A rapid machine-learning approach for detecting bulk composition and quality of marine biomass using rapid evaporative ionisation mass spectrometry

Jesse Wood[1], Bach Hoai Nguyen[1], Bing Xue[1], Mengjie Zhang[1], and Daniel Killeen[2]

[1] Victoria University of Wellington, Te Herenga Waka, Wellington, New Zealand
{jesse.wood, hoai.bach.nguyen, bing.xue, mengjie.zhang}@ecs.vuw.ac.nz
[2] New Zealand Institute for Plant and Food Research Limited, Nelson, New Zealand
daniel.killeen@plantandfood.co.nz

Abstract. Marine biomass composition analysis traditionally requires time-consuming processes and domain expertise. This study demonstrates the effectiveness of Rapid Evaporative Ionisation Mass Spectrometry (REIMS) combined with advanced machine learning techniques for rapid and accurate marine biomass composition determination. Using fish species, body parts, oil contamination, and cross-species contamination as model systems representing diverse biochemical profiles, the paper investigates various machine learning methods, including novel unsupervised pre-training strategies for transformers. The transformer-based model consistently outperformed traditional machine learning and other deep learning approaches across all tasks. The paper further explores the explainability of the best-performing and mostly black-box models using Local Interpretable Model-agnostic Explanations (LIME). REIMS analysis with machine learning proves to be a fast, accurate, and explainable technique for real-time marine biomass compositional analysis, with potential applications in marine-based industry quality control, product optimisation, and food safety monitoring.

Keywords: AI applications · explainable AI · machine learning · marine biomass · mass spectrometry · multidisciplinary AI

1 Introduction

Marine biomass composition analysis plays a crucial role in various industries, including food production, quality control, and environmental monitoring. Traditional approaches for analyzing marine biomass composition, such as Gas Chromatography Mass Spectroscopy [43], Nuclear Magnetic Resonance Spectroscopy [3], and Genomic Profiling [32], are often time-consuming, labour-intensive, and require significant domain expertise. In response to these challenges, Rapid Evaporative Ionisation Mass Spectrometry (REIMS) has emerged as a promising technique for rapid and accurate analysis of biological samples [4, 5, 20, 45].

However, REIMS data analysis faces several limitations. The rapid nature of REIMS necessitates equally rapid inference of its results, as traditional analytical chemistry techniques are too slow. Furthermore, current analytical methods for REIMS data often require domain expertise in chemistry and fish processing, which doesn't match the speed of rapid evaporation mass spectrometry. REIMS also produces high-dimensional data, with this particular dataset having 1023 mass-to-charge ratios as features, but there are limited training instances due to the time-consuming and expensive task of sample preparation. Additionally, for industry applications require fast, accurate, and interpretable models that can be verified and troubleshooted in real-world scenarios.

To address these challenges, this paper introduces several innovative approaches. This paper employs machine learning techniques that provide rapid inference and automation, eliminating the need for human-in-the-loop domain expertise in chemistry or fish processing. To handle the high-dimensionality of REIMS data, this paper utilizes deep learning [8,42] and evolutionary computation methods [40,41] that are well-suited for complex feature interactions in mass spectra. To mitigate the limited number of training instances, the paper implements data augmentation and unsupervised pretraining techniques. Finally, the paper employs Local Interpretable Model-agnostic Explanations (LIME) [30] to provide interpretable outputs that identify important features and quantify their impact, making our models more accessible to domain experts in chemistry and fish processing.

2 Datasets

Figure 1 gives the two wild-caught fish species - Hoki and Mackerel - that are the subject of this study. These are two important fish in New Zealand's seafood industry, especially given New Zealand's largest fishery is hoki [34]. For illustrative purposes the different fish body parts, which are shared across both species of fish, are given in fig. 2.

(a) Mackerel (b) Hoki

Fig. 1: Mackerel (left) Hoki (right) fish species

Following our introduction to the challenges and potential of REIMS-based analysis, this section now focuses on the critical foundation of our study: the

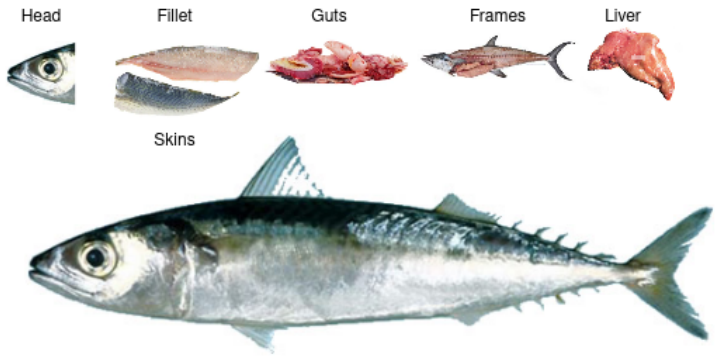


Fig. 2: Fish body parts

dataset. The dataset used in this study consists of REIMS spectra collected from two fish species and seven body parts. Additionally, samples were prepared to simulate oil and cross-species contamination scenarios. The data collection process involved:

