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

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

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

Abstract. Marine biomass compositional analysis traditionally requires time-consuming processes and domain expertise. This study demonstrates the effectiveness of Rapid Evaporative Ionisation Mass Spectrometry (REIMS) combined with advanced machine learning techniques for rapid and accurate marine biomass composition determination. Using fish species, body parts, oil contamination, and cross-species contamination as model systems representing diverse biochemical profiles, the paper employed various machine learning methods, including decision trees, genetic programming, and novel unsupervised pre-training strategies for transformers. Our research achieved remarkable results: 99.58% accuracy in fish speciation, 63.33% accuracy for fish body parts classification, 42.56% accuracy in oil contamination detection, and 86.24% accuracy in cross-species contamination detection. The transformer-based model consistently outperformed traditional machine learning and other deep learning approaches across all tasks. The paper explores the explainability of the best-performing and mostly black-box models using Local Interpretable Model-agnostic Explanations (LIME). REIMS analysis with machine learning proves to be a fast, accurate, and interpretable technique for real-time marine biomass compositional analysis, with potential applications in marine-based industry quality control, product optimisation, and food safety monitoring.

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

1 Introduction

Marine biomass compositional analysis plays a crucial role in various industries, including food production, quality control, and environmental monitoring. Traditional approaches for analyzing marine biomass composition, such as Gas Chromatography Mass Spectroscopy [31], Nuclear Magnetic Resonance

Spectroscopy [2], and Genomic profiling [25], are often time-consuming, labour-intensive, and require significant domain expertise. In response to these challenges, Rapid Evaporative Ionisation Mass Spectrometry (REIMS) has emerged as a promising technique for rapid and accurate analysis of biological samples [3, 4, 15, 33].

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

To address these challenges, this paper introduces several innovative approaches. This paper employs machine learning techniques that provide rapid inference and automation, eliminating the need for human-in-the-loop domain expertise in chemistry or fish processing. To handle the high-dimensionality of REIMS data, this paper utilizes deep learning [7, 9, 13, 17, 20, 23, 30] and evolutionary computation methods [14, 16, 28, 29] that are well-suited for complex feature interactions in mass spectra. To mitigate the limited number of training instances, the paper implements data augmentation and unsupervised pretraining techniques. Finally, the paper employs Local Interpretable Model-agnostic Explanations (LIME) [24] to provide interpretable outputs that identify important features and quantify their impact, making our models more accessible to domain experts in chemistry and fish processing.

2 Dataset

(a) Mackerel (b) Hoki

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

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

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

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

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

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

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

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

3 Classification

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

1. **Traditional machine learning methods:** Random Forest (RF) [12], K-Nearest Neighbors (KNN) [8], Decision Trees (DT) [5], Naive Bayes (NB) [10], Logistic Regression (LR) [18], Support Vector Machines (SVM) [6], and Linear Discriminant Analysis (LDA) [1].
2. **Ensemble method** [11]: A combination of the above traditional methods.
3. **Deep learning methods:** Transformer [7, 30], Long Short-Term Memory (LSTM) [13], Variational Autoencoder (VAE) [17], Kolmogorov-Arnold Networks (KAN) [23], Convolutional Neural Network (CNN) [19–22], and Mamba [9].
4. **Evolutionary computation:** Genetic Algorithms (GA) [14], Particle Swarm Optimization (PSO) [16], and Multiple Class Independent Feature Construction (MCIFC) [28, 29]

They use default settings from sklearn [26], except SVM with a linear kernel and LR set to 2,000 max iterations, these exceptions were found experimentally with trial and error. The ensemble voting classifier uses hard voting. Additionally, more advanced classification techniques, such as deep learning and evolutionary computation methods, are detailed below.

