An automated machine-learning approach for detecting cross-species and oil contamination 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*—The seafood processing industry faces critical challenges in quality control, including fraudulent species substitution and contamination risks. While marine biomass composition analysis can address these challenges, traditional methods require 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 rapid marine biomass analysis. Focusing on oil contamination and cross-species contamination detection, we investigate various machine learning methods, including unsupervised pre-training strategies for transformers. Our deep learning approaches outperformed traditional methods, achieving 91.97% accuracy in cross-species detection and 59.33% accuracy in oil contamination detection. Using Local Interpretable Model-agnostic Explanations (LIME), we identified key mass spectrometry features driving classification decisions. REIMS analysis with machine learning provides an accurate, rapid, and explainable technique for real-time marine biomass compositional analysis, with immediate applications in 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 seafood processing industry faces numerous challenges in maintaining product quality, safety, and authenticity throughout its complex supply chain. From harvesting to final packaging, fish products undergo multiple processing steps including cleaning, gutting, filleting, freezing, and packaging. Each stage presents opportunities for quality issues such as species substitution, contamination, or improper handling. A major concern is fraudulent practices like species substitution and adulteration, where high-value products are mixed with lower-quality biomass such as offal or cheaper species [1]. This not only deceives consumers but can also pose health risks. Studies have shown widespread mislabelling in seafood products [2], undermining consumer trust and industry integrity. Quality control is further complicated by potential contamination during processing. Oil contamination, whether from processing equipment or boat engine oil, can compromise product safety and quality. While the use of food-grade lubricants [3] helps mitigate risks during processing, effective detection methods remain crucial. Moreover, accurate species identification and stock monitoring are essential for both regulatory compliance and sustainable resource management. Processing facilities need reliable methods to verify species identity and maintain accurate counts of processed fish, supporting both quality control and conservation efforts [4]. To address these challenges, we investigate the application of advanced analytical and computational methods to seafood processing quality control. Specifically, we explore two critical classification tasks using Rapid Evaporative Ionization Mass Spectrometry (REIMS):

- Oil contamination detection to ensure product safety
- Cross-species contamination identification to prevent adulteration

These tasks represent key control points where automated analysis can significantly improve quality assurance. By applying machine learning techniques to REIMS data, we can provide rapid, accurate assessments that enhance product integrity throughout the processing pipeline. Our approach offers particular advantages in detecting cross-species adulteration [1], even at low concentrations, and maintaining accurate species-level inventory control in processing facilities.

This study analyzes mass spectrometry datasets from New Zealand Plant and Food Research's Cyber-Marine project, which develops quality assurance systems for marine biomass processing. Using Rapid Evaporative Ionization Mass Spectrometry (REIMS), an electrosurgical knife creates an aerosol from seafood samples that is analyzed by a mass spectrometer. Multiple 3-5 second incisions per sample generate detailed chemometric profiles spanning mass-to-charge ratios (m/z) from 77.04 - 999.32, providing rich biochemical fingerprints for quality assessment.

REIMS analysis of marine biomass faces several critical challenges. Traditional approaches require samples to be sent for offline analysis by chemistry experts, a process that conflicts with REIMS's inherent speed capabilities [5]. This mismatch between rapid data collection and slow analysis creates a bottleneck in fish processing facilities. The data itself presents additional challenges: each sample contains 2080 mass-to-charge ratios, creating a high-dimensional dataset [6], yet the expensive and time-consuming nature of sample preparation limits the number of training instances. Furthermore, the industrial setting demands models that are not only fast and accurate but also interpretable and easily validated in real-

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 Zeland

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

world conditions. To address these challenges, we aim to develop automated methods that can provide immediate "online" analysis directly on the production line, eliminating the need for offsite expert analysis while maintaining reliability.

To address these challenges, we propose an integrated machine learning approach combining deep learning [7], [8] and evolutionary computation [9], [10] techniques. Our solution leverages the power of transformers with attention mechanisms to capture complex, non-linear relationships in high-dimensional REIMS spectra, while genetic programming efficiently explores the feature space to identify crucial molecular signatures. To overcome the limited training data, we implement two complementary strategies: data augmentation to create synthetic training samples, and unsupervised pretraining to learn robust feature representations from unlabeled data before fine-tuning on the labeled dataset. For interpretability, we employ Local Interpretable Model-agnostic Explanations (LIME) [11], allowing domain experts to understand which molecular features drive model decisions. This comprehensive approach enables automated analysis while maintaining the transparency needed for real-world applications in fish processing facilities.

The main contributions of the paper are:

1) **Novel Pre-training Strategy for Mass Spectrometry Data:** The paper introduces a unique adaptation of BERT-style progressive masking for mass spectrometry data, treating mass-to-charge ratios as tokens. This represents the first application of NLP-inspired pre-training techniques to REIMS data, enabling better feature learning from limited samples.
2) **First Deep Learning Benchmark Suite for REIMS Analysis:** The study provides the first comprehensive evaluation of modern deep learning architectures (Transformer, LSTM, VAE, KAN, CNN, Mamba) against traditional methods (OPLS-DA) for REIMS data analysis. Results demonstrate significant improvements over the industry standard OPLS-DA method, with deep learning achieving 91.97% accuracy in cross-species detection compared to OPLS-DA's 79.96%.
3) **Chemical-Interpretable Deep Learning:** The research bridges the gap between deep learning and analytical chemistry by mapping LIME explanations to specific mass-to-charge ratios, providing chemically meaningful interpretations of model decisions. This allows domain experts to validate model predictions based on known biochemical markers.

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. The relevant background required for the deep learning and evolutionary computation is presented here too.