1. **Fish speciation:** Samples from multiple fish species (i.e. Hoki and Mackerel) were collected and analysed using REIMS. A dataset with 106 samples, with 44.44% Hoki and 55.56% Mackerel. The goal is to build a model that can accurately detect fish species from REIMS data.
2. **Fish body parts:** Various parts of fish (e.g. fillets, heads, livers, skins, guts, and frames) were isolated and analysed. A dataset with 30 samples, with 20% fillets, 20% heads, 10% livers, 20% skins, 20% guts and 10% frames. The goal is to differentiate between varying marine biomass samples using REIMS.
3. **Oil contamination:** Samples with varying concentrations of oil contamination (i.e. 50%, 25%, 10%, 5%, 1%, 0.1% or none 0%) to simulate contamination scenarios. A dataset with 126 samples, with 14.28% belonging to each class. The goal is to detect the presence of oil and quantify the concentration of oil contamination present.
4. **Cross-species contamination:** Samples were prepared by intentionally mixing tissues from different fish species (i.e. Hoki, Mackerel or Hoki-Mackerel mixed) to simulate contamination scenarios. A dataset with 144 samples, with 29.41% Hoki, 39.21% Mackerel and 31.32% mixed. The goal is to identify cross-species contamination where marine biomass has been adulterated by mixing products from two different species.
5. **Instance Recognition**: Samples from different batches of processed fish are used to differentiate between individual batches during fish processing. A pair-wise comparison task, where each pair of samples either belongs to the same or different batches of fish. A dataset with 72 samples, with 50% same, and 50% different. The goal is to detect processed fish that came from the same individual batch using REIMS.

The REIMS spectra were normalised to be within $x \in [0, 1]$ with L_2 normalization where ϵ is a small value to avoid division by zero (default: 1e-12), which gives

$$v = \frac{v}{max(||v||_2, \epsilon)} \tag{1}$$

$$||v||_2 = \sqrt{\sum_{k=1}^{n} |v_k|^2} \tag{2}$$

The dataset was then split into training and testing sets, with 80% training and 20% test, for each classification task.

3 Methods

With the datasets established, this section moves on to the heart of our analytical approach: the classification methods that extract meaningful insights from the REIMS spectra. This paper evaluates a diverse range of machine learning techniques to classify the REIMS spectra:

- **Traditional machine learning methods:** Random Forest (RF) [18], K-Nearest Neighbors (KNN) [9], Decision Trees (DT) [6], Naive Bayes (NB) [13], Logistic Regression (LR) [23], Support Vector Machines (SVM) [7], and Linear Discriminant Analysis (LDA) [2].
- **Ensemble method** [14]: A combination of the above traditional methods.
- **Deep neural networks:** Transformer [8, 42], Long Short-Term Memory (LSTM) [19], Variational Autoencoder (VAE) [22], Kolmogorov-Arnold Networks (KAN) [28], Convolutional Neural Network (CNN) [24–27], and Mamba [12].
- **Evolutionary computation:** Multiple Class Independent Feature Construction (MCIFC) [40, 41]

3.1 Traditional Machine Learning Methods

Experiments use the default settings from sklearn [33], except SVM with a linear kernel, and LR set to 2,000 for the maximum number of iterations, these exceptions were found experimentally with trial and error. The ensemble voting classifier uses hard voting, i.e. uses predicted class labels for majority rule voting. Additionally, more advanced classification techniques, such as deep learning and evolutionary computation methods, are detailed below.

3.2 Transformers

This paper adapts the transformer architecture, see fig. 3, introduced by Vaswani et al. [42] for machine translation, to marine biomass analysis. The architecture

features stacked encoder-decoder layers with residual connections [16]. Preliminary experiments explored various weight initialization methods - Xavier [10], Kaiming [15], and orthogonal [36]) - but found that the default weight initialization strategy to be the best. This work uses pre-norm [1,44] layer normalization,

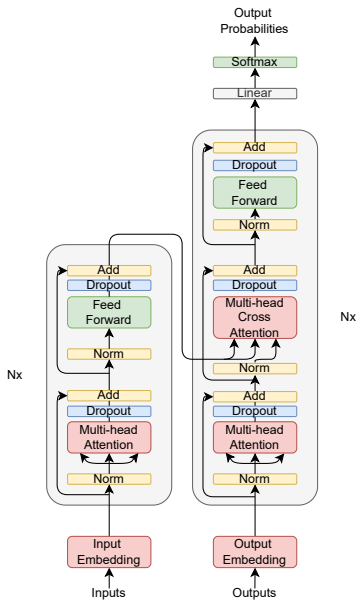
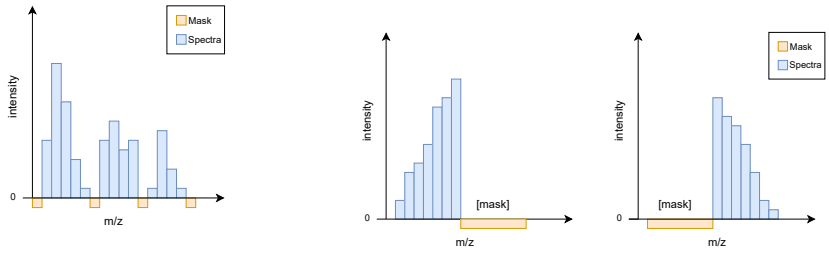


Fig. 3: Transformer Architecture

This paper introduces two novel unsupervised pre-training methods for mass spectrometry data, illustrated in fig. 4, which were inspired by techniques used in the BERT model [8]. These methods allow the model to learn general patterns from larger, unlabeled mass spectrometry datasets, creating useful embeddings that improve performance on smaller, fine-tuned datasets for specific downstream tasks.