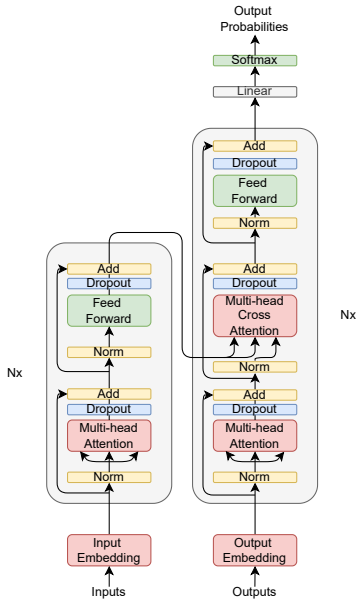
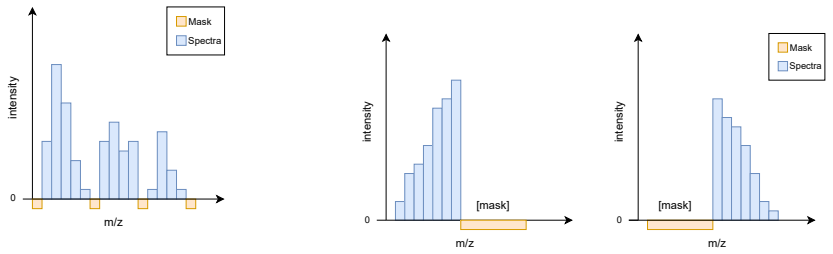


Fig. 2: Transformer Architecture

In previous work, [32] introduced two novel unsupervised pre-training methods for mass spectrometry data, inspired by BERT [7]: Masked Spectra Modelling (MSM) and Next Spectra Prediction (NSP). MSM adapts masked language modelling to mass spectrometry by masking and predicting mass-to-charge ratios in spectra. NSP splits spectra in half and predicts whether two halves belong to the same spectrum. These methods enable the model to learn general patterns from larger, unlabeled datasets, creating useful embeddings for improved performance on smaller, fine-tuned datasets for specific tasks. Please refer to that original paper for more information.



(a) MSM (b) NSP

Fig. 3: Masked spectra modelling (left) Next spectra prediction (right)

Long Short-Term Memory (LSTM) [13] is a type of recurrent neural network designed to handle long-term dependencies in sequential data. Variational Autoencoders (VAE) [17] are generative models that learn to encode data into a latent space and decode it back, allowing for both data compression and generation. Kolmogorov-Arnold Networks (KAN) [23] is a novel type of neural network architecture inspired by the Kolmogorov-Arnold representation theorem, aimed at improving function approximation capabilities. Convolutional Neural Networks (CNN) [19–22] are specialized neural networks that use convolution operations to process grid-like data, particularly effective for image analysis tasks. Mamba [9] is a recent state-space model architecture designed as an alternative to transformers, offering efficient processing of sequential data.

Evolutionary computation (EC) methods of Genetic Algorithms (GA) [14], Particle Swarm Optimization (PSO) [16], and Multiple Class Independent Feature Construction (MCIFC) [28, 29] are employed. Particle Swarm Optimization [16] is a population-based algorithm inspired by bird flocking, where particles move through the search space guided by their own best-known position and the swarm's best-known position. Genetic Algorithms [14] mimic natural selection by evolving a population of potential solutions through selection, crossover, and mutation operations. Genetic Programming [28, 29] is an extension of Genetic Algorithms that evolves computer programs or expressions, typically rep-

resented as tree structures, to solve problems. These three techniques are all nature-inspired metaheuristics used for optimization and search problems

4 Results

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 1 gives the results of the classifiers on the test set, giving the average over 30 independent runs, with the best-performing model on the test set given in **bold**, and second-best in *italics*.

