Hook, Line and Spectra: Machine Learning for Fish Species and Part Classification using Rapid Evaporative Ionization Mass Spectrometry

Jesse Wood, *Member, IEEE,* Bach Hoai Nguyen, *Member, IEEE,*, Bing Xue *Fellow,*, Mengjie Zhang *Fellow,*, and Daniel Killeen, *Member, IEEE,*

***Abstract*—Marine biomass composition analysis traditionally requires time-consuming processes and domain expertise. This study demonstrates the effectiveness of Rapid Evaporative ionization Mass Spectrometry (REIMS) combined with advanced machine learning techniques for accurate marine biomass composition determination. Using fish species and body parts as model systems representing diverse biochemical profiles, we investigate various machine learning methods, including unsupervised pretraining strategies for transformers. The deep learning approaches consistently outperformed traditional machine learning across all tasks. We further explored the explainability of the best-performing and mostly black-box models using Local Interpretable Model-agnostic Explanations to find important features driving decisions behind each of the top-performing classifiers. REIMS analysis with machine learning can be accurate and potentially explainable technique for automated marine biomass compositional analysis. It has potential applications in marine-based industry quality control, product optimization, and food safety monitoring.**

***Index Terms*—AI applications, explainable AI, machine learning, marine biomass, mass spectrometry, multidisciplinary AI**

I. Introduction

THE fish processing industry forms a critical component of the global seafood supply chain, transforming raw marine biomass into consumer products through multiple stages. This process typically involves species sorting, cleaning, filleting, packaging, and quality control at various checkpoints. Each stage presents unique challenges that can benefit from artificial intelligence and machine learning (AI/ML) solutions. The traditional fish processing workflow begins with the arrival of catch, where workers must rapidly sort different species - a task prone to human error particularly with similar-looking fish. The catch then moves through cleaning and filleting stations, where different body parts are separated for various products - from premium fillets to processed fish meal. Quality control occurs throughout, checking for freshness, proper handling, and accurate labeling. Finally, products are packaged and prepared for distribution. Several critical challenges exist within this workflow:

1) Quality control: Mislabeling and fraud remain persistent issues in the seafood industry [1], with economic and food safety implications. Studies have shown significant rates of species substitution in various markets [2].
2) Product Optimization: Different fish species and fish body parts have varying commercial values and uses, with some parts commanding premium prices in specific markets. Accurate classification of these parts ensures optimal resource utilization and maximizes economic value across the supply chain [3].
3) Safety Monitoring: Accurate tracking of processed species volumes is essential for both regulatory compliance and stock management [4].

These areas provide opportunities where automated analysis can signficantly improve fish processing. Specifically, we explore the application of machine learning to Rapid Evaporative Ionization Mass Spectrometry (REIMS) data across two critical classification tasks: fish species identification and body part classification. REIMS technology, combined with machine learning algorithms, aiming to offer a promising solution for real-time, accurate analysis during processing operations. Our focus on these specific tasks is driven by their direct impact on industry pain points:

(1) Species classification helps combat fraud and ensures proper resource management. (2) Body part identification helps to optimize processing efficiency and product value. (3) Accurate species counting supports both regulatory compliance and sustainability efforts.

This paper demonstrates how ML techniques applied to REIMS data can enhance the efficiency and accuracy of these critical fish processing operations, while supporting broader goals of sustainability and food security in the seafood industry.

This study utilizes datasets provided by New Zealand Plant and Food Research as part of Cyber-Marine [5]. REIMS can be used to optimize the value obtained from seafood resources. The dataset consists of mass spectrometry samples collected using REIMS, where an electrosurgical knife is used to create an aerosol from the samples. This aerosol is then directed into a mass spectrometer, where ionization occurs, allowing for mass-to-charge ratio analysis. Each sample undergoes multiple incisions lasting 3-5 seconds, providing detailed chemometric data in the mass range of m/z 77.04 - 999.32.

REIMS marine biomass analysis faces several challenges: time-consuming manual “offline” analysis, costly domain expertise required, high-dimensionality [6], and few training samples, and the need for automated “online” inference.

Jesse Wood, Bach Hoai Nguyen, Bing Xue, and Mengjie Zhang are with Victoria University of Wellington, Te Herenga Waka, Wellington, New Zealand

Daniel Killeen is with Plant and Food Research Limited, Nelson, New Zealand

Manuscript received October 20, 2024; revised October 20,2024.

Note, "online" inference in the domain of chemistry and fish processing, not to be confused with "online" learning from machine learning. The rapid nature of REIMS necessitates equally rapid inference of its results, as traditional analytical chemistry techniques which take several hours are too slow [7]. Furthermore, current analytical methods for REIMS data often require domain expertise in chemistry and fish processing, which does not match the speed of REIMS. Traditionally, samples would be sent away for "offline" analysis by domain experts in chemistry, we look to develop methods capable of automated inference for "online" analysis on the production line of a fish processing factory. REIMS also produces high-dimensional data, with this particular dataset having 2080 mass-to-charge ratios as features, but there are limited training instances due to the time-consuming and expensive task of sample preparation. Additionally, industry applications require fast, accurate, and interpretable models that can be verified and troubleshooted in real-world scenarios.

To address the above challenges, this paper proposes several innovative machine learning approaches that provide automated inference, eliminating the need for domain expertise in chemistry and fish processing. To handle the high-dimensionality of REIMS data, this paper utilizes deep learning [8], [9] and evolutionary computation [10], [11] that can address complex feature interactions in mass spectra with limited training instances. Techniques like BERT [8] and attention mechanisms [9] can capture complex, non-linear relationships between features in high-dimensional data. To mitigate the limited number of training samples, we implement unsupervised pretraining. The unsupervised pretraining approach involves training the model on a large amount of unlabeled data before fine-tuning it on the limited labelled dataset. The model learns general features and patterns from the unlabeled data, which can then be transferred to the specific task at hand. This can significantly improve performance when labelled data is scarce. Finally, we employ Local Interpretable Model-agnostic Explanations (LIME) [12] 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.

The main contributions of the paper are:

1) **Real-time Marine Biomass Analysis:** The paper demonstrates the use of REIMS combined with advanced machine learning techniques to enable automated analysis of marine biomass. This represents a significant improvement over traditional, time-consuming methods.
2) **Machine learning on sequential data:** The paper demonstrates that deep learning approaches, particularly transformers with and without progressive masking pre-training, consistently outperform traditional machine learning methods for analyzing sequential REIMS data.
3) **Feature importance:** Identify the important mass-to-charge ratios for the best performing models. This enhances domain knowledge in fish processing and chemistry.