1. **Masked Spectra Modelling (MSM)**: adapts masked language modelling to mass spectrometry. MSM randomly masks mass-to-charge ratios in spectra and predicts the missing values, evaluated as a regression task using mean square error (MSE) as the loss function.
2. **Next Spectra Prediction (NSP)**: is inspired by next sentence prediction, NSP splits spectra in half and predicts whether two halves belong to the same spectrum, evaluated as a pair-wise comparison using categorical cross entropy (CCE) as the loss function.



(a) MSM (b) NSP

Fig. 4: Masked spectra modelling (left) Next spectra prediction (right)

Table 1: Transformer parameter settings

Learning rate	1E-5	Epochs	100
Dropout	0.2	Label smoothing	0.1
Early stopping patience	5	Optimiser	AdamW
Loss: MSM	MSE	Loss: NSP & Speciation	CCE
Input dimensions	1023	Hidden dimensions	128
Output dimensions: MSM	1023	Output dimensions: NSP & Speciation	2
Output dimensions: Part	6	Number of layers	3
Number of heads	3		

Table 1 outlines the transformer configuration used. The transformer model in this paper employed the AdamW optimiser [29], this improves the Adam [21] by decoupling weight decay from the learning rate. Dropout [38] approximates a bagged ensemble of neural networks efficiently. For regularisation, label smoothing [39] softens class label targets by combining one-hot encodings with a uniform distribution. The transformer network utilises the Gaussian error linear unit (GELU) activations [17]. Data augmentation inflates the number of training instances, it duplicates each instance five times and injects noise [37], effectively expanding the training set five-fold. Early stopping [31] saves model parameters whenever validation loss improves, effectively tuning the hyperparameter of epochs [11].

3.3 Other Deep Neural Networks

Long Short-Term Memory (LSTM) [19] is a type of recurrent neural network designed to handle long-term dependencies in sequential data. Variational Autoencoders (VAE) [22] are generative models that learn to encode data into a

latent space and decode it back, allowing for both data compression and generation. Kolmogorov-Arnold Networks (KAN) [28] is a novel type of neural network architecture inspired by the Kolmogorov-Arnold representation theorem, aimed at improving function approximation capabilities. Convolutional Neural Networks (CNN) [24–27] use convolution operations to process grid-like data, particularly effective for image analysis tasks, because of spatial connectivity in pixels. In our case, mass spectrometry can be treated as a 1-D image, because neighbouring mass-to-charge ratios also share spatial connectivity. Mamba [12] is a recent state-space model architecture designed as an alternative to transformers, offering efficient processing of sequential data.

3.4 Evolutionary Computation

Complementing the deep learning approach of transformers, we next examine the use of multi-tree genetic programming, which offers a different paradigm for feature construction and classification. We employ multiple class independent feature construction (MCIFC) [41] in a novel application for marine biomass analysis using REIMS. Algorithm 1 gives the pseudocode for MCIFC.

Algorithm 1 GP-based multiple feature construction

Input : train_set, m;
Output : Best set of m trees;
Initialise a population of GP individuals. Each individual is an array of m trees;
best_inds ← the best e individuals;
while Maximum 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;

MCIFC represents candidate solutions as multiple trees, with one subtree per class (construction ratio of 1). This approach serves both feature construction and classification, using a winner-takes-all strategy for class prediction. MCIFC uses a two-tree representation when applied to the fish species classification task since there are two classes, i.e. Hoki and Mackerel. Additionally, MCIFC uses a six-tree representation when applied to the fish body parts classification task since there are six classes, i.e. fillet, heads, livers, skins, guts, and frames.

The genetic operators in MCIFC are crossover (80% probability) and mutation (20% probability) - both adapted from conventional genetic programming -

and Tournament selection with a size of 7 is used for parent selection. Crossover operates only between trees of the same class, while mutation randomly alters one subtree.

The fitness evaluation combines accuracy with a distance regularisation term:

$$\alpha \cdot \text{balanced\_accuracy} + (1-\alpha)(\beta \cdot \text{inter} + (1-\beta) \cdot (1-\text{intra}))$$

where $\alpha = 0.8$ - prioritising accuracy and $\beta = 0.5$ - balancing inter and intra-class distances. The interclass and intraclass distances are calculated as:

$$\text{inter} = \frac{1}{|S|} \sum_{i,j} d(i,j) \quad \forall \quad i \neq j \quad \text{and} \quad \text{class}(i) \neq \text{class}(j)$$

$$\text{intra} = \frac{1}{|S|} \sum_{i,j} d(i,j) \quad \forall \quad i \neq j \quad \text{and} \quad \text{class}(i) = \text{class}(j)$$

where $d(i,j)$ is the Euclidean distance between points i and j, and $|S|$ is the total number of pairs. This approach aims to maximise interclass distances while minimising intraclass distances, promoting better class separation in the constructed feature space.