A. Marine Biomass

Quality control in seafood processing faces two critical challenges: species authentication and contamination detection. Mislabelling is particularly prevalent, with genomic profiling studies revealing an average global seafood mislabelling rate of 30% [2]. High-profile cases, such as the 2016 Melbourne restaurant serving catfish as dory [12], highlight the industry-wide impact of species substitution fraud. Contamination presents another significant challenge. Cross-species contamination, where high-value products are adulterated with cheaper species or offcuts, undermines product integrity and can pose health risks. Oil contamination occurs through multiple pathways: from processing equipment despite the use of food-grade lubricants [3], from boat engine oil during harvesting, and from non-food-grade machinery during processing. While food-grade lubricants help mitigate some risks, comprehensive detection methods are needed for all contamination sources. Rapid Evaporative Ionization Mass Spectrometry (REIMS) combined with machine learning offers promising solutions to both challenges. REIMS analysis can detect species substitution in both raw and cooked biomass, providing rapid, accurate authentication. For contamination detection, REIMS's ability to generate detailed biochemical profiles enables the identification of both cross-species adulteration and oil contamination, enhancing existing quality control procedures.

B. REIMS

The seafood industry has traditionally relied on several established analytical methods for marine biomass composition analysis. These include Gas Chromatography-Mass Spectroscopy [13], Nuclear Magnetic Resonance Spectroscopy [14], and Genomic Profiling [2]. While effective, these methods face significant practical limitations: they require extensive preparation time, demand specialized technical expertise, and involve complex sample processing procedures. Such constraints have driven the search for faster, more accessible analytical techniques in fish processing. Rapid Evaporative Ionization Mass Spectrometry (REIMS) has emerged as a breakthrough technology addressing these limitations. First introduced by Balog et al. [15], REIMS has demonstrated remarkable capabilities across diverse applications. Notable achievements include detecting horse offal in beef mince at concentrations as low as 1%-5% [4], and successfully identifying fish species and catch methods to prevent fraud [1]. The conventional approach to REIMS data analysis has centered on Orthogonal Partial Least Squares Discriminant Analysis (OPLS-DA) [16]–[18], coupled with Principal Component Analysis (PCA) for dimensionality reduction [19]. However, this PCA-OPLS-DA methodology has notable drawbacks, particularly its dependence on manually-configured outlier thresholds for adulteration detection, requiring extensive chemistry expertise. Our work introduces two key advances: automated methods with self-learning parameters that eliminate the need for chemical domain expertise, and novel deep learning and evolutionary computation approaches that surpass the performance of traditional OPLS-DA methods.

C. Deep Learning

The complexity and high dimensionality of Rapid Evaporative Ionization Mass Spectrometry (REIMS) data require sophisticated models capable of handling sequential dependencies and intricate spectral patterns. We investigate several architectures particularly suited to these challenges: Transformers [7], [8] leverage self-attention mechanisms to weigh the importance of different mass-to-charge ratios in the spectra, similar to their success in processing sequential data in language tasks. This makes them particularly effective at identifying critical regions in REIMS spectra for classification. Long Short-Term Memory (LSTM) networks [20] excel at capturing long-range dependencies across mass-to-charge values, essential for detecting complex spectral patterns that may span wide ranges in the data. Their ability to maintain and utilize information over long sequences is crucial for identifying subtle contamination signatures. Variational Autoencoders (VAE) [21] learn compressed representations of REIMS spectra, making them valuable for detecting anomalies and cross-species contamination through their ability to model and reconstruct complex spectral distributions. For handling non-linear patterns, we employ Kolmogorov-Arnold Networks (KAN) [22], which efficiently approximate complex functions in the spectral data that might indicate species differences or contamination levels. Convolutional Neural Networks (CNN) [23] exploit local dependencies in the spectra, treating them as one-dimensional signals with important neighboring relationships. Finally, the Mamba architecture [24] provides an efficient state-space approach to processing sequential REIMS data, offering a balance between computational performance and modeling capability crucial for real-time analysis.

D. Evolutionary Computation

Genetic programming (GP), pioneered by Koza [25], provides a powerful framework for automated feature engineering through simulated evolution. In GP, solutions evolve as tree structures that undergo crossover and mutation operations to optimize their performance on specific tasks. This approach has proven particularly valuable for handling complex, high-dimensional classification problems. Recent advances by Tran et al. [9], [10] have enhanced GP's capabilities through Multiple Class-Independent Feature Construction (MCIFC). This method employs a multi-tree representation where each tree evolves features specific to different classes, enabling simultaneous feature construction and selection. MCIFC's ability to learn class-specific transformations makes it especially suitable for REIMS data analysis, where different species or contamination types may exhibit distinct spectral patterns. For marine biomass analysis, MCIFC offers unique advantages in capturing non-linear relationships between mass-to-charge ratios that characterize different fish species or identify contamination. By constructing features tailored to each class, MCIFC can potentially identify subtle spectral signatures that might be overlooked by traditional feature selection methods.

III. Methods

Having established the background, let's examine the core analytical approaches used to extract meaningful insights from REIMS spectra.

- **Traditional machine learning methods:** Random Forest (RF) [26], K-Nearest Neighbors (KNN) [27], Decision Trees (DT) [28], Naive Bayes (NB) [29], Logistic Regression (LR) [30], Support Vector Machines (SVM) [31], and Linear Discriminant Analysis (LDA) [18].
- **Ensemble method:** [32]: A combination of the above traditional methods.
- **Benchmark method:** Orthogonal Partial Least Squares Disrciminant Analysis (OPLS-DA) [17].
- **Deep neural networks:** Transformer [7], [8], Long Short-Term Memory (LSTM) [20], Variational Autoencoder (VAE) [21], Convolutional Neural Network (CNN) [23], [33]–[35], Kolmogorov-Arnold Networks (KAN) [22] and Mamba [24].
- **Evolutionary computation:** Multiple Class Independent Feature Construction (MCIFC) [9], [10]

A. Transformer

1) Architecture: Our approach adapts the transformer architecture [8] for mass spectrometry data analysis. While transformers traditionally process sequential data using self-attention mechanisms, we modify the architecture to handle REIMS spectral data without positional embeddings [36], as mass-to-charge ratio positions are inherently meaningful. The model employs an encoder-only structure with stacked layers, where each layer consists of multi-head self-attention followed by feed-forward networks.

As shown in Figure 1, we incorporate residual connections [37] around each encoder block. These connections create direct paths for gradient flow during training, enabling deeper networks while preventing vanishing and exploding gradients. This design proves particularly effective for REIMS data, where detecting subtle spectral patterns requires deep architectural features.