II. Related works

Building upon the foundation laid in the introduction, this section delves deeper into the existing body of research on marine biomass analysis, exploring both traditional methods and recent advancements in REIMS technology, while highlighting the gaps and challenges that our proposed approach aims to address. We also provide the necessary background on deep learning required.

A. Marine Biomass

Mislabelling is a significant issue in the global seafood industry, with a meta-analysis of genomic profiling methods finding an average mislabelling rate of 30% worldwide [2]. Machine learning methods using REIMS data offer a promising solution to this problem by enabling more accurate fish species classification. For example, in 2016, a restaurant in Melbourne was accused of serving catfish instead of dory [13], highlighting the need for better species detection techniques. REIMS technology, which works on both raw and cooked biomass, can combat fraud by ensuring species authenticity. Approximately 40% of a fish is edible fillet, while the remaining 60% can be repurposed into products like fertilizers or high-value pharmaceutical-grade omega-3 concentrates. Fish oil, rich in omega-3 polyunsaturated fatty acids [14], is nutritionally essential but increasingly scarce in Western diets [15]. REIMS-based machine learning methods in fish processing also help identify high-value parts for repurposing into valuable products, contributing to the rising consumer demand for omega-3 supplements made from diverse marine biomass [16].

B. REIMS

Traditional approaches for analyzing marine biomass composition have long been the cornerstone of research and quality control in the seafood industry. These methods include Gas Chromatography-Mass Spectrometry [17], Nuclear Magnetic Resonance Spectroscopy [18], and Genomic Profiling [2]. While these techniques have proven valuable, they often come with significant drawbacks. They are typically time-consuming, requiring extensive sample preparation and analysis time. Additionally, they are labour-intensive, demanding skilled technicians to operate complex equipment and interpret results. Perhaps most importantly, these methods necessitate substantial domain expertise, limiting their accessibility and scalability in real-world applications. These limitations have spurred the search for rapid techniques for marine biomass analysis in fish processing.

In recent years, REIMS has emerged as a promising technique for rapid and accurate analysis of biological samples, addressing many of the limitations of traditional methods. Since its introduction in the original paper by Balog et al. [19], REIMS has demonstrated its versatility and effectiveness across various applications. For instance, it has been successfully employed to detect horse offal mixed into beef mince at concentrations as low as 1%-5%, showcasing its potential in addressing food fraud [1]. In the realm of seafood,

REIMS has been applied to binary classification tasks for detecting fish species and catch methods, further illustrating its utility in combating fish fraud [20]. Historically, REIMS biomass analysis has primarily relied on Orthogonal Partial Least Squares Discriminant Analysis (OPLS-DA) [21]–[23] with Principal Component Analysis (PCA) for dimensionality reduction [24]. However, this PCA-OPLS-DA approach has limitations, particularly in its reliance on outlier thresholding for adulteration detection, which requires manually defined hyperparameters set by domain experts in chemistry. This work proposes automated methods with learnable parameters that do not require domain expertise in chemistry to be configured. Additionally, this work proposes deep learning and evolutionary computation methods from machine learning that outperform the traditional OPLS-DA approach.

C. Machine Learning for REIMS

Deep learning models were selected for REIMS marine biomass analysis because of their ability to handle complex, high-dimensional data with sequential or structured dependencies, which are inherent in REIMS data. Transformers [8], [9], known for their powerful self-attention mechanisms, excel at weighing the importance of different features in sequential data, making them well-suited for identifying patterns in REIMS spectra. Since REIMS data, like sequences in language, consists of ordered data points (mass-to-charge ratios) with varying degrees of importance, the transformer's attention mechanism allows it to dynamically focus on critical parts of the spectrum for classification or prediction. Long Short-Term Memory (LSTM) networks [25], a type of recurrent neural network (RNN), are also effective for REIMS data because they capture long-term dependencies in sequential data. This is crucial for REIMS analysis, as spectral data might contain dependencies across distant mass-to-charge values. LSTM's ability to store and retrieve information over long sequences enhances its performance in such tasks, especially when the signal patterns may not be immediately adjacent. Variational Autoencoders (VAEs) [26] offer an effective approach to handling the complexity and variability of REIMS data by learning a compressed, latent representation of the spectral information. VAEs can also reconstruct this data, making them ideal for tasks like species and body parts classification, where they can model and detect small anomalies or deviations in the spectral data. Kolmogorov-Arnold Networks (KAN) [27] are highly efficient at approximating complex functions, which is essential in REIMS data analysis, where subtle differences in spectra can indicate different classes such as species and body parts. KAN's ability to improve function approximation makes it especially powerful for handling non-linear patterns in mass spectrometry data, which traditional models may struggle to capture. Convolutional Neural Networks (CNN) [28]–[31], although primarily used in image processing, are highly effective for REIMS data due to the spatial connectivity in mass spectra. Just as neighbouring pixels in images share spatial relationships, neighbouring mass-to-charge ratios in REIMS data also exhibit dependencies. CNNs can exploit this structure to identify patterns in one-dimensional data, treating mass spectra similarly to 1D images. Finally, Mamba [32], a state-space model, offers an efficient alternative to transformers for sequential data processing. Mamba is designed for high-performance handling of complex time-series data, making it an excellent fit for REIMS analysis, where both computational efficiency and the ability to model sequential dependencies are essential for automated or large-scale biomass analysis.

III. Methods

With the background 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:

- **Benchmark method:** Orthogonal Partial Least Squares Disrciminant Analysis (OPLS-DA) [22]. OPLS-DA is a supervised multivariate analysis technique that separates predictive from non-predictive variation in complex datasets to improve model interpretability and identify variables that drive class separation.
- **Traditional machine learning methods:** Random Forest (RF) [33], K-Nearest Neighbors (KNN) [34], Decision Trees (DT) [35], Naive Bayes (NB) [36], Logistic Regression (LR) [37], Support Vector Machines (SVM) [38], and Linear Discriminant Analysis (LDA) [23].
- **Ensemble method:** [39]: A combination of the above traditional methods. A hard voting ensemble classifier combines multiple base classifiers by having each classifier make a prediction and taking the most common predicted class label as the final output through majority voting.
- **Deep neural networks:** Transformer [8], [9], Long Short-Term Memory (LSTM) [25], Variational Autoencoder (VAE) [26], Convolutional Neural Network (CNN) [28]–[31], Kolmogorov-Arnold Networks (KAN) [27] and Mamba [32].
- **Genetic Programming:** Multiple Class Independent Feature Construction (MCIFC) [10], [11]. The MCIFC algorithm represents candidate solutions as multiple trees, with one subtree per class. This structure serves feature construction and classification purposes, employing a winner-takes-all strategy for class prediction.