Table 1: Test accuracy classification results

	Fish speciation	Fish part	Oil	Cross-species
RF	95.88% ± 4.47%	40.00% ± 15.27%	38.73% ± 8.15%	81.04% ± 5.67%
KNN	83.69% ± 6.91%	31.66% ± 14.49%	31.94 ± 9.34%	68.68% ± 6.89%
DT	*99.13% ± 1.72%*	27.22% ± 13.25%	28.17% ± 7.34%	69.16% ± 5.59%
NB	87.97% ± 9.57%	45.00% ± 15.60%	32.50% ± 6.84%	55.70% ± 8.34%
LR	96.72% ± 4.75%	56.66% ± 15.27%	30.91% ± 8.32	*86.18% ± 5.03%*
SVM	95.97% ± 5.06%	56.11% ± 14.58%	35.63% ± 7.80%	85.53% ± 5.84%
LDA	96.47% ± 3.67%	45.55% ± 16.06%	31.86% ± 6.65%	81.37% ± 6.60%
Ensemble	98.16% ± 3.00%	51.66% ± 15.72%	37.26% ± 8.69%	84.34% ± 5.84%
Transformer	**99.58% ± 1.31%**	**63.33% ± 24.59%**	**42.56% ± 12.03%**	**86.24% ± 6.27%**
LSTM	96.81% ± 3.74%	58.33% ± 8.78%	31.53% ± 5.37%	83.22% ± 4.51%
VAE	98.18% ± 2.34%	50.00% ± 13.60%	35.38% ± 8.06%	77.41% ± 9.37%
KAN	97.27% ± 2.34%	*60.00% ± 17.91%*	21.92% ± 4.81%	69.67% ± 5.08%
CNN	97.72% ± 3.21%	59.99% ± 14.05%	38.46% ± 9.25%	77.41% ± 7.60%
Mamba	94.09% ± 5.27%	46.66% ± 10.54%	*40.76% ± 6.83%*	81.29% + 7.57%
MCIFC	94.54% ± 10.38%	55.45% ± 19.19%	38.32% ± 8.72%	72.39% ± 9.79%
GA	91.06% ± 6.13%	39.07% ± 10.23%	23.73% ± 6.17%	67.18% ± 8.31%
PSO	78.13% ± 12.31%	28.73% ± 9.85%	16.87% ± 5.46%	46.65% ± 7.40%

5 Discussion

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. This discussion is separated into two sections, one section for dataset-specific analysis and another for model-specific analysis.

5.1 Dataset-specific Analysis

Fish Speciation: In the fish speciation task, most models performed exceptionally well, with the transformer model achieving the highest accuracy of 99.58%.

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

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

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

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

5.2 Model-specific Analysis

Transformer: Consistently outperformed other models across all tasks, likely due to its ability to capture complex, long-range dependencies in the spectral data. Its self-attention mechanism allows it to focus on the most relevant parts of the spectrum for each task.

Traditional Machine Learning Models (RF, KNN, DT, NB, LR, SVM, LDA): Performed well on fish speciation and cross-species contamination tasks, indicating that these tasks have relatively clear decision boundaries. Struggled more with body part classification and oil detection, suggesting these tasks require more complex feature interactions.

Deep Learning Models (LSTM, VAE, KAN, CNN, Mamba): Generally performed well, but often fell short of the transformer's accuracy. CNN's strong performance, particularly in oil detection and cross-species contamination, suggests that local spectral patterns are important for these tasks. Mamba models excel at multi-class classification due to their ability to efficiently handle long-range dependencies and extract adaptive features from complex input sequences.

Ensemble Method: Performed consistently well across tasks, demonstrating the value of combining multiple models to leverage their individual strengths.

Evolutionary Computation: The evolutionary computation methods showed moderate performance across tasks. Multiple Class-Independent Feature Construction (MCIFC) feature construction approach can capture some relevant patterns in the spectral data. Genetic Algorithm (GA) and Particle Swarm Optimisation (PSO) both consistently achieved lower accuracy, underfitting the training dataset, and failing to capture patterns that generalize to unseen data. These population-based search methods struggle to capture relevant patterns in the data when compared to traditional machine and deep learning techniques.

Further discussion: The varying performance of different models across tasks highlights the importance of selecting appropriate algorithms for specific analytical challenges in marine biomass analysis. While the transformer model consistently excelled, simpler models like logistic regression demonstrated competitive performance in certain tasks, offering potential advantages in terms of interpretability and computational efficiency. The challenges faced in oil contamination detection and, to a lesser extent, body part classification, point to areas where further research is needed. This might include exploring more advanced feature extraction techniques, increasing the size and diversity of the training dataset, or developing specialised model architectures tailored to these specific tasks. Overall, our results demonstrate the potential of combining REIMS with machine learning for rapid and accurate marine biomass analysis, while also highlighting areas for future improvement and research.