A notable aspect of the transformer architecture used in this work is the choice of pre-norm layer normalization [38], [39], 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 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

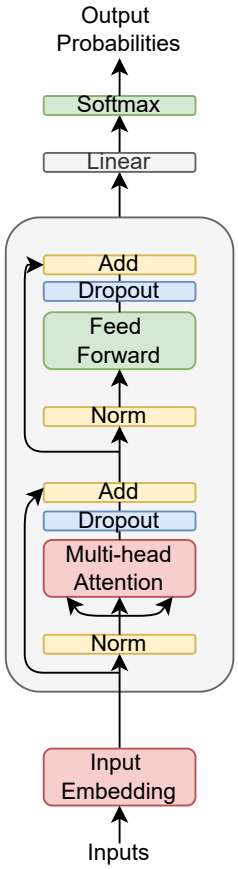


Fig. 1: Transformer Architecture.

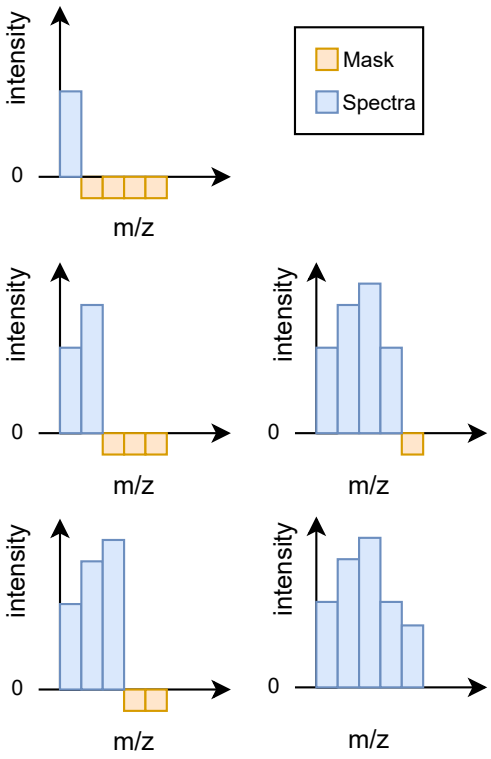


Fig. 2: Masked language modelling.

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 Rapid Evaporative Ionization Mass Spectrometry 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: This paper adopts unsupervised pre-training inspired by BERT [7] 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.

B. Evolutionary Computation

Complementing the deep learning approach of transformers, we now explore multi-tree genetic programming as an alternative paradigm for feature construction and classification in

marine biomass analysis using REIMS. This approach, known as multiple class independent feature construction (MCIFC) [10], offers a novel application in this domain. Here we extend MCIFC for classification with a winner-takes-all strategy. The MCIFC algorithm, detailed in Algorithm 1, 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. In our cross-species contamination detection classification task, MCIFC uses a three-tree representation for the three classes (Hoki, Mackerel and Hoki-Mackerel mix), while the oil contamination detection task utilizes a seven-tree representation for its seven classes (50%, 25%, 10%, 5%, 1%, 0.1% and 0.0%).

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 $\leftarrow$ the best e individuals;
while Maximum generation is not reached **do**
  for $i = 1$ to Population Size **do**
    $transf\_train \leftarrow$ Calculate constructed features of individual i on train_set;
    $fitness \leftarrow$ 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;

The genetic operators in MCIFC [9], [10] include crossover (80% probability) and mutation (20% probability), adapted from conventional genetic programming [25]. Genetic programming, popularized by John Koza in 1994 [25], is a type of evolutionary algorithm that simulates the process of natural selection to solve problems by evolving computer programs. It operates by generating a population of potential solutions, each represented as a tree structure. Over successive generations, these solutions are refined through genetic operations such as mutation and crossover. Mutation involves making small, random changes to parts of an individual solution to explore new possibilities. Crossover combines portions of two parent solutions to produce offspring. These processes, along with selection mechanisms, guide the population toward better solutions over time. Tournament selection with a size of 7 is used for parent selection. Importantly, crossover only occurs between trees of the same class, while mutation randomly alters one subtree. The fitness evaluation in MCIFC combines accuracy with a distance term:

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

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

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

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

Here, $d(i,j)$ represents the Euclidean distance between two instances i and j, and $|S|$ is the total number of pairs of instance. This approach aims to maximize interclass distances while minimizing intraclass distances, thereby promoting better class separation in the constructed feature space.

IV. Experimental Setup

Having detailed our methodology, we now describe the experimental framework used to evaluate these approaches. We compare a comprehensive range of machine learning techniques, from traditional methods, advanced deep learning architectures and evolutionary computation methods, assessing their ability to analyze REIMS data.

A. Benchmark Technique

To evaluate the performance of the proposed methods, OPLS-DA is used as a benchmark to compare new approaches to the existing methods for REIMS analysis. Each method is evaluated and the average is given over 30 independent runs. Stratified k-fold cross-validation with $k = 5$ 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 five-fold cross-validation, the model is tested across various scenarios, improving its generalization to unseen data and providing a comprehensive evaluation of its performance.

B. Datasets

Following our introduction to the challenges and potential of REIMS-based analysis, this section now focuses on the critical foundation of our study: the dataset. 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 [40].

The dataset used consists of REIMS spectra collected from two fish species and six body parts. Particularly, we will have two different datasets corresponding to two different (following) tasks:

1) **Oil contamination:** A dataset with 126 samples, with 14.28% belonging to each class (50%, 25%, 10%, 5%, 1%, 0.1%, 0%). Oil contamination detection is crucial in fish processing as contamination can occur at multiple stages: from boat engine oil during harvesting, processing equipment lubricants, or mechanical systems in the factory. The seven concentration levels were chosen to cover the full range from heavy contamination (50%) to trace amounts (0.1%), allowing the model to detect contamination across all practically relevant scenarios. Each class has equal representation to ensure the model learns to distinguish between different contamination levels with equal importance.
2) **Cross-species contamination:** The dataset contains 144 samples, with 39.21% Hoki, 31.37% Mackerel and 29.41% mixed. Cross-species contamination detection addresses a critical food fraud issue where high-value fish products are diluted with cheaper species for economic gain. The dataset focuses on Hoki and Mackerel as they represent different price points in New Zealand's seafood industry, with Hoki being the country's largest and most valuable fishery. The slightly higher proportion of pure samples (70.58% combined) versus mixed samples (29.41%) reflects real-world conditions where adulteration is less common than authentic products, helping the model learn realistic detection scenarios.