A. Transformer

1) Architecture: The transformer model, originally proposed in the seminal paper by Vaswani et al. [9], revolutionized natural language processing and other tasks involving sequence data by relying entirely on self-attention mechanisms instead of recurrent or convolutional layers. Our transformer model consists of an encoder only structure, where encoders are stacked as layers. Each encoder layer comprises multi-head self-attention followed by position-wise feed-forward layers. We implement a transformer without positional embeddings (NoPE) [40].

In the architecture used for this work (fig. 1), the encoder blocks are equipped with residual connections [41], allowing gradients to flow efficiently during backpropagation. These

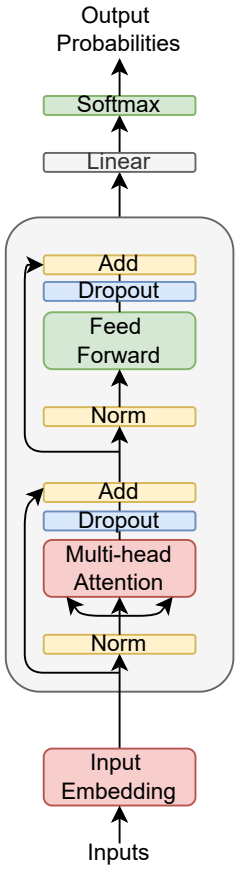


Fig. 1: Transformer Architecture.

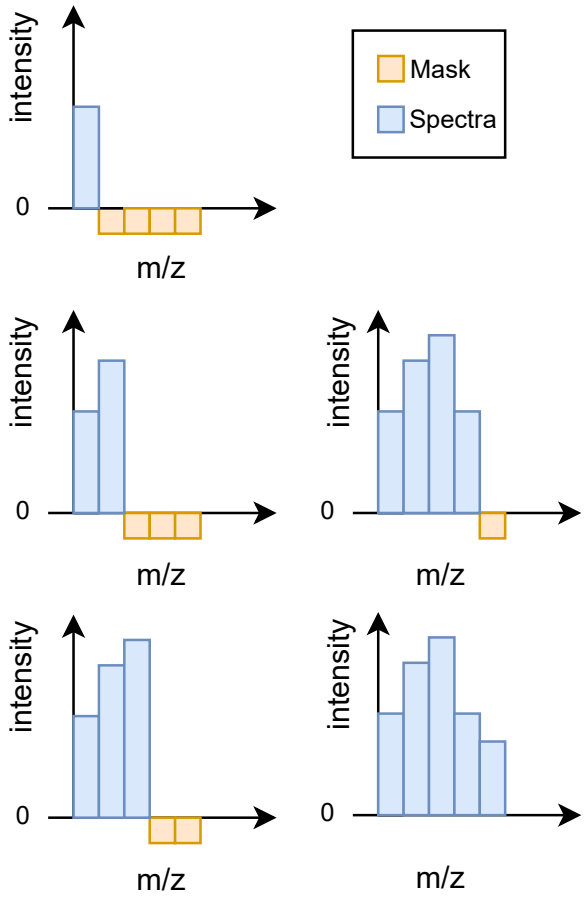


Fig. 2: Masked language modelling.

residual connections act as "gradient superhighways", enabling deeper models without the risk of vanishing or exploding gradients, thus allowing better training stability.

A notable aspect of the transformer architecture used in this work is the choice of pre-norm layer normalization [42], [43], where the layer normalization is applied before multi-head attention and feed-forward layers. This approach contrasts with post-norm layer normalization (used after the attention block), as it stabilizes training and improves the convergence of deep transformers by ensuring more consistent gradients across layers. By normalizing before the main components of each layer, the pre-norm structure helps maintain better gradient flow across the network, contributing to more effective training of the encoder layers.

2) Progressive masking: Figure 2 illustrates the concept of progressive masking in pre-training transformer models. At the bottom right we see the original mass spectra. On the top left, we see the first mask, which applies a mask to all spectra except the first one. On from that we see masks that slowly shrink down until we reach the original spectra. These patterns demonstrate how the masking process evolved, starting with masking just one spectra and progressively unmasking more spectra in the sample. Mask 1 shows only the first spectra, with the rest masked. Mask 2 reveals two spectra, masking the remainder. Mask 3 unmasks one more spectra, showing three spectra. The final mask shows all the spectra except the final one. This progressive masking technique creates multiple training examples from a single spectra, effectively increasing the amount and diversity of training data for the transformer model. In this work, we apply left-to-right progressive masking to REIMS data. Instead of sentences in natural language processing, we are masking mass spectra, and pre-training has the model predict the masked spectra. This amortized the limited number of training samples by creating 2080 masked spectra per instance to train from, resulting in a training set of 2080 features $\times$ 72 samples = 149,760 instances.

3) Pre-trained Transformers: Pre-training is an extension of transformers that allows them to be pre-trained on a general task, then transfer the pretrained weights to a transformer model to be fine-tuned on a downstream task. This paper adopts unsupervised pre-training inspired by BERT [8] to improve the performance of transformer models on mass spectrometry tasks. Unsupervised pre-training offers significant benefits, particularly for models working with limited labelled data. By training on large-scale, unlabeled datasets, the model learns general patterns that capture the underlying structure of the data, resulting in useful embeddings that can be fine-tuned for specific downstream tasks with smaller, labelled datasets. This approach mitigates the need for extensive labelled data while still providing high-quality results.

This approach is an adaptation of the masked language

modelling (MLM) task used in BERT to handle mass spectrometry data. In MLM, tokens in a sentence are progressively masked, and the model is trained to predict these masked tokens. Analogously, in masked spectra modelling (MSM), mass-to-charge ratios in spectra are progressively masked, and the model learns to predict the missing values. This is framed as a regression task, where the loss function is the mean squared error (MSE). By learning to predict missing mass-to-charge ratios, the model develops a robust understanding of the relationships between features in the spectra, making it well-suited for downstream tasks. We use left-to-right progressive masking to amortize the limited number of training instances.

By pre-training on this task, the transformer network learns valuable domain-specific representations. When fine-tuned on smaller, labelled datasets, the model can leverage these pre-trained weights, resulting in improved accuracy, faster convergence, and better generalization. This approach is particularly advantageous in fields like mass spectrometry, where labeled data is limited but large amounts of unlabeled data are readily available.

