. Introduction
2. Background
    - Mass Spectrometry
    - REIMS
    - Classifcation / Regression
    - Interpretable ML
3. Preparations
    - Exploratory Data Analysis
    - Preprocessing
4. Applications
    - Classification
    - Contamination Detection
    - Individual Identification
    - Auto ML
5. Discussion
6. Conclusion

7. Resources

In this section you will detail any resource requirements such as hardware, software or access to subjects.

- Hardware
  - ECS Grid Compute
  - Rapoi
  - Niwa HPC - via Auckland University
- Software
  - Repository - Github
  - Project Management - Github Projects
  - Programming language - Python
  - Documentation - Read the Docs
- Experience
  - Field-trip to Callaghan Innovation to see REIMS
  - Field-trip to NZ Plant and Food Research (if necessary for future datasets).

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 [14] . 2, 3

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 [15] . 2–5

DDIM Denoising Diffusion Implicit Models. 3

DDPM Denoising Diffusion Probabilistic Models. 3

detection Detection finds if something is hidden in a sample. It does not have specify what was hidden, only that sample had something hiding. E.g., it can detect some form of adulteration, cross-species contamination, mineral oil in a fish sample. . 2

DIMS Direct Infusion Mass Spectrometry. 2, 4

fish fraud Food fraud, simply put, is the selling of food products with a misleading label, description or promise [16] . 2

GC-MS Gas Chromatography - Mass Spectrometry. 2

identification Different to detection, identification involves detecting the presence of a phenomena in a sample, and then specifying what the phenomena was. E.g., an identification system can find cross-species contamination, and identify the both species in the contamination. . 2, 4

KL Kullback-Leibler. 3

lipidomics Lipidomics is the study of reaction pathways involved in lipid metabolism within biological systems. The lipidome consists of the lipid profile of a particular sample such as cell, tissue or organism, which can be integrated as a metabolome sub-set [17] . 2

MS Mass-Spectrometry. 2, 4

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

REIMS Rapid Evaporative Ionisation Mass Spectrometry. 2–6

significant markers Significant Markers (or important variables) are ions that are unique to a specific offal cut, and present in all samples [4]. . 5

speciation Differentiating between distinct species [5] . 4

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] e. a. Wood J, Nguyen B, "Automated fish classification using unprocessed fatty acid chromatographic data: A machine learning approach," *AI 2021: Advances in Artificial Intelligence*, vol. 35, no. 52, pp. 12536–12544, 2022.

[4] 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.

[5] 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.

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

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

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

[9] 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.

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

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

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

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

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

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

[16] "Tackling seafood fraud."

[17] E. Fahy, S. Subramaniam, R. C. Murphy, M. Nishijima, C. R. Raetz, T. Shimizu, F. Spener, G. van Meer, M. J. Wakelam, and E. A. Dennis, "Update of the lipid maps comprehensive classification system for lipids1," *Journal of lipid research*, vol. 50, pp. S9–S14, 2009.