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

C. Parameter Settings

Experiments use the default settings from sklearn [41], 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 mode. 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 [42] decouples weight decay from the learning rate, an improvement over the popular Adam optimizer [43]. Dropout [44] turns off neurons at random during training to efficiently approximate a bagged ensemble of sub-neural networks. Label smoothing [45] 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) [46] for activation functions. Data augmentation replicates each instance five times and injects noise [47] expanding the training set five-fold - data augmentation inflates the number of training instances. Early stopping [48] 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 [49]. 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: NSP & Speciation	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 II: MCFIC parameter settings

Function Set	$+, -, \times$, protectedDiv
Terminal Set	$x_1, x_2, ..., x_n$
Maximum Tree Depth	6
Population size	1 * 1023 ($= 1\times$ #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

D. Transformer parameters

Table I gives the configuration of hyperparameters for the transformer.

E. Genetic Programming parameters

Table II provides a comprehensive overview of the parameter settings used in the MCIFC method, including the construction ratio, which denotes the number of trees per class.

V. Results and Discussion

After implementing our classification strategies, we now present and analyze how these various machine learning techniques performed on the REIMS datasets.

Table III and table IV give 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.

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.

TABLE III: Classification results for oil contamination detection

Method	Train	Test
KNN	55.30% ± 0.00%	25.00% ± 0.00%
DT	100.00% ± 0.00%	23.40% ± 2.06%
LR	100.00% ± 0.00%	20.89% ± 0.00%
LDA	70.07% ± 0.00%	22.86% ± 0.00%
NB	65.48% ± 0.00%	25.54% ± 0.00%
RF	100.00% ± 0.00%	32.45% ± 2.32%
SVM	100.00% ± 0.00%	20.54% ± 0.00%
Ensemble	100.00% ± 0.00%	28.13% ± 1.00%
OPLS-DA	86.48% ± 6.31%	26.43% ± 6.00%
Transformer	100.00% ± 0.00%	50.95% ± 5.90%
Pre-trained	100.00% ± 0.00%	*56.32% ± 6.62%*
LSTM	100.00% ± 0.00%	53.87% ± 11.83%
VAE	60.90% ± 11.49%	54.70% ± 7.15%
KAN	100.00% ± 0.00%	**59.33% ± 10.92%**
CNN	100.00% ± 0.00%	55.75% ± 6.56%
Mamba	100.00% ± 0.00%	56.29% ± 7.94%
MCIFC	54.06% ± 4.80%	18.33% ± 8.33%

TABLE IV: Classification results for cross-species identification.

Method	Train	Test
KNN	81.50% ± 0.00%	59.38% ± 0.00%
DT	100.00% ± 0.00%	63.51% ± 1.72%
LR	100.00% ± 0.00%	70.82% ± 0.00%
LDA	100.00% ± 0.00%	70.82% ± 0.00%
NB	72.12% ± 0.00%	50.35% ± 0.00%
RF	100.00% ± 0.00%	63.97% ± 2.13%
SVM	100.00% ± 0.00%	69.51% ± 0.00%
Ensemble	100.00% ± 0.00%	68.76% ± 1.24%
OPLS-DA	100.00% ± 0.00%	79.96% ± 7.50%
Transformer	100.00% ± 0.00%	90.91% ± 5.65%
Pre-trained	100.00% ± 0.00%	**91.97% ± 4.55%**
LSTM	100.00% ± 0.00%	*91.62% ± 5.74%*
VAE	93.31% ± 7.92%	89.84% ± 4.98%
KAN	100.00% ± 0.00%	90.92% ± 6.62%
CNN	100.00% ± 0.00%	83.03% ± 5.48%
Mamba	100.00% ± 0.00%	90.38% ± 6.92%
MCIFC	82.92% ± 4.45%	65.85% ± 13.66%

A. Oil Contamination

The Kolmogorov-Arnold Network (KAN) performed best for oil contamination detection with an accuracy of 59.33%, followed by the pre-trained Transformer at 56.32%. These models likely excel at this task because oil contamination introduces subtle and nuanced changes in the ionization patterns that require models capable of detecting minor anomalies or deviations in the data. By leveraging the Kolmogorov-Arnold theorem, KAN provides a mathematical guarantee that any continuous function can be approximated by this network structure. This is a powerful property, making KANs an effective universal approximator for complex, high-dimensional functions.

Traditional machine learning models struggle with this task, with accuracies ranging from 20.89% for Logistic Regression to 32.45% for Random Forest. This poor performance indicates that the signal variations introduced by oil contamination are not easily captured using simple, linear decision boundaries or tree-based splits. The complexity of the task requires more sophisticated feature extraction, something deep learning models like KAN and Transformer are better equipped to handle.

The KAN model (59.33%) outperforms the OPLS-DA method (26.43%) by a significant margin for oil contamination multi-class classification. Notably, deep learning methods have bested the traditional approach in the field of REIMS analysis.

B. Cross-species Contamination

The pre-trained transformer achieves the highest test accuracy (91.97%) on the cross-species contamination dataset. The self-attention mechanism in transformers allows each data point (or peak) in the mass spectrum to attend to all other points. This global context awareness is particularly useful in mass spectrometry, where relationships between peaks across the entire spectrum can be important for identification and quantification. The LSTM is the second-best performing model (91.62%). Mass spectrometry data is inherently sequential, with peaks occurring at different mass-to-charge ratios (m/z). LSTMs are designed to handle sequential data effectively.