IV. Experimental setup

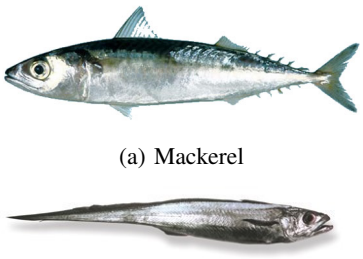
Having outlined our various machine learning approaches for analyzing REIMS data, we now describe the experimental setup used to evaluate these methods, including the benchmark technique, datasets, and parameter settings used in our evaluation

A. Benchmark technique

To evaluate the performance of the proposed methods, Orthogonal Partial Least Squares Disrciminant Analysis (OPLS-DA) [22] is used as a benchmark to compare new approaches to the existing methods for REIMS analysis. OPLS-DA is the standard technique for biomass analysis using REIMS, as supported by prevelant use in the literature [1], [7], [19], [20]. Therefore, we use OPLS-DA as a benchmark for comparative performance, showing the contrast between traditional and new approaches for biomass analysis with REIMS. OPLS-DA is considered a machine learning technique, specifically a supervised dimensionality reduction and classification method. It falls into the category of linear supervised machine learning algorithms, similar to PLS-DA (Partial Least Squares Discriminant Analysis) and LDA (Linear Discriminant Analysis). However, its primary strength lies in its ability to separate systematic variation into predictive and orthogonal (non-predictive) components, which makes it particularly useful for both classification and biomarker identification in fields like metabolomics and chemometrics.

B. Experimental Settings

Each method is evaluated and the average is given over 30 independent runs. Stratified k-fold cross-validation, with $k = 5$ for fish species and $k = 3$ for body parts, is particularly beneficial for evaluating model performance on datasets with limited training samples and imbalanced classes. This method ensures that each fold maintains a class distribution similar to the entire dataset, which helps the model learn effectively from both majority and minority classes. By doing so, it reduces the variance of performance estimates, leading to more stable and reliable metrics. Additionally, it maximizes the use of available data, allowing each sample to contribute to both training and validation, which is crucial for small datasets. With three and five-fold cross-validations, the model is tested across various scenarios, improving its generalization to unseen data and providing a comprehensive evaluation of its performance.



(a) Mackerel

(b) Hoki

Fig. 3: Mackerel (left) Hoki (right) fish species.

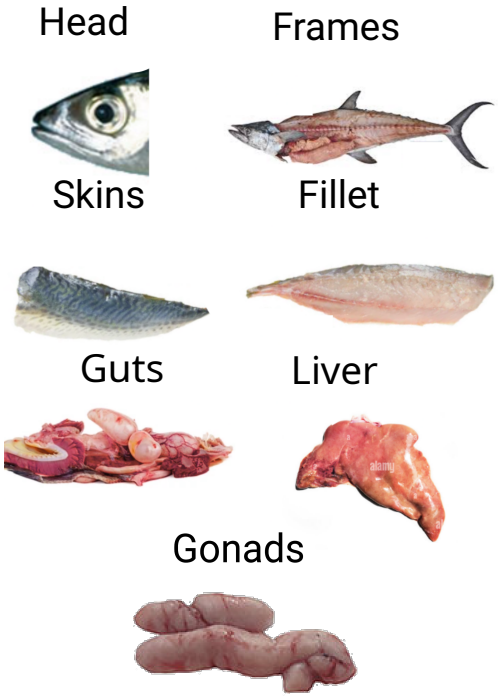


Fig. 4: Fish body parts.

C. Datasets

Figure 3 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 [44].

For illustrative purposes, the different fish body parts, which are shared across both species of fish, are given in Figure 4.

The dataset used consists of REIMS spectra collected from two fish species and seven body parts. Particularly, we will

have two different datasets corresponding to two different (following) tasks:

1) **Species classification:** The task is to distinguish between two species of fish (Hoki and Mackerel) based on 2080 features derived from REIMS analysis. This classification is crucial for food authentication and quality control in the seafood industry, helping prevent species substitution fraud and ensure accurate product labeling. We focus on pure (non-contaminated/non-mixed) samples to establish a reliable baseline for species identification. The dataset contains 106 samples, with a relatively balanced distribution of 44.44% Hoki and 55.56% Mackerel. These proportions reflect the natural availability of samples while maintaining sufficient representation for both species to train a robust classifier.
2) **Body parts classification:** This multi-class classification task aims to identify seven distinct fish parts (fillets, heads, livers, skins, gonads, guts, and frames) using REIMS data. The classification supports process automation by enabling automated sorting and processing in seafood production lines, while helping maximize the value of each fish part, such as using fillets for premium products and frames for fish meal. Furthermore, precise classification ensures proper tracking and documentation of different fish components throughout the supply chain. The dataset contains 33 samples with a distribution of 16.66% each for fillets, heads, livers, skins, and guts, and 8.33% each for gonads and frames. The relatively small sample size per class is attributed to a limited number of annotated samples for each class of body part.

The REIMS spectra were normalised to be within $x \in [0, 1]$, fitted to the training set of each fold. Let $X = \{x_1, x_2, \ldots, x_n\}$ be a dataset containing n elements. The normalized value x_i' for each element x_i is given by:

$$x_i' = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

where:

- $x_{\min}$ is the minimum value in the dataset X
- $x_{\max}$ is the maximum value in the dataset X

D. Parameter settings

Experiments use the default settings from sklearn [45], except SVM with a linear kernel, and LR set to 2,000 for the maximum number of iterations. The ensemble voting classifier combines all the traditional machine learning methods into one model. The ensemble uses hard voting, i.e. uses predicted class labels for majority rule voting.

The deep learning models all use the following parameters. The AdamW optimizer [46] decouples weight decay from the learning rate, an improvement over the popular Adam optimizer [47]. Dropout [48] turns off neurons at random during training to efficiently approximate a bagged ensemble of subneural networks. Label smoothing [49] softens class labels by combining the one-hot encodings with a uniform distribution, adding noise to the class labels. The deep learning networks uses Gaussian error linear units (GELU) [50] for activation functions. Early stopping [51] is one of the most common forms of regularization which saves the model parameters when the validation loss improves, it tunes the hyperparameter of epochs [52]. To allow fair comparison, each model has the same hyperparameters; a hidden dimension of 128, trained for 100 epochs, a learning rate of 1e-5, a batch size of 64, 4 layers (where applicable), dropout of $p = 0.2$ and label smoothing of 0.1.

TABLE I: 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: Speciation & Part	CCE
Input dimensions	2080
Hidden dimensions	128
Output dimensions: MSM	2080
Output dimensions: Speciation	2
Output dimensions: Part	7
Number of layers	4
Number of heads	4

Table I gives the configuration of hyperparameters for the transformer - these settings were derived through trial and error via experimentation.

We follow the original paper for the parameter settings for MCIFC [11]. We use a construction ratio of 1, allowing for one tree per class.

V. Results and Discussions

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 II and table III gives the results of the classifiers on the training and test set, with the best-performing model on the test set given in **bold**, and the second-best are given in *italics*. Note that the method *pre-trained* indicates the transformer with progressive left-to-right masked pre-training. The transformer was pre-trained on the training data of each individual fold during stratified k-fold cross-validation.