Table 2 outlines the parameter settings of the MCIFC method. The construction ratio is the number of trees per class.

Table 2: MCFIC parameter settings

Function Set	$+, -, \times, \cos, \sin, \tan, -1*$	Terminal Set	$x_1, x_2, ..., x_n$
Maximum Tree Depth	6	Population size	1 * 1023 (= 1× #features)
Initial Population	Ramped Half-and-Half	Generations	400
Crossover	0.8	Mutation	0.2
Elitism	0.1	Selection	Tournament
Tournament Size	7	Construction ratio	1
Fitness weighting α	0.8	Distance weighting β	0.5

4 Results and Discussion

Having outlined our classification strategies, this section now presents and interprets the outcomes of applying these various machine learning techniques to the REIMS datasets. Table 3 gives the results of the classifiers on the test set, giving the average over 30 independent runs, with the best-performing model on the test set given in **bold**, and second-best in *italics*.

Table 3: Test accuracy classification results

	Fish speciation	Fish part	Oil	Cross-species	Instance Recognition
RF	95.88% ± 4.47%	40.00% ± 15.27%	38.73% ± 8.15%	81.04% ± 5.67%	49.33% ± 4.37%
KNN	83.69% ± 6.91%	31.66% ± 14.49%	31.94% ± 9.34%	68.68% ± 6.89%	*87.50% ± 0.00%*
DT	*99.13% ± 1.72%*	27.22% ± 13.25%	28.17% ± 7.34%	69.16% ± 5.59%	35.91% ± 4.30%
NB	87.97% ± 9.57%	45.00% ± 15.60%	32.50% ± 6.84%	55.70% ± 8.34%	55.00% ± 0.00%
LR	96.72% ± 4.75%	56.66% ± 15.27%	30.91% ± 8.32	*86.18% ± 5.03%*	50.00% ± 0.00%
SVM	95.97% ± 5.06%	56.11% ± 14.58%	35.63% ± 7.80%	85.53% ± 5.84%	50.00% ± 0.00%
LDA	96.47% ± 3.67%	45.55% ± 16.06%	31.86% ± 6.65%	81.37% ± 6.60%	52.50% ± 0.00%
Ensemble	98.16% ± 3.00%	51.66% ± 15.72%	37.26% ± 8.69%	84.34% ± 5.84%	49.50% ± 1.49%
Transformer	**99.58% ± 1.31%**	**63.33% ± 24.59%**	**42.56% ± 12.03%**	**86.24% ± 6.27%**	74.44% ± 9.74%
LSTM	96.81% ± 3.74%	58.33% ± 8.78%	31.53% ± 5.37%	83.22% ± 4.51%	63.99% ± 13.72%
VAE	98.18% ± 2.34%	50.00% ± 13.60%	35.38% ± 8.06%	77.41% ± 9.37%	66.66% ± 8.43%
KAN	97.27% ± 2.34%	*60.00% ± 17.91%*	21.92% ± 4.81%	69.67% ± 5.08%	68.00% ± 4.98%
CNN	97.72% ± 3.21%	59.99% ± 14.05%	38.46% ± 9.25%	77.41% ± 7.60%	72.66% ± 11.71%
Mamba	94.09% ± 5.27%	46.66% ± 10.54%	*40.76% ± 6.83%*	81.29% + 7.57%	64.66% ± 9.45%
MCIFC	94.54% ± 10.38%	55.45% ± 19.19%	38.32% ± 8.72%	72.39% ± 9.79%	**90.17% ± 9.01%**

Our study employed a diverse range of machine learning techniques to classify REIMS spectra for various tasks related to marine biomass analysis. The results demonstrate varying levels of success across different models and tasks, providing insights into the strengths and limitations of each approach.

4.1 Fish Speciation:

In the fish speciation task, most models performed exceptionally well, with the transformer model achieving the highest accuracy of 99.58%. The decision tree model also performed remarkably well (99.13%), suggesting that the spectral features distinguishing different fish species are highly distinct and can be effectively captured by both complex and simpler models. The high performance across various models indicates that REIMS spectra contain clear, discriminative information for fish species identification.

4.2 Fish Body Part:

The increased difficulty of fish body parts classification stems from the greater similarity in biochemical composition among different body parts compared to distinct species, making it a more complex problem for the models to solve. The transformer model outperformed others with 63.33% accuracy in this challenging task, likely due to the transformer's ability to capture subtle, long-range dependencies in the spectral data. The KAN followed this with 60.00% test accuracy, KAN embodies the universal approximation theorem, demonstrating that a feedforward network with sufficient hidden units can approximate any continuous multivariate function on compact subsets of $\mathbb{R}^N$.