The lower accuracy of this task compared to fish species classification [50] suggests that mixed-species samples introduce additional complexity, making it harder to classify. Cross-species contamination likely involves overlapping patterns from multiple species, which traditional machine learning models find difficult to distinguish, requiring advanced models that can disentangle complex, multi-source signals.

The pre-trained transformer (91.97%) outperforms the OPLS-DA method (79.96%) for the task of cross-species contamination. For this second task, deep learning methods exceed the traditional approach of REIMS analysis used in the literature.

C. Summary:

Our experiments reveal three significant findings about machine learning approaches for REIMS data analysis. First, deep learning methods consistently outperform the traditional OPLS-DA method [1], [4], [5], [15], with the pre-trained transformer achieving 91.97% accuracy in cross-species detection and KAN reaching 59.33% accuracy in oil contamination detection. This represents a substantial improvement over OPLS-DA's 79.96% and 26.43% respectively, suggesting that deep learning should be considered the new standard for REIMS analysis. Second, different architectures show task-specific strengths. The transformer architecture excels at cross-species detection, likely due to its ability to capture long-range dependencies in mass spectra, while KAN proves more effective for oil contamination detection through its enhanced function approximation capabilities. Third, our results highlight areas requiring further development. While cross-species detection achieves high accuracy, oil contamination detection remains challenging. This performance gap suggests that detecting subtle chemical changes from oil contamination requires more sophisticated approaches than identifying distinct species-specific molecular signatures.

VI. Further Analysis on the Explainability

While our deep learning models achieve high accuracy, their practical adoption in fish processing facilities requires interpretability. Domain experts in chemistry and fish processing need to understand and validate model decisions based

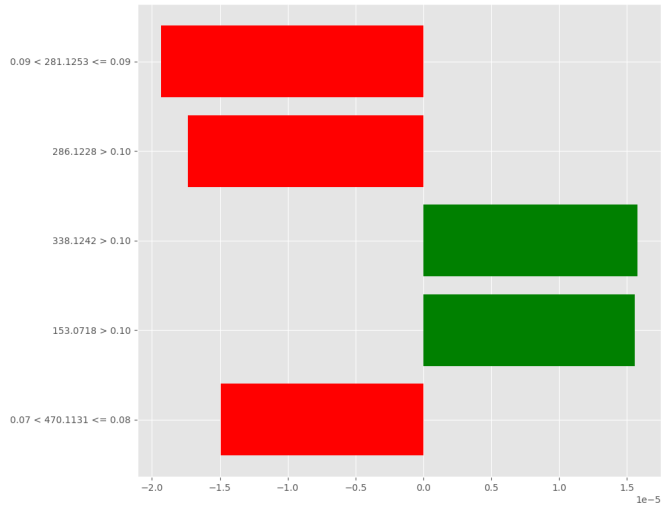


Fig. 3: LIME explanation for KAN for classification of oil contamination in 50% concentration.

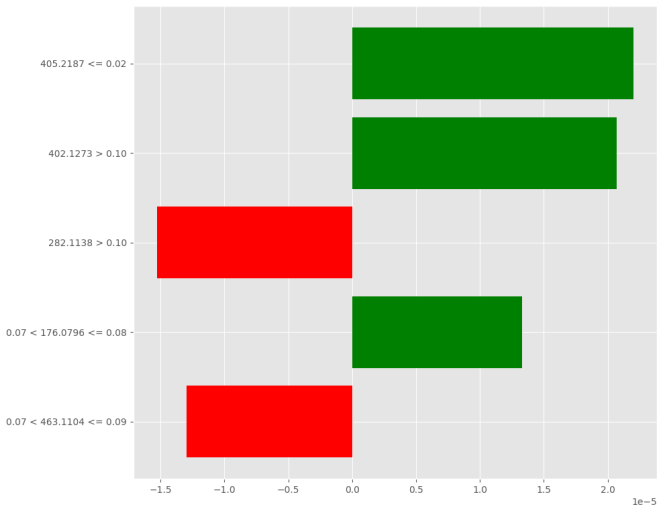


Fig. 4: LIME explanation for KAN for classification of oil contamination in 25% concentration.

on known molecular markers. To achieve this transparency, we employ Local Interpretable Model-agnostic Explanations (LIME) [51]. LIME reveals how our models use specific mass spectrometry features (mass-to-charge ratios) to make classifications. For each prediction, LIME creates multiple variations of the input by perturbing feature values and observes how these changes affect the model's output. This sampling process helps construct a simpler, interpretable model that approximates our complex model's behavior in the local region around the prediction of interest. We visualize these LIME explanations as bar charts showing the five most influential mass-to-charge ratios for each classification:

- Green bars indicate features whose presence supports the predicted class.
- Red bars show features whose presence contradicts the predicted class.
- Bar length represents the strength of each feature's influence.
- The y-axis specifies the mass-to-charge ratio (m/z) and intensity threshold values.

This approach lets chemists verify if the model bases its decisions on chemically meaningful markers. For example, when detecting oil contamination, they can confirm if the model identifies mass-to-charge ratios associated with known oil compounds.

A. Oil Contamination

The KAN model achieves the best classification accuracy (59.33%) for oil contamination. Figure 3 gives the LIME explanation for the KAN for oil contamination when the concentration is 50%. The most important feature, and strongest red bar, is when the mass-to-charge ratio 281.1253 m/z is within the normalized intensity range $0.09 < y \leq 0.09$. This suggests that average amounts of this molecule indicate a sample is not contaminated with large amounts of oil.

Figure 4 gives the LIME explanation for the KAN for oil contamination when the concentration is 25%. The most important feature, and strongest green bar, is when the mass-to-charge ratio 405.2187 m/z is less than or equal to the normalized intensity threshold of 0.02. This indicates that this molecule is a likely indicator of small quantities of oil present.