A. Fish Species Classification

For the task of fish species classification, the best-performing models were the pre-trained Transformer (99.62%). This model excels in capturing the intricate patterns in the REIMS data, which provides distinct signatures for different fish species. The high performance of the decision tree model (99.17%) shows that even traditional machine learning methods are highly effective in this domain. The tree-based models like Decision Trees and Random Forests work well because they can split the data based on highly discriminative features, capturing non-linear relationships effectively. For a decision tree, while individual splits are linear (axis-parallel), their combination creates non-linear decision boundaries.

TABLE II: Classification results for fish species identification.

Method	Train	Test
OPLS-DA	98.91% ± 0.74%	96.39% ± 4.44%
KNN	95.76% ± 0.00%	79.37% ± 0.00%
DT	100.00% ± 0.00%	99.17% ± 0.00%
LR	100.00% ± 0.00%	85.21% ± 0.00%
LDA	98.54% ± 0.00%	92.29% ± 0.00%
NB	89.17% ± 0.00%	66.67% ± 0.00%
RF	100.00% ± 0.00%	90.05% ± 0.56%
SVM	100.00% ± 0.00%	84.58% ± 0.00%
Ensemble	100.00% ± 0.00%	87.84% ± 0.40%
Transformer	100.00% ± 0.00%	*99.17% ± 1.67%*
Pre-trained	100.00% ± 0.00%	**99.62% ± 1.15%**
LSTM	100.00% ± 0.00%	*98.84% ± 1.76%*
VAE	100.00% ± 0.00%	98.64% ± 1.94%
KAN	100.00% ± 0.00%	97.41% ± 2.45%
CNN	100.00% ± 0.00%	96.87% ± 3.24%
Mamba	100.00% ± 0.00%	98.27% ± 2.14%
MCIFC	100.00% ± 0.00%	97.89% ± 2.59%

TABLE III: Classification results for fish body parts identification.

Method	Train	Test
OPLS-DA	80.11% ± 2.86%	51.17% ± 22.16%
KNN	43.06% ± 0.00%	39.17% ± 0.00%
DT	100.00% ± 0.00%	35.50% ± 4.35%
LR	100.00% ± 0.00%	59.58% ± 0.00%
LDA	74.31% ± 0.00%	52.92% ± 0.00%
NB	100.00% ± 0.00%	48.33% ± 0.00%
RF	100.00% ± 0.00%	61.67% ± 0.00%
SVM	100.00% ± 0.00%	52.33% ± 2.57%
Ensemble	100.00% ± 0.00%	52.33% ± 2.57%
Transformer	100.00% ± 0.00%	**84.06% ± 6.42%**
Pre-trained	100.00% ± 0.00%	*83.94% ± 7.12%*
LSTM	100.00% ± 0.00%	82.11% ± 9.15%
VAE	85.43% ± 6.28%	74.81% ± 13.84%
KAN	100.00% ± 0.00%	73.06% ± 9.58%
CNN	100.00% ± 0.00%	70.41% ± 13.75%
Mamba	100.00% ± 0.00%	80.67% ± 8.73%
MCIFC	97.95% ± 1.61%	55.45% ± 19.19%

The consistently high test accuracy across all models suggests that the REIMS dataset for fish species contains strong, distinguishable signals that can be effectively exploited by various machine learning techniques. This makes the classification task easier for both deep learning models and traditional methods. The models excel at this task because the REIMS data likely provides clear, consistent, and high-dimensional representations of species differences, which can be leveraged by the deep architectures for feature extraction and by traditional methods for decision-making.

All the deep learning models consistently outperform the traditional OPLS-DA method - with the pre-trained transformer getting 96.39% test accuracy - from the literature for REIMS analysis. The research field of REIMS analysis should consider deep learning methods for other applications, as they offer superior performance.

B. Fish Body Part

The transformer without pre-training performed the best in the task of classifying fish body parts, achieving a test accuracy of 84.06%. These models are well-suited for this task because they can handle complex and multi-dimensional input data like REIMS, capturing the subtle differences between body parts through advanced feature extraction and context awareness. LSTMs, with their ability to capture sequential dependencies, also perform well (82.11%), suggesting some temporal or positional dependencies in the ionization patterns that relate to specific body parts.

Traditional machine learning methods, however, show lower performance compared to the fish species classification task. This suggests that classifying fish body parts is inherently more complex due to less distinct signal differences between body parts, making it harder for simpler models to differentiate between classes. This increased difficulty likely arises from overlapping chemical compositions between different parts of the same species. Previous work [17] on fish species and body parts classification with gas chromatography data illustrated the increased difficulty of body parts classification.

Again, all the deep learning methods - with the transformer achieving the best test accuracy at 84.06% - outperform the OPLS-DA method (51.17%). For the second task, deep learning methods have proven to be superior to the traditional approach from the literature.

C. Summary:

Across all tasks, the deep learning methods offered superior performance to the OPLS-DA method that dominates the literature [1], [7], [19], [20] on REIMS analysis. Future work in the field for other applications of REIMS analysis should consider deep learning methods as a viable alternative. 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 decision trees demonstrated competitive performance in certain tasks, offering potential advantages in terms of interpretability and computational efficiency. The challenges faced in 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 automated and accurate marine biomass analysis, while also highlighting areas for future improvement and research.

VI. Further Analysis on Feature Importance

While the performance of our vanilla and pretrained transformers is promising, understanding how they arrive at their predictions is crucial for building trust and gaining insights. It is important to identify important features driving decisions made by black-box models, such that these models can be understood, trusted, and verified by domain experts in chemistry and fish processing. To address this, we employ Local Interpretable Model-agnostic Explanations (LIME), a technique used to explain predictions made by complex black-box machine learning models [53]. We analyze the top 5 most important features of the top-performing models that have been identified by LIME. 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 influenced the prediction. LIME explanations, or feature importance 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 x-axis:** The length of each bar 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 x-axis represents the feature importance.
- **The y-axis:** This represents the mass-to-charge (m/z) ratios and their intensity thresholds from the mass spectrometry data. The y-axis represents the important features.

A. *Fish Species Classification*

The pre-trained transformer achieves the best classification accuracy (99.62%) for fish species classification. Figure 5 gives the LIME explanation for the pre-trained transformer for the fish species Mackerel. The most important feature, and the strongest green bar, is when the mass-to-charge ratio 794.0990 m/z is within the normalized intensity range $0.28 < y \leq 0.47$. Suggesting that large amounts of this molecule are present in the fish species Mackerel.