6 Explainable AI

While the performance of our models is promising, understanding how they arrive at their predictions is crucial for building trust and gaining insights. To address this, the paper employs Local Interpretable Model-agnostic Explanations

(LIME), a technique used to explain predictions made by complex black-box machine learning models [27]. LIME approximates a complex model's behaviour with a simpler and interpretable model (e.g. linear regression) for a specific instance in a local area to be understood. LIME creates and evaluates many altered versions through perturbations of an instance in the input data to see how those perturbations change the prediction. Through perturbations and their observed changes to the prediction, this information is used to generate a local explanation that highlights which features (e.g. mass-to-charge ratios) influenced the prediction. LIME charts are used to explain the predictions of machine learning models by showing which features (in this case, specific mass-to-charge ratios) are most influential for a particular prediction. In these LIME charts:

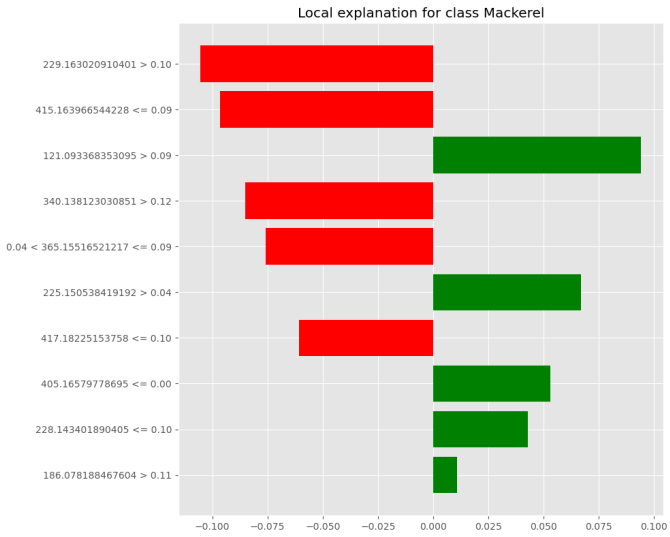
- Green bars: These represent features (mass-to-charge ratios) that contribute positively towards the predicted class. In other words, the presence or higher intensity of these features increases the likelihood of the sample being classified as the predicted class.
- Red bars: These represent features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

6.1 Transformer on Fish Speciation

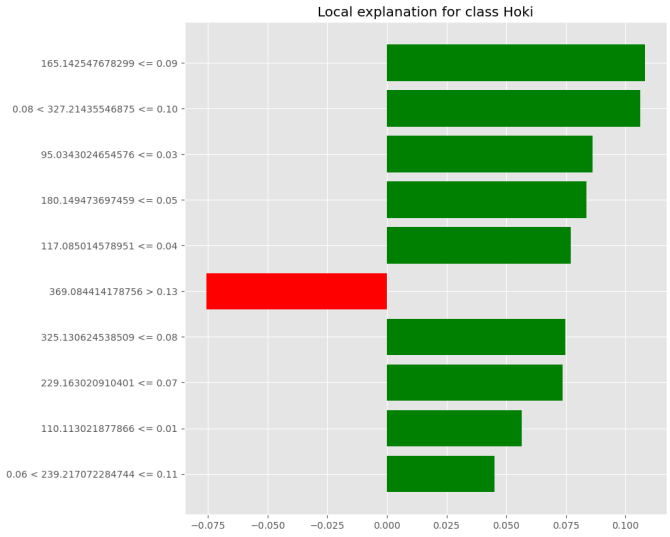
The transformer achieves the best classification accuracy on all datasets, but it notably performs best at fish species classification. Figure 4 gives the LIME explanations for the transformer model for fish speciation, for the first Hoki and Mackerel instances, respectively. For Mackerel, there are several green bars in the 185-405 and 340-420 m/z ranges, indicating that these spectral features strongly support the classification of the sample as Mackerel. One of the largest red bars (i.e. negative correlation) is when the molecule at mass-to-charge ratio 229.1630 m/z is greater than the normalised intensity of 0.10, this suggests it is not commonly found in Mackerel. For Hoki, there are prominent green bars in the 95-370 m/z ranges, suggesting these features are important for identifying Hoki. There is one feature with a red bar (i.e. negative correlation) suggesting that when the mass-to-charge ratio 369.0844 m/z is greater than a normalised intensity of 0.13, this represents a molecule not commonly found in Mackerel.