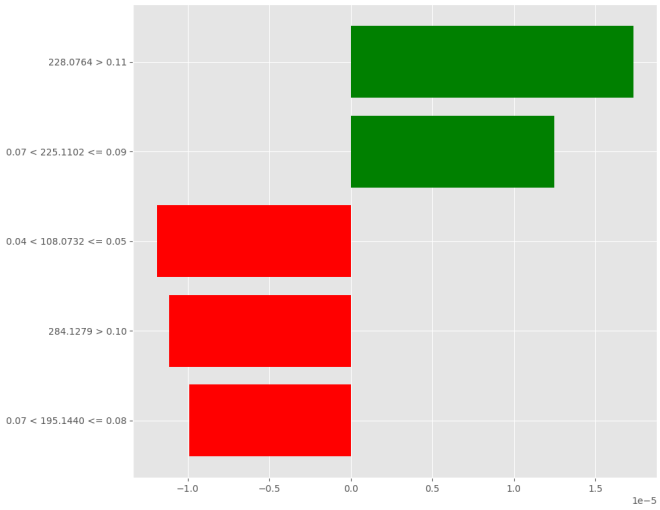


Fig. 5: LIME explanation for KAN for classification of oil contamination in 10% concentration.

Figure 5 gives the LIME explanation for the KAN for oil contamination in concentrations of 10%. The most important feature, and strongest green bar, is when the mass-to-charge ratio 228.0764 m/z exceeds the normalized intensity threshold of 0.11. This indicates that this molecule is likely associated with small amounts of oil present.

Figure 6 gives the LIME explanation for the KAN for oil contamination in concentrations of 5%. The most important feature, and strongest red bar, is when the mass-to-charge ratio 277.1248 m/z is greater than or equal to the normalized intensity threshold of 0.07. This indicates that average to large amounts of this molecule indicate that minor oil contamination is present.

Figure 7 gives the LIME explanation for the KAN for oil

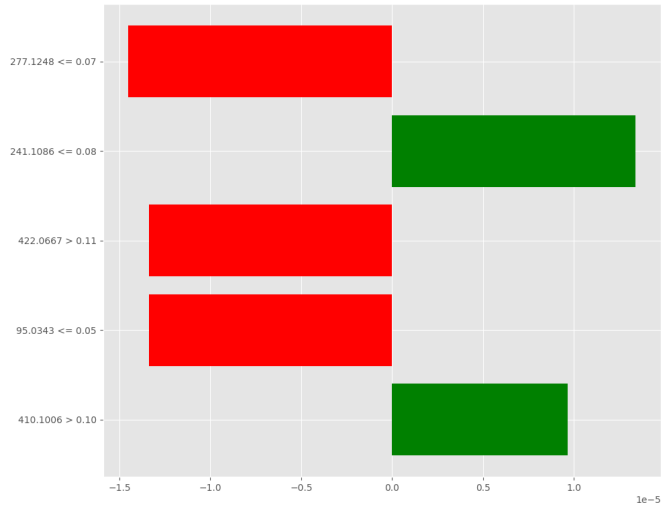


Fig. 6: LIME explanation for KAN for classification of oil contamination in 5% concentration.

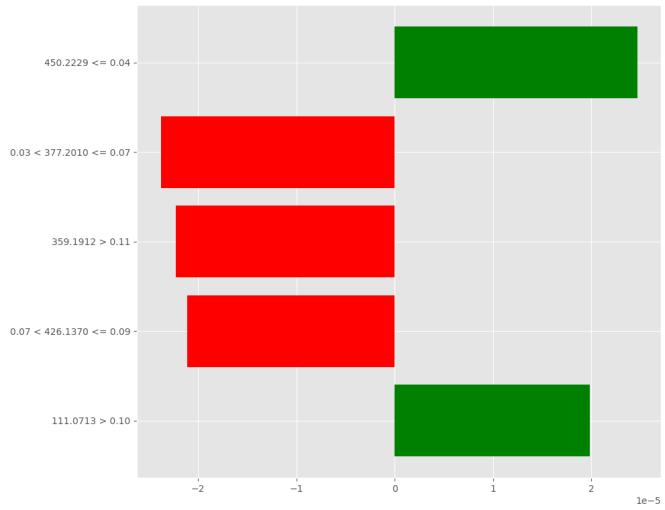


Fig. 7: LIME explanation for KAN for classification of oil contamination in 1% concentration.

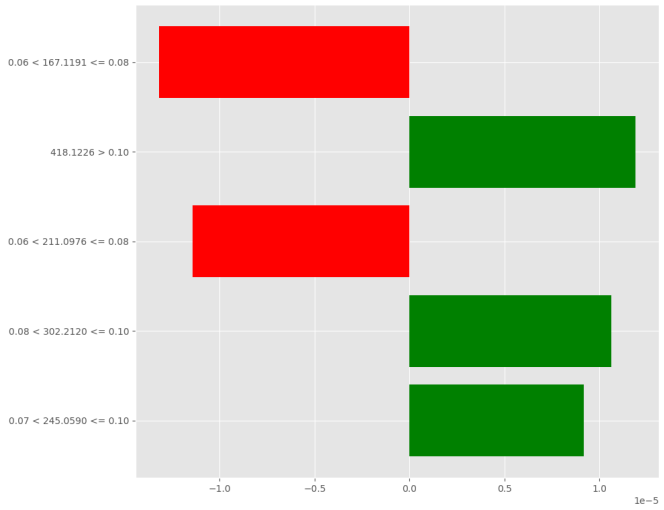


Fig. 8: LIME explanation for KAN for classification of oil contamination in 0.1% concentration.

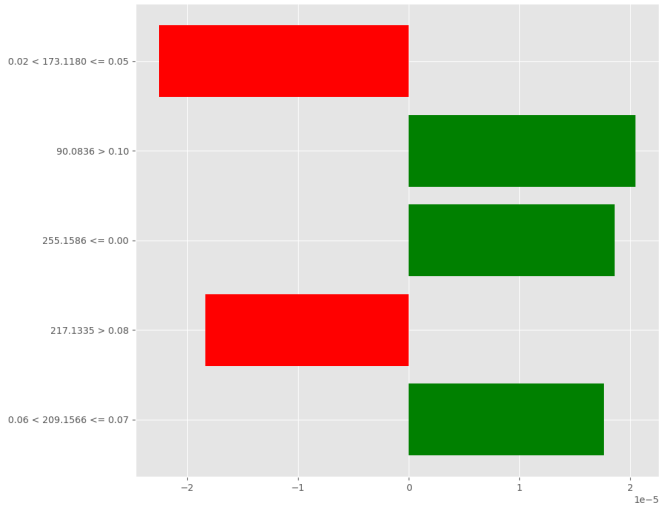


Fig. 9: LIME explanation for KAN for classification of oil contamination in 0% concentration.

contamination in concentrations of 1%. The most important feature, and strongest green bar, is when the mass-to-charge ratio 450.2229 m/z is less than or equal to the normalized intensity threshold of 0.04. This indicates that small amounts of this molecule indicate very small amounts of oil contamination present.