4.3 Oil Contamination:

Oil contamination detection was the most challenging task - because it is a multi-class classification task the highest number of classes for any task - with the transformer model achieving the highest accuracy at 42.56%. The Mamba achieves the second-best test accuracy of 40.76%. Mamba is designed to capture long-range dependencies in sequential data efficiently. Unlike traditional LSTMs (31.53%) that struggle with long sequences, Mamba's selective state space formulation allows it to maintain and update relevant information over long distances. The random forest achieves the third-best test accuracy of 38.73%. The KAN performs the worst with 21.92% test accuracy. The relatively poor performance across all models suggests that oil contamination may not have a straightforward spectral signature, or that the chosen concentration levels were too subtle to create distinct patterns. This task might benefit from additional feature engineering or more sophisticated preprocessing techniques.

4.4 Cross-species Contamination:

For cross-species contamination detection, several models performed well, with the transformer (86.24%), logistic regression (86.18%), and SVM (85.53%) achieving the highest accuracies. The strong performance of logistic regression, a linear model, suggests that the features distinguishing cross-species contamination are largely linearly separable. This is encouraging for practical applications, as it indicates that even simple models can effectively detect cross-species contamination.

4.5 Instance Recognition:

For instance recognition for same batch detection, Mutliple Class-Independent Feature Construction (MCIFC), performing significantly better (90.17%) than the Transformer (74.44%) at third best. The strong performance of Convolutional Neural Networks (72.66%) for this task is worth mentioning, showing that the CNN can capture local feature interactions between important mass-to-charge ratios. Out of the traditional machine learning methods, the k-Nearest Neighbour with $k = 1$ (87.5%) performed best, memorizing what is essentially a look-up table for detecting individual instances from the same batch is an effective approach.

4.6 Further discussion:

The varying performance of different models across tasks highlights the importance of selecting appropriate algorithms for specific analytical challenges in marine biomass analysis. While the transformer model consistently excelled, simpler models like logistic regression demonstrated competitive performance in certain tasks, offering potential advantages in terms of interpretability and computational efficiency. The challenges faced in oil contamination detection and,

to a lesser extent, body part classification, point to areas where further research is needed. This might include exploring more advanced feature extraction techniques, increasing the size and diversity of the training dataset, or developing specialised model architectures tailored to these specific tasks. Overall, our results demonstrate the potential of combining REIMS with machine learning for rapid and accurate marine biomass analysis, while also highlighting areas for future improvement and research.

5 Further Analysis on the Explainability

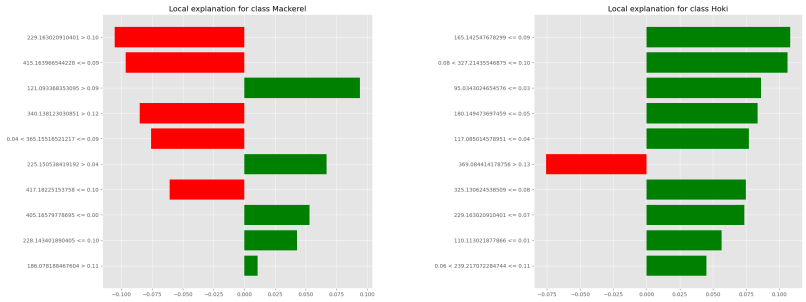
While the performance of our transformers and MCIFC is promising, understanding how they arrive at their predictions is crucial for building trust and gaining insights. To address this, the paper employs Local Interpretable Model-agnostic Explanations (LIME), a technique used to explain predictions made by complex black-box machine learning models [35]. LIME approximates a complex model's behaviour with a simpler and interpretable model (e.g. linear regression) for a specific instance in a local area to be understood. LIME creates and evaluates many altered versions through perturbations of an instance in the input data to see how those perturbations change the prediction. Through perturbations and their observed changes to the prediction, this information is used to generate a local explanation that highlights which features (e.g. mass-to-charge ratios) influenced the prediction. LIME charts are used to explain the predictions of machine learning models by showing which features (in this case, specific mass-to-charge ratios) are most influential for a particular prediction. In these LIME charts:

- Green bars: These represent features (mass-to-charge ratios) that contribute positively towards the predicted class. In other words, the presence or higher intensity of these features increases the likelihood of the sample being classified as the predicted class.
- Red bars: These represent features that contribute negatively towards the predicted class. The presence or higher intensity of these features decreases the likelihood of the sample being classified as the predicted class.
- The length of each bar: This indicates the magnitude of the feature's importance. Longer bars (whether green or red) signify that the corresponding feature strongly influences the model's prediction.
- The y-axis: This represents the mass-to-charge (m/z) ratios and their intensity thresholds from the mass spectrometry data
- The x-axis: This typically represents each feature's relative importance or contribution to the prediction.

5.1 Transformer on Fish Speciation

The transformer achieves the best classification accuracy on all datasets, but it notably performs best at fish species classification. Figure 5 gives the LIME

explanations for the transformer model for fish speciation, for the first Hoki and Mackerel instances, respectively. For Mackerel, there are several green bars in the 185-405 and 340-420 m/z ranges, indicating that these spectral features strongly support the classification of the sample as Mackerel. One of the largest red bars (i.e. negative correlation) is when the molecule at mass-to-charge ratio 229.1630 m/z is greater than the normalised intensity of 0.10, this suggests it is not commonly found in Mackerel. For Hoki, there are prominent green bars in the 95-370 m/z ranges, suggesting these features are important for identifying Hoki. There is one feature with a red bar (i.e. negative correlation) suggesting that when the mass-to-charge ratio 369.0844 m/z is greater than a normalised intensity of 0.13, this represents a molecule not commonly found in Mackerel.