6.2 KAN on Fish Body Part

The KAN performs the second best on the fish parts dataset. This is another difficult multi-class classification task, this time, however, there are only six



(a) Mackerel



(b) Hoki

Fig. 4: LIME explanations for transformer on fish species of Mackerel (left) Hoki (right)

classes. Here, fig. 5, is the LIME explanation for the KAN for the fish parts dataset. The strongest green bar is when the mass-to-charge ratio 365.3074 m/z is greater than the normalised intensity of 0.23, this molecule is likely highly correlated with the guts of fish body part. The strongest red bar (i.e. negative correlation) is when the mass-to-charge ratio 168.1298 m/z is within the range of normalised intensity $0.15 < y <= 0.21$, indicating when this molecule is present, it is likely not a guts fish body part sample.

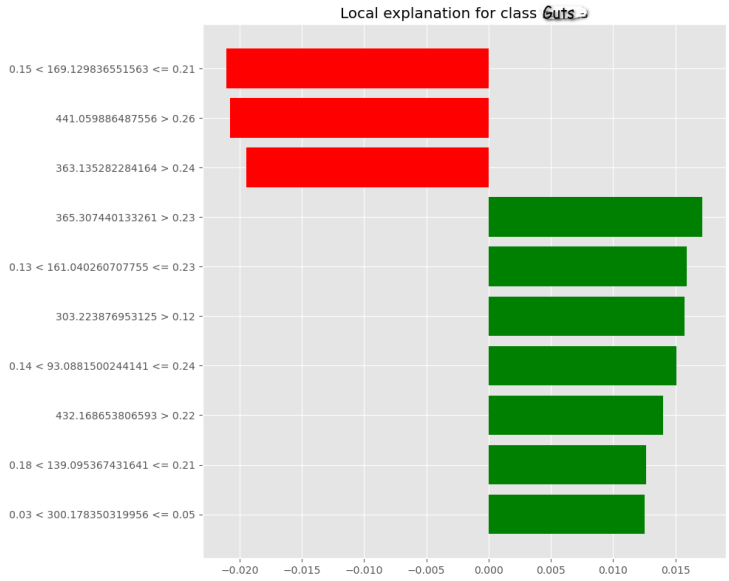


Fig. 5: LIME explanation for KAN for classification of the fish part of guts

6.3 Mamba on Oil Contamination

The Mamba performs second best in oil contamination detection and profiling. This is a difficult multi-class classification problem with seven different classes. Here, in fig. 6, is the LIME explanation for the Mamba for the oil dataset. The strongest green bar is when the mass-to-charge ratio of 80.0119 m/z is greater than a normalised intensity of 0.11, indicating this molecule is likely highly correlated with large concentrations of oil. There are only two red bars (i.e. negative correlation) when the mass-to-charge ratio 269.2526 m/z is greater than the normalised intensity of 0.13, and, when the mass-to-charge ratio 95.1034 m/z is less than or equal to the normalised intensity of 0.05, these molecules are likely associated with fish and not oil.