Figure 8 gives the LIME explanation for the KAN for oil contamination in concentrations of 0.1%. The most important feature, and strongest red bar, is when the mass-to-charge ratio 167.1191 m/z is within the normalized intensity range $0.06 < yle 0.08$. Indicating that average amounts of this molecule indicate a sample is not likely contaminated with very little oil. Therefore, this molecule is either a sign of no contamination or plenty of contamination.

Figure 9 gives the LIME explanation for the KAN for oil contamination in concentrations of 0.0%, or none. The most important feature, and largest red bar, is when the mass-to-charge ratio 173.1180 m/z is within the normalized intensity range $0.02 < 173.1180 \leq 0.05$. This indicates that average amounts of this molecule indicate that a sample is contaminated by oil. This molecule likely corresponds to the oil itself.

B. Cross-Species Contamination

The transformer model performs the best (91.97%) on the cross-species contamination dataset.

Figure 10 gives the LIME explanation for the pre-trained transformer for detecting cross-species contamination for the class of Hoki-Mackerel mix. The most important feature, and strongest red bar, is when the mass-to-charge ratio of 251.1667 m/z is less than or equal to the normalized intensity threshold of 0.03. This indicates this molecule is a strong indicator of no cross-species contamination, and likely belongs solely to the Hoki or Mackerel fish species.

Figure 10 gives the LIME explanation for the pre-trained transformer for detecting cross-species contamination for the

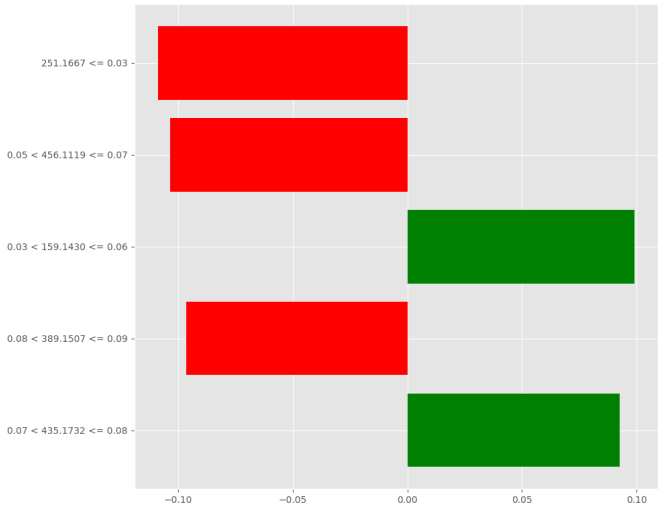


Fig. 10: LIME explanation for pre-trained transformer for cross-species contamination of Hoki-Mackerel class.

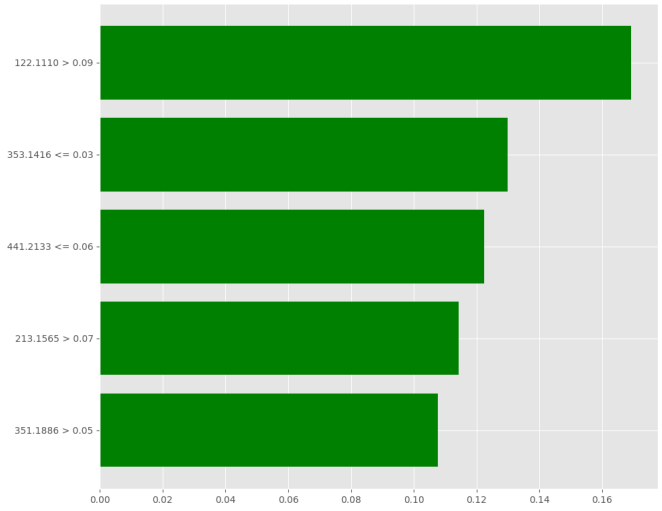


Fig. 11: LIME explanation for pre-trained transformer for cross-species contamination of Hoki class.

class of Hoki. The most important feature, and strongest green bar, is when the mass-to-charge ratio 122.1110 m/z exceeds the normalized intensity threshold of 0.09. This indicates that large amounts of this molecule are likely abundant in Hoki, and Hoki alone.

Figure 10 gives the LIME explanation for the pre-trained transformer for detecting cross-species contamination for the class of Mackerel. The most important feature, and strongest red bar, is when the mass-to-charge ratio of 367.0835 m/z is less than or equal to the normalized intensity threshold of 0.07. This indicates that this molecule is not likely found in Mackerel at all, and its presence indicates the sample is either a Hoki or Hoki-Mackerel mix.