(a) Mackerel (b) Hoki

Fig. 5: LIME explanations for transformer on fish species of Mackerel (left) Hoki (right)

5.2 KAN on Fish Body Part

The KAN performs the second best on the fish parts dataset. This is another difficult multi-class classification task, this time, however, there are only six classes. Here, fig. 6, is the LIME explanation for the KAN for the fish parts dataset. The strongest green bar is when the mass-to-charge ratio 365.3074 m/z is greater than the normalised intensity of 0.23, this molecule is likely highly correlated with the guts of fish body part. The strongest red bar (i.e. negative correlation) is when the mass-to-charge ratio 168.1298 m/z is within the range of normalised intensity $0.15 < y <= 0.21$, indicating when this molecule is present, it is likely not a guts fish body part sample.

5.3 Mamba on Oil Contamination

The Mamba performs second best in oil contamination detection and profiling. This is a difficult multi-class classification problem with seven different classes.

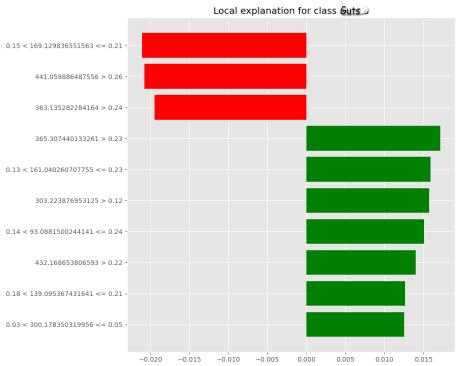


Fig. 6: LIME explanation for KAN for classification of the fish part of guts

Here, in fig. 7, is the LIME explanation for the Mamba for the oil dataset. The strongest green bar is when the mass-to-charge ratio of 80.0119 m/z is greater than a normalised intensity of 0.11, indicating this molecule is likely highly correlated with large concentrations of oil. There are only two red bars (i.e. negative correlation) when the mass-to-charge ratio 269.2526 m/z is greater than the normalised intensity of 0.13, and, when the mass-to-charge ratio 95.1034 m/z is less than or equal to the normalised intensity of 0.05, these molecules are likely associated with fish and not oil.

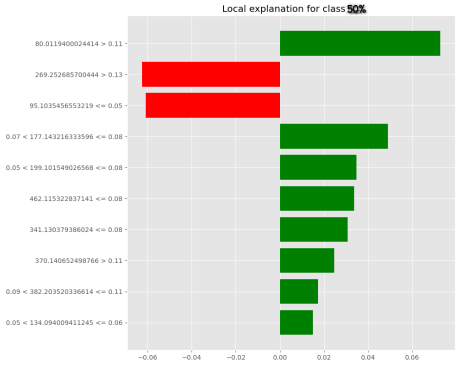


Fig. 7: LIME explanation for Mamba for oil contamination at 50%

5.4 Logistic Regression on Cross-Species Contamination

The Logistic Regression (LR) performs second best on the cross-species dataset. This is a multi-class classification task with three classes. Here, fig. 8, is the LIME explanation for LR for the cross-species dataset. The strongest green bar (i.e. positive correlation) is when the mass-to-charge ratio 295.2214 m/z is in the

normalised intensity range of $0.04 < y < 0.06$. Suggesting that small amounts of this molecule are indicative of adultered marine biomass that contains cross-species contamination. The strongest negative bar (i.e. negative correlation) is when the mass-to-charge ratio 281.1253 m/z is in the normalised intensity range $0.03 < y <= 0.06$. Suggesting that small amounts of this molecule are indicative of an unadulterated sample that consists of purely one species of fish.

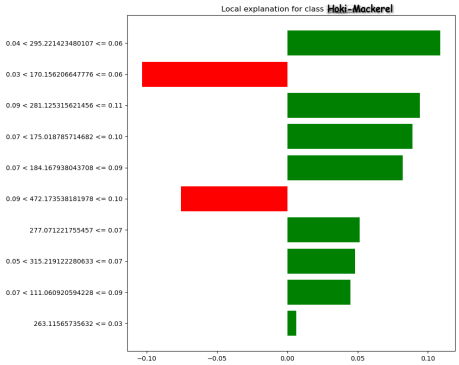
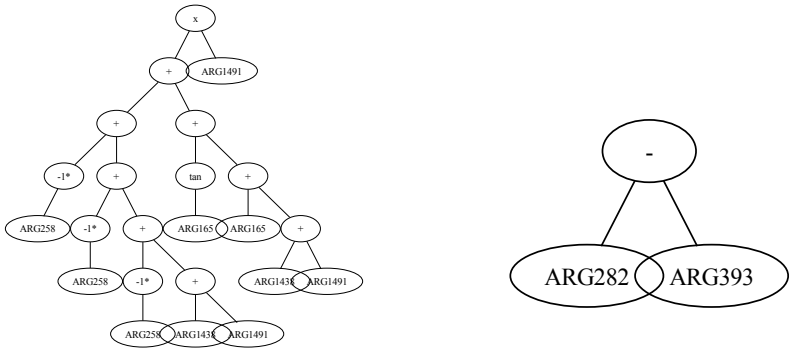


Fig. 8: LIME explanation for LR for cross-species contamination classification with Hoki-Mackerel mix

5.5 Genetic Programming Tree for Instance Recognition

Multiple Class-Independent Feature Construction performed the best on the instance recognition task. This is a pair-wise comparison task where to instances belong to the same or different batches of processed fish.