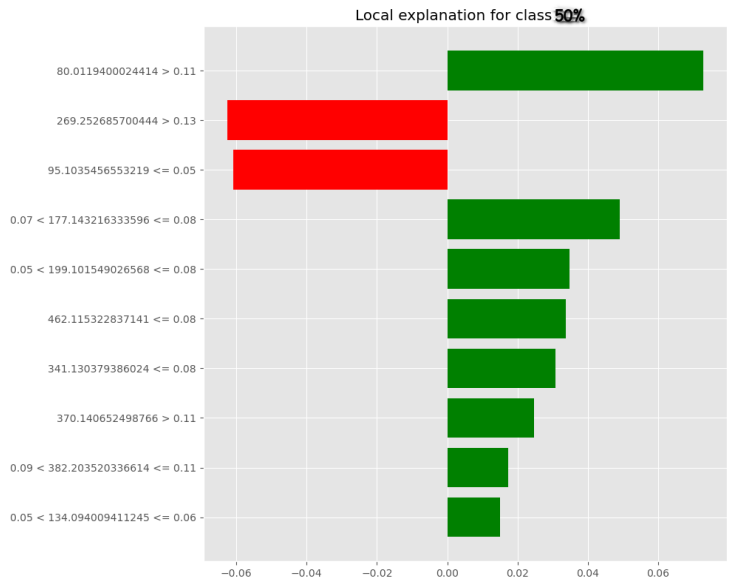


Fig. 6: LIME explanation for Mamba for oil contamination at 50%

6.4 Logistic Regression on Cross-Species Contamination

The Logistic Regression (LR) performs second best on the cross-species dataset. This is a multi-class classification task with three classes. Here, fig. 7, is the LIME explanation for LR for the cross-species dataset. The strongest green bar (i.e. positive correlation) is when the mass-to-charge ratio 295.2214 m/z is in the normalised intensity range of $0.04 < y < 0.06$. Suggesting that small amounts of this molecule are indicative of adultered marine biomass that contains cross-species contamination. The strongest negative bar (i.e. negative correlation) is when the mass-to-charge ratio 281.1253 m/z is in the normalised intensity range $0.03 < y <= 0.06$. Suggesting that small amounts of this molecule are indicative of an unadulterated sample that consists of purely one species of fish.

7 Conclusion

Having examined both the performance and interpretability of our models, this section can now draw overall conclusions about the effectiveness of combining REIMS with advanced machine learning techniques for marine biomass analysis.

This study demonstrates the effectiveness of combining REIMS with advanced machine learning techniques for rapid and accurate marine biomass compositional analysis. The transformer-based model consistently outperformed other methods across all four classification tasks: fish speciation, body part classification, oil contamination detection, and cross-species contamination detection.

The high accuracy achieved in fish speciation (99.58%) showcases the potential of this approach for quality control applications in the fishing industry. While

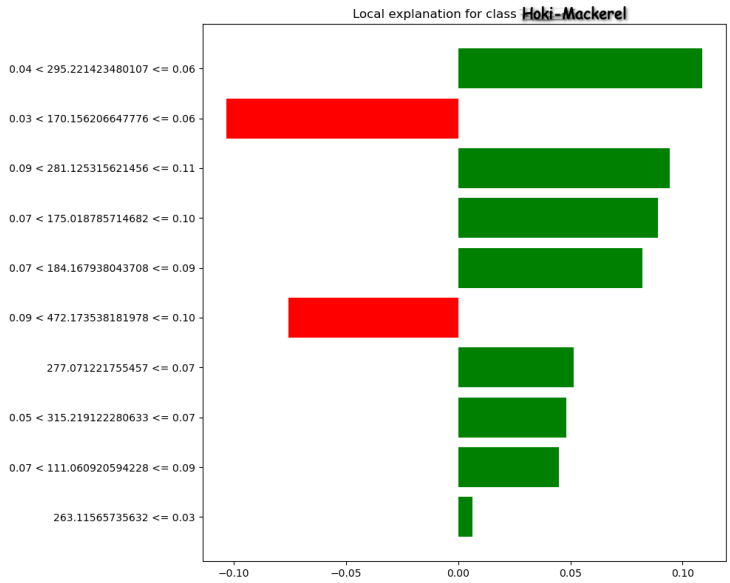


Fig. 7: LIME explanation for LR for cross-species contamination classification with Hoki-Mackerel mix

the performance on fish body part classification (63.33%) and oil contamination detection (42.56%) was lower, it still represents a significant improvement over traditional analysis methods in terms of speed and automation.

Notably, in the cross-species contamination detection task, both the transformer model and logistic regression performed exceptionally well, achieving accuracies of 86.24% and 86.18% respectively. The strong performance of logistic regression, a relatively simple linear model, suggests that the features extracted from the REIMS spectra for this task are highly informative and linearly separable.

The application of explainable AI techniques, particularly LIME (Local Interpretable Model-agnostic Explanations), provided valuable insights into the decision-making processes of our models. These explanations revealed specific mass-to-charge ratios that strongly influence classifications, enhancing our understanding of the biochemical markers associated with different fish species, body parts, and contamination levels. For instance, the LIME analysis for fish speciation highlighted distinct spectral regions that differentiate Mackerel from Hoki, while the analysis for oil contamination pointed to specific molecular markers associated with high oil concentrations.

This interpretability not only increases confidence in the model's predictions but also opens up possibilities for new scientific insights into the biochemical composition of marine biomass. It demonstrates that our approach can provide both accurate classifications and meaningful, chemically relevant explanations for those classifications.