VII. Conclusion and Future Work

Our research demonstrates that deep learning models can revolutionize REIMS-based marine biomass analysis. The

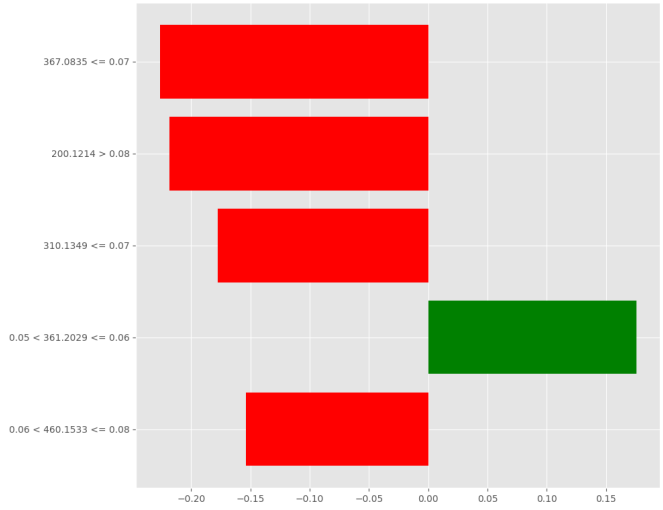


Fig. 12: LIME explanation for pre-trained transformer for cross-species contamination of Mackerel class.

pre-trained transformer achieved 91.97% accuracy in cross-species detection and KAN achieved 59.33% accuracy in oil contamination detection, both significantly outperforming the traditional OPLS-DA approach used in industry. Three key findings emerge from this work: First, pre-training strategies adapted from natural language processing prove remarkably effective for mass spectrometry data, suggesting that self-supervised learning could address the limited training data challenge in analytical chemistry. Second, deep learning models consistently outperform traditional methods in detecting subtle contamination patterns, with transformers excelling at cross-species detection and KAN at oil contamination detection. Third, LIME analysis reveals that our models base their decisions on chemically meaningful mass-to-charge ratios, providing interpretability that aligns with domain expertise. These results establish a foundation for automated, real-time quality control in fish processing facilities, replacing time-consuming manual analysis while maintaining reliability and transparency. The combination of high accuracy and chemical interpretability suggests that deep learning could become a standard tool in analytical chemistry, particularly for rapid screening applications requiring immediate results.

While our study has yielded promising results, several important areas warrant further investigation: (1) Transfer Learning Across Species - Investigate whether models trained on Hoki and Mackerel can transfer their learned features to detect contamination in other fish species, potentially reducing the need for extensive new training data. (2) Multi-modal Integration - Combine REIMS data with other analytical techniques (e.g., Near-infrared spectroscopy or DNA barcoding) to create more robust classification systems that leverage complementary information sources. This could improve accuracy in challenging cases where single-modality analysis is insufficient.

ACKNOWLEDGMENT

The authors would like to thank project leader Sue Marshall from Plant and Food Research. This work is supported in part MBIE Fund on Research Program under the contract of C11X2001.

REFERENCES

[1] 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.

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

[3] E. Moens, "Production and use of food-grade lubricants," *Trends in Food Science & Technology*, vol. 14, no. 4, pp. 157–162, 2003.

[4] 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.

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

[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] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

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

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

[22] 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.

[23] 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.

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

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

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

[27] 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.

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

[29] D. J. Hand and K. Yu, "Idiot's bayes—not so stupid after all?" *International statistical review*, vol. 69, no. 3, pp. 385–398, 2001.

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

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

[32] 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.

[33] 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.

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

[35] 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.

[36] 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.

[37] 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.

[38] 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 transformer architecture," in *International Conference on Machine Learning*. PMLR, 2020, pp. 10 524–10 533.

[39] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[40] 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.

[41] 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.

[42] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[44] 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.

[45] 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.

[46] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[47] J. Sietsma and R. J. Dow, "Creating artificial neural networks that generalize," *Neural networks*, vol. 4, no. 1, pp. 67–79, 1991.

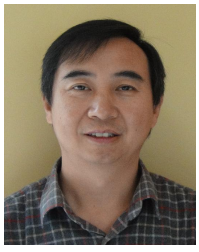
[48] N. Morgan and H. Bourlard, "Generalization and parameter estimation in feedforward nets: Some experiments," *Advances in neural information processing systems*, vol. 2, 1989.

[49] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.
[50] J. Wood, B. H. Nguyen, B. Xue, M. Zhang, and D. Killeen, "A rapid machine-learning approach for detecting fish species and body parts using rapid evaporative ionisation mass spectrometry," in *Australasian Joint Conference on Artificial Intelligence*. Springer, 2024, pp. 516–529.
[51] 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.

Jesse Wood (M'23) received his BE with First Class Honours in software engineering at Victoria University of Wellington (VUW), New Zealand in 2023. He is currently a PhD student in the Center of Data Science and Artificial Intelligence & the School of Engineering and Computer Science at Victoria University of Wellington. His research interests are in deep learning, evolutionary computation, transfer learning, explainable AI, AI applications and interdisciplinary AI. He is currently the webmaster of the IEEE New Zealand Central Section.

Bach Hoai Nguyen (M'14) received his B.Sc. with First Class Honours and the PhD in computer science at Victoria University of Wellington (VUW), New Zealand (NZ) in 2015 and 2018, respectively. He is currently a Lecturer in the Centre of Data Science and Artificial Intelligence & the School of Engineering and Computer Science at Victoria University of Wellington. His research interests are in evolutionary computation, feature selection, feature construction, and transfer learning. He is currently the Chair of the IEEE Task Force on Evolutionary Feature Selection and Construction, and a member of the IEEE CIS.

Bing Xue (M'10-SM'21-F'23) is currently a Professor of Artificial Intelligence, Deputy Director of Centre for Data Science and Artificial Intelligence, and Deputy Head of School in the School of Engineering and Computer Science at Victoria University of Wellington. She has over 300 papers published in fully refereed international journals and conferences and her research focuses mainly on evolutionary computation and machine learning. Dr Xue is currently the Chair of IEEE CIS Evolutionary Computation Technical Committee and Editor of IEEE CIS Newsletter. She has also served as associate editor of several international journals, such as IEEE CIM, IEEE TEVC and ACM TELO. She is also a Fellow of Engineering New Zealand.



Mengjie Zhang (M'04-SM'10-F'19) is currently a Professor of Computer Science, the Director of Centre for Data Science and Artificial Intelligence, Victoria University of Wellington, New Zealand. His current research interests include genetic programming, image analysis, feature selection and reduction, job-shop scheduling, and evolutionary deep learning and transfer learning. He has published over 800 research papers in refereed international journals and conferences. He is a Fellow of the Royal Society of New Zealand, a Fellow of Engineering New Zealand, a Fellow of IEEE, and an IEEE Distinguished Lecturer.

Daniel Killeen Biography text here