The class prediction of the GP tree is the arg max of the multiple output trees constructed by MCIFC - a winner-takes-all strategy. Figure 9 (left) gives the tree constructed by MCIFC for the pair-wise comparison which returns the different class. The constructed tree makes use of argument 258 - which corresponds to mass to charge ratio 168.1431 m/z - three times. This suggests it is an important variable for differentiating between individual batches. This value is negated by the negation operator and summed together three times, this argument will push the output of the tree towards negative values, decreasing the chance of the different class being predicted. Figure 9 (right) gives the tree constructed by MCIFC for the pair-wise comparison which returns the same class. The constructed tree is very simple, it makes use of two arguments - at a mass-to-charge ratio of 177.0343 m/z and 212.154 m/z - an arithmetic expression where the presence of the former pushes the output of the tree towards higher values, increasing the change of the class prediction of same. The presence of the latter pushes the output value of the tree to lower values, decreasing the chance of the class prediction the same.



(a) class: different (b) class: same

Fig. 9: GP tree for instance recognition: class different (left) same (right)

5.6 Decision Tree on Fish Species

To begin our exploration of model interpretability, we first examine the structure and decision-making process of the decision tree classifier, which offers clear insights into the key features driving classification.

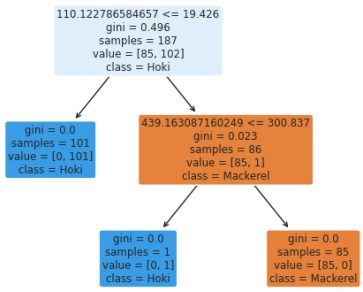


Fig. 10: Fish speciation: decision tree

Figure 10 gives the decision tree. This tree splits data when key mass-to-charge ratios exceed a threshold. For when 110.1228 m/z exceeds intensity 19.426, and later when 439.1631 m/z exceeds intensity 300.837. The intensity threshold for 439.1631 m/z is much greater than 110.1228 m/z, suggesting a large abundance of that molecule in Mackerel.

Assigning compounds to high-resolution mass spectrometry (HRMS) is challenging. This is due to the enormous amounts of metabolites present in homogenised fish tissues. The feature at 110.1128 m/z is consistent with a diphenol group. The feature 439.1631 m/z is consistent with a fragmented phospho-

lipid (1-Lauroyl-2-hydroxy-sn-glycero-3-phosphocholine). These compounds are known to vary between fish species and tissues [3]. This means there can be a dramatic variation in the molecules such as the lipid profiles, between samples of the same species or body part

6 Conclusion and Future Work

This study demonstrates the effectiveness of combining REIMS with advanced machine learning techniques for rapid and accurate marine biomass compositional analysis. The transformer-based model consistently outperformed other methods across all four classification tasks: fish speciation, body part classification, oil contamination detection, and cross-species contamination detection. The high accuracy achieved in fish speciation (99.58%) and instance recognition (90.17%) showcases the potential of this approach for quality control applications in the fishing industry. While the performance on fish body part classification (63.33%) oil contamination detection (42.56%) was lower, it still represents a significant improvement over traditional analysis methods in terms of speed and automation. Notably, in the cross-species contamination detection task, both the transformer model and logistic regression performed exceptionally well, achieving accuracies of 86.24% and 86.18% respectively. The strong performance of logistic regression, a relatively simple linear model, suggests that the features extracted from the REIMS spectra for this task are highly informative and linearly separable.

The application of explainable AI techniques, particularly LIME (Local Interpretable Model-agnostic Explanations), provided valuable insights into the decision-making processes of our models. These explanations revealed specific mass-to-charge ratios that strongly influence classifications, enhancing our understanding of the biochemical markers associated with different fish species, body parts, and contamination levels. For instance, the LIME analysis for fish speciation highlighted distinct spectral regions that differentiate Mackerel from Hoki, while the analysis for oil contamination pointed to specific molecular markers associated with high oil concentrations. This interpretability not only increases confidence in the model's predictions but also opens up possibilities for new scientific insights into the biochemical composition of marine biomass. It demonstrates that our approach can provide both accurate classifications and meaningful, chemically relevant explanations for those classifications.

Overall, this research opens up new possibilities for real-time, accurate, and interpretable analysis in marine biomass compositional studies, with significant implications for quality control, product optimization, and food safety in marine-based industries.