Figure 6 gives the LIME explanation for the pre-trained transformer for the fish species Hoki. The most important feature, and the strongest red bar, is when the mass-to-charge ratio 229.0710 m/z is within the normalized intensity range $0.26 < y \leq 0.36$. Suggesting that large amounts of this molecule indicate a sample does not belong to the fish species Hoki.

Figure 7 gives the decision tree with near-perfect accuracy, showing how a simple model with two splits, can classify fish species, giving both a highly accurate and interpretable model. The figure shows the two mass-to-charge ratios and their intensity thresholds for which they based their decision boundaries on.

B. *Fish Body Part*

The transformer performs the best (83.94%) on the fish parts dataset. Figure 8 gives the LIME explanation for the

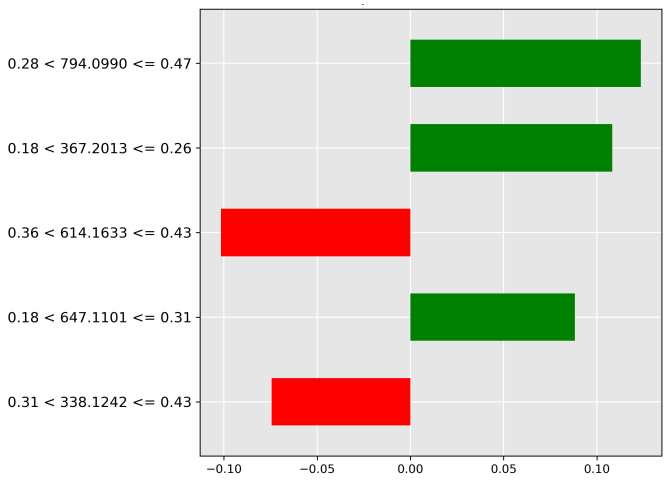


Fig. 5: Lime explanation for **pre-trained transformer** for classification of fish species **Mackerel**.

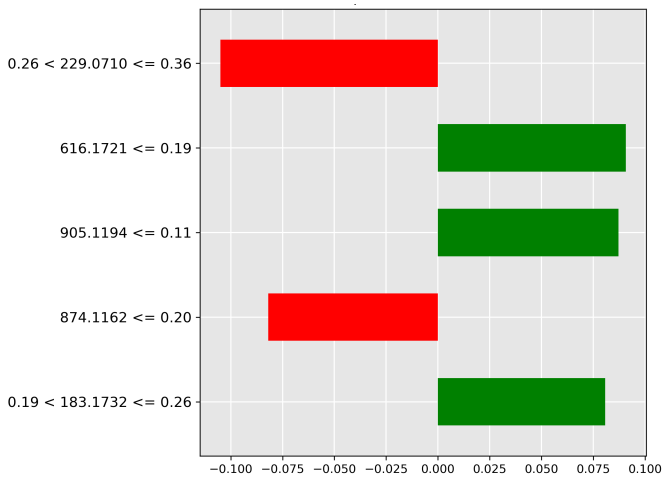


Fig. 6: Lime explanation for **pre-trained transformer** for classification of fish species **Hoki**.

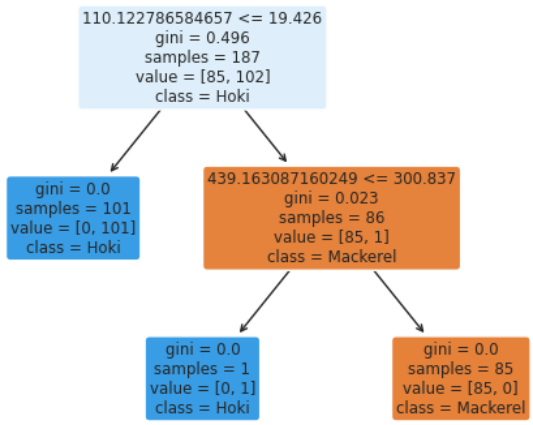


Fig. 7: **Decision tree** for fish species.

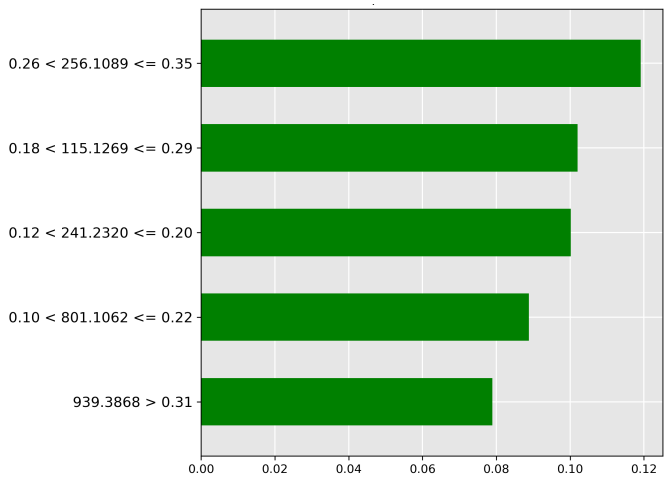


Fig. 8: Lime explanation for **transformer** for classification of fish part **head**.

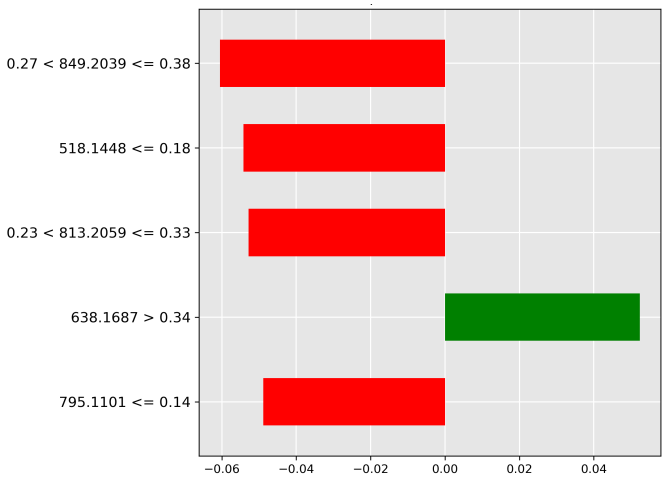


Fig. 10: Lime explanation for **transformer** for classification of fish part **liver**.

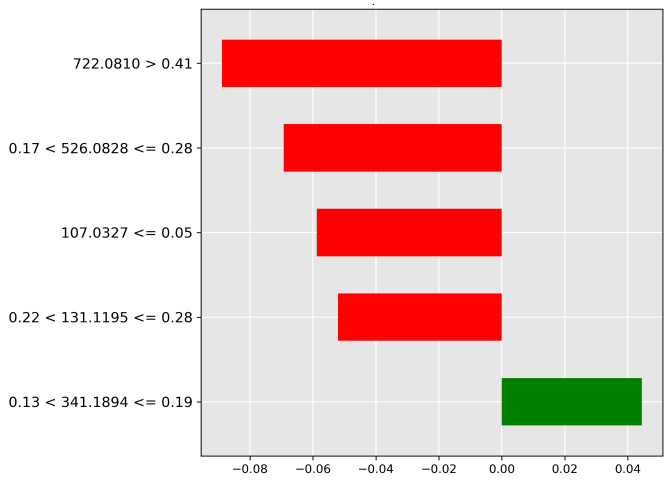


Fig. 9: Lime explanation for **transformer** for classification of fish part **fillet**.