While our study has yielded promising results, it also opens up numerous avenues for further research and development. These are potential directions for expanding and refining our approach. Those directions for future work include:

- Real-time analysis: Develop a system for real-time REIMS data acquisition and analysis, allowing for immediate classification results in industrial settings.
- Adaptive sampling: Develop intelligent sampling strategies that use model interpretations to guide the collection of new data, focusing on areas of uncertainty or where additional samples could provide the most informative insights.
- Interpretability-driven model improvement: Use the insights gained from explainable AI techniques to refine model architectures, feature selection, or preprocessing steps, potentially leading to more accurate and robust classifications.
- Regulatory compliance: Work with regulatory bodies to ensure that the developed methods meet or exceed current standards for marine biomass analysis and food safety monitoring.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
2. Balakrishnama, S., Ganapathiraju, A.: Linear discriminant analysis-a brief tutorial. Institute for Signal and information Processing **18**(1998), 1–8 (1998)
3. Bettjeman, B.I., Hofman, K.A., Burgess, E.J., Perry, N.B., Killeen, D.P.: Seafood phospholipids: extraction efficiency and phosphorous nuclear magnetic resonance spectroscopy (31p nmr) profiles. Journal of the American Oil Chemists' Society **95**(7), 779–786 (2018)
4. Black, C., Chevallier, O.P., Cooper, K.M., Haughey, S.A., Balog, J., Takats, Z., Elliott, C.T., Cavin, C.: Rapid detection and specific identification of offals within minced beef samples utilising ambient mass spectrometry. Scientific reports **9**(1), 1–9 (2019)
5. Black, C., Chevallier, O.P., Haughey, S.A., Balog, J., Stead, S., Pringle, S.D., Riina, M.V., Martucci, F., Acutis, P.L., Morris, M., et al.: A real time metabolomic profiling approach to detecting fish fraud using rapid evaporative ionisation mass spectrometry. Metabolomics **13**(12), 1–13 (2017)
6. Breiman, L.: Classification and regression trees. Routledge (2017)
7. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3), 273–297 (1995)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
9. Fix, E., Hodges, J.L.: Discriminatory analysis. nonparametric discrimination: Consistency properties. International Statistical Review/Revue Internationale de Statistique **57**(3), 238–247 (1989)
10. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 249–256. JMLR Workshop and Conference Proceedings (2010)
11. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016), `http://www.deeplearningbook.org`

12. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
13. Hand, D.J., Yu, K.: Idiot's bayes—not so stupid after all? International statistical review **69**(3), 385–398 (2001)
14. Hansen, L.K., Salamon, P.: Neural network ensembles. IEEE transactions on pattern analysis and machine intelligence **12**(10), 993–1001 (1990)
15. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. pp. 1026–1034 (2015)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
17. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
18. Ho, T.K.: Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition. vol. 1, pp. 278–282. IEEE (1995)
19. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
20. Jha, S.N.: Rapid detection of food adulterants and contaminants: theory and practice. Academic Press (2015)
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
22. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
23. Kleinbaum, D.G., Dietz, K., Gail, M., Klein, M., Klein, M.: Logistic regression. Springer (2002)
24. LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., Jackel, L.: Handwritten digit recognition with a back-propagation network. Advances in neural information processing systems **2** (1989)
25. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural computation **1**(4), 541–551 (1989)
26. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
27. LeCun, Y., et al.: Generalization and network design strategies. Connectionism in perspective **19**(143-155), 18 (1989)
28. Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T.Y., Tegmark, M.: Kan: Kolmogorov-arnold networks. arXiv preprint arXiv:2404.19756 (2024)
29. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
30. McCann, S., Lowe, D.G.: Local naive bayes nearest neighbor for image classification. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3650–3656. IEEE (2012)
31. Morgan, N., Bourlard, H.: Generalization and parameter estimation in feedforward nets: Some experiments. Advances in neural information processing systems **2** (1989)
32. Pardo, M.Á., Jiménez, E., Pérez-Villarreal, B.: Misdescription incidents in seafood sector. Food Control **62**, 277–283 (2016)

33. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
34. for Primary Industries, M.: Hoki: New zealand's largest fishery. `https://www.mpi.govt.nz/fishing-aquaculture/fisheries-management/fish-stock-status/hoki-new-zealands-largest-fishery/` (2024)
35. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
36. Saxe, A.M., McClelland, J.L., Ganguli, S.: Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. arXiv preprint arXiv:1312.6120 (2013)
37. Sietsma, J., Dow, R.J.: Creating artificial neural networks that generalize. Neural networks **4**(1), 67–79 (1991)
38. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014)
39. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
40. Tran, B., Xue, B., Zhang, M.: Genetic programming for feature construction and selection in classification on high-dimensional data. Memetic Computing **8**(1), 3–15 (2016)
41. Tran, B., Xue, B., Zhang, M.: Genetic programming for multiple-feature construction on high-dimensional classification. Pattern Recognition **93**, 404–417 (2019)
42. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
43. Wood, J., Nguyen, B.H., Xue, B., Zhang, M., Killeen, D.: 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)
44. Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., Liu, T.: On layer normalization in the transformer architecture. In: International Conference on Machine Learning. pp. 10524–10533. PMLR (2020)
45. Zhang, R., Ross, A.B., Jacob, N., Agnew, M., Staincliffe, M., Farouk, M.M.: Rapid evaporative ionisation mass spectrometry fingerprinting can discriminate lamb meat due to different ageing methods and levels of dehydration. Journal of Proteomics **272**, 104771 (2023)