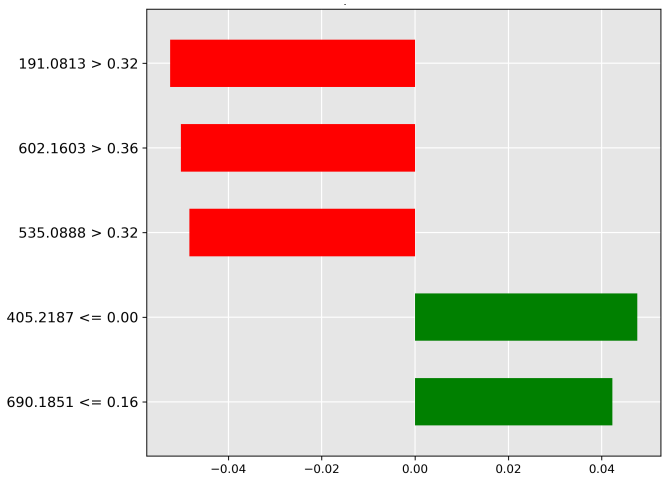


Fig. 11: Lime explanation for **transformer** for classification of fish part **skins**.

transformer for the fish parts classification for fish heads. The most important feature, and the strongest green bar, is when the mass-to-charge ratio 256.1089 m/z is within the normalized intensity range $0.26 < y \leq 0.35$. This indicates that large amounts of this molecule are likely present in fish heads.

Figure 9 gives the LIME explanation for the transformer for the fish body part of the fillet. The most important feature, and strongest red bar, is when the mass-to-charge ratio 722.0810 m/z is greater than the normalized intensity 0.41. This suggests that large amounts of this molecule are not expected in fish fillets.

Figure 10 gives the LIME explanation for the transformer for the fish body part of the liver. The most important feature, and strongest red bar, is when the mass-to-charge ratio 849.2039 m/z is within the normalized intensity range of $0.27 < y \leq 0.38$. This indicates that large amounts of this molecule are not likely to be found in fish liver.

Figure 11 gives the LIME explanation for the transformer for the fish body part of the skins. The most important feature, and the strongest red bar, is when the mass-to-charge ratio 191.0813 m/z is greater than the normalized intensity 0.32. This indicates that large amounts of this molecule are not usually found in fish skins.

Figure 12 gives the LIME explanation for the transformer for the fish body part of the guts. The most important feature, and the strongest red bar, is when the mass-to-charge ratio 675.1786 m/z is less than or equal to the normalized intensity 0.11. This suggests that small amounts of this molecule are not usually found fish guts.

Figure 13 gives the LIME explanation for the transformer for the fish body part of frames. The most important feature, and strongest green bar, is when the mass-to-charge ratio 533.161 m/z is within the normalized intensity range of $0.25 < y \leq 0.24$. This indicates that average to large amounts of this chemical are not expected to be found in fish frames.

Figure 14 gives the LIME explanation for the transformer for the fish body part of gonads. The most important feature, and strongest red bar, is when the mass-to-charge ratio 93.0882 m/z is less than or equal to the normalized intensity threshold of $0.09 < y \leq 0.18$. Biochemically, this could suggest that

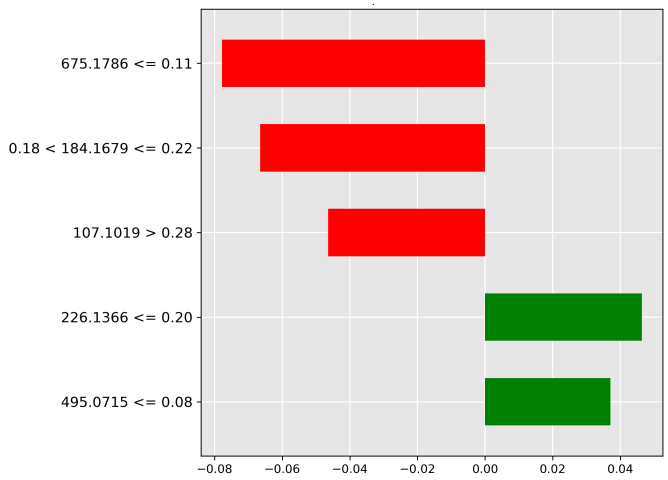


Fig. 12: Lime explanation for **transformer** for classification of fish part **guts**.

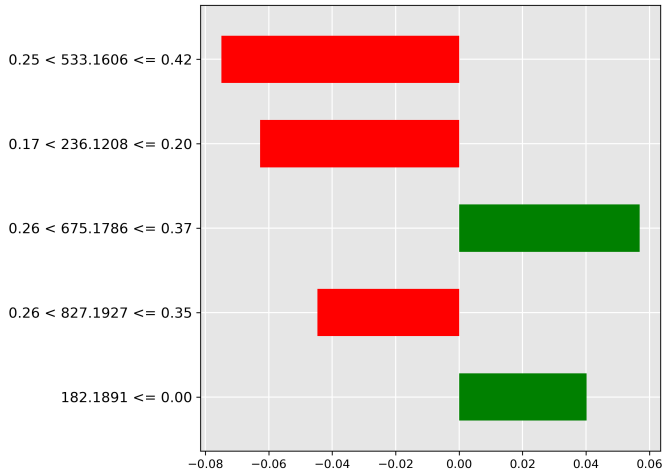


Fig. 13: Lime explanation for **transformer** for classification of fish part **frames**.

m/z 93.0882 might correspond to a compound not found in fish gonads.

VII. Conclusion and Future Work

The results from these classification tasks demonstrate that deep learning models, particularly Transformer is well-suited for handling the complex, high-dimensional data generated by REIMS data. These models consistently outperform traditional machine learning methods, especially for tasks involving subtle or overlapping signal differences, such as body parts detection. The pre-trained transformer outperforms the regular transformer on fish species classification, suggesting that pre-training captures meaningful embeddings that improve the performance of downstream classification tasks. While traditional models like decision trees show excellent performance in simpler tasks like fish species classification, their performance drops significantly for more challenging tasks, highlighting the need for advanced feature extraction and representation learning that deep learning models provide. The overall strong performance across the board suggests that REIMS

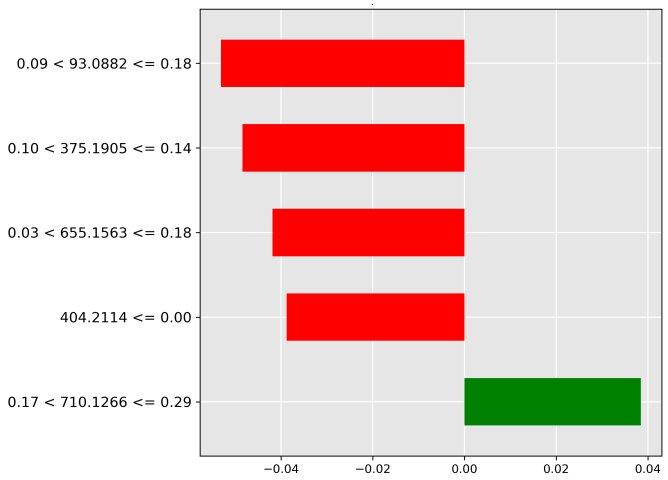


Fig. 14: Lime explanation for **transformer** for classification of fish part **gonads**.

data provides rich, discriminative information, particularly for fish species classification. However, body part identification requires more sophisticated modelling approaches, where deep learning shines due to its ability to capture complex patterns and subtle signal deviations.

The application of explainable AI techniques, i.e. LIME, 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 and body parts. For instance, the LIME analysis for fish speciation highlighted distinct spectral regions that differentiate Mackerel from Hoki. 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 automated, 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: (1) develop a system for real-time REIMS data acquisition and analysis, allowing for immediate classification results in industrial settings, and (2) work with regulatory bodies to ensure that the developed methods meet or exceed current standards for marine biomass analysis and food safety monitoring.

Declarations

A. *Availability of data and material*

The data used in this study is not publicly available due to confidentiality requirements.

B. Competing interests

The authors declare that they have no competing interests.

C. Funding

This work was supported by the MBIE Fund on Research Program under contract C11X2001.

D. Authors' contributions

JW conceived and implemented the methodology, conducted experiments, analyzed results, and wrote the manuscript. BN, BX, MZ and DK provided supervision, proofreading, and feedback on the research and manuscript. All authors read and approved the final manuscript.

E. Acknowledgements

The authors would like to thank Sue Marshall from Plant and Food Research for her leadership on this project.

REFERENCES

[1] 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.

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

[3] A. Ghaly, V. Ramakrishnan, M. Brooks, S. Budge, and D. Dave, "Fish processing wastes as a potential source of proteins," *Amino acids and oils: A critical review. J. Microb. Biochem. Technol*, vol. 5, no. 4, pp. 107–129, 2013.

[4] D. Pauly and D. Zeller, "Catch reconstructions reveal that global marine fisheries catches are higher than reported and declining," *Nature communications*, vol. 7, no. 1, p. 10244, 2016.

[5] Plant and F. Research, "New research to maximise value from seafood resources - plant & food research," https://www.plantandfood.com/en-nz/article/new-research-to-maximise-value-from-seafood-resources, 2020.

[6] M. Köppen, "The curse of dimensionality," in *5th online world conference on soft computing in industrial applications (WSC5)*, vol. 1, 2000, pp. 4–8.

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

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[10] 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.

[11] ——, "Genetic programming for multiple-feature construction on high-dimensional classification," *Pattern Recognition*, vol. 93, pp. 404–417, 2019.

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

[13] H. P. F. D. M. Australia, "Melbourne restaurant hunky dory accused of serving catfish to customers instead of dory," May 2016. [Online]. Available: https://www.dailymail.co.uk/news/article-3611999/Melbourne-restaurant-Hunky-Dory-accused-serving-catfish-customers-instead-dory.html

[14] A. P. Simopoulos, "Evolutionary aspects of diet: the omega-6/omega-3 ratio and the brain," *Molecular neurobiology*, vol. 44, no. 2, pp. 203–215, 2011.

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

[16] M. L. Panse and S. D. Phalke, "World market of omega-3 fatty acids," *Omega-3 Fatty Acids*, pp. 79–88, 2016.

[17] 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*. Springer, 2022, pp. 516–529.

[18] B. I. Bettjeman, K. A. Hofman, E. J. Burgess, N. B. Perry, and D. P. Killeen, "Seafood phospholipids: extraction efficiency and phosphorous nuclear magnetic resonance spectroscopy (31p nmr) profiles," *Journal of the American Oil Chemists' Society*, vol. 95, no. 7, pp. 779–786, 2018.

[19] 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.

[20] 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.

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

[22] M. Bylesjö, M. Rantalainen, O. Cloarec, J. K. Nicholson, E. Holmes, and J. Trygg, "Opls discriminant analysis: combining the strengths of pls-da and simca classification," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 20, no. 8-10, pp. 341–351, 2006.

[23] S. Balakrishnama and A. Ganapathiraju, "Linear discriminant analysis-a brief tutorial," *Institute for Signal and information Processing*, vol. 18, no. 1998, pp. 1–8, 1998.

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

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[27] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljačić, T. Y. Hou, and M. Tegmark, "Kan: Kolmogorov-arnold networks," *arXiv preprint arXiv:2404.19756*, 2024.

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

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

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

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

[32] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

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

[34] 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.

[35] L. Breiman, *Classification and regression trees*. Routledge, 2017.

[36] 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.

[37] D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, and M. Klein, *Logistic regression*. Springer, 2002.

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

[39] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE transactions on pattern analysis and machine intelligence*, vol. 12, no. 10, pp. 993–1001, 1990.

[40] J. Wang, T. Ji, Y. Wu, H. Yan, T. Gui, Q. Zhang, X. Huang, and X. Wang, "Length generalization of causal transformers without position encoding," *arXiv preprint arXiv:2404.12224*, 2024.

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

[42] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T. Liu, "On layer normalization in the trans-

former architecture," in *International Conference on Machine Learning*. PMLR, 2020, pp. 10 524–10 533.
[43] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.
[44] M. for Primary Industries, "Hoki: New zealand's largest fishery," https://www.mpi.govt.nz/fishing-aquaculture/fisheries-management/fish-stock-status/hoki-new-zealands-largest-fishery/, 2024.
[45] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[46] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.
[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[48] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.
[49] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.
[50] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.
[51] N. Morgan and H. Bourlard, "Generalization and parameter estimation in feedforward nets: Some experiments," *Advances in neural information processing systems*, vol. 2, 1989.
[52] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.
[53] M. T. Ribeiro, S. Singh, and C. Guestrin, ""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*, 2016, pp. 1135–1144.