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

8 Future Work

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

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

References

1. Balakrishnama, S., Ganapathiraju, A.: Linear discriminant analysis-a brief tutorial. Institute for Signal and information Processing **18**(1998), 1–8 (1998)
2. Bettjeman, B.I., Hofman, K.A., Burgess, E.J., Perry, N.B., Killeen, D.P.: Seafood phospholipids: extraction efficiency and phosphorous nuclear magnetic resonance spectroscopy (31p nmr) profiles. Journal of the American Oil Chemists' Society **95**(7), 779–786 (2018)
3. Black, C., Chevallier, O.P., Cooper, K.M., Haughey, S.A., Balog, J., Takats, Z., Elliott, C.T., Cavin, C.: Rapid detection and specific identification of offals within minced beef samples utilising ambient mass spectrometry. Scientific reports **9**(1), 1–9 (2019)
4. Black, C., Chevallier, O.P., Haughey, S.A., Balog, J., Stead, S., Pringle, S.D., Riina, M.V., Martucci, F., Acutis, P.L., Morris, M., et al.: A real time metabolomic profiling approach to detecting fish fraud using rapid evaporative ionisation mass spectrometry. Metabolomics **13**(12), 1–13 (2017)
5. Breiman, L.: Classification and regression trees. Routledge (2017)
6. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3), 273–297 (1995)

7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Fix, E., Hodges, J.L.: Discriminatory analysis. nonparametric discrimination: Consistency properties. International Statistical Review/Revue Internationale de Statistique **57**(3), 238–247 (1989)
9. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
10. Hand, D.J., Yu, K.: Idiot's bayes—not so stupid after all? International statistical review **69**(3), 385–398 (2001)
11. Hansen, L.K., Salamon, P.: Neural network ensembles. IEEE transactions on pattern analysis and machine intelligence **12**(10), 993–1001 (1990)
12. Ho, T.K.: Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition. vol. 1, pp. 278–282. IEEE (1995)
13. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
14. Holland, J.H.: Genetic algorithms. Scientific american **267**(1), 66–73 (1992)
15. Jha, S.N.: Rapid detection of food adulterants and contaminants: theory and practice. Academic Press (2015)
16. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of ICNN'95-international conference on neural networks. vol. 4, pp. 1942–1948. IEEE (1995)
17. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
18. Kleinbaum, D.G., Dietz, K., Gail, M., Klein, M., Klein, M.: Logistic regression. Springer (2002)
19. LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., Jackel, L.: Handwritten digit recognition with a back-propagation network. Advances in neural information processing systems **2** (1989)
20. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural computation **1**(4), 541–551 (1989)
21. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
22. LeCun, Y., et al.: Generalization and network design strategies. Connectionism in perspective **19**(143-155), 18 (1989)
23. Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T.Y., Tegmark, M.: Kan: Kolmogorov-arnold networks. arXiv preprint arXiv:2404.19756 (2024)
24. McCann, S., Lowe, D.G.: Local naive bayes nearest neighbor for image classification. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3650–3656. IEEE (2012)
25. Pardo, M.Á., Jiménez, E., Pérez-Villarreal, B.: Misdescription incidents in seafood sector. Food Control **62**, 277–283 (2016)
26. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
27. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)

28. Tran, B., Xue, B., Zhang, M.: Genetic programming for feature construction and selection in classification on high-dimensional data. Memetic Computing **8**(1), 3–15 (2016)
29. Tran, B., Xue, B., Zhang, M.: Genetic programming for multiple-feature construction on high-dimensional classification. Pattern Recognition **93**, 404–417 (2019)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
31. Wood, J., Nguyen, B.H., Xue, B., Zhang, M., Killeen, D.: Automated fish classification using unprocessed fatty acid chromatographic data: A machine learning approach. In: Australasian Joint Conference on Artificial Intelligence. pp. 516–529. Springer (2022)
32. Wood, J., Nguyen, B.H., Xue, B., Zhang, M., Killeen, D.: 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. pp. 516–529. Springer (2024)
33. Zhang, R., Ross, A.B., Jacob, N., Agnew, M., Staincliffe, M., Farouk, M.M.: Rapid evaporative ionisation mass spectrometry fingerprinting can discriminate lamb meat due to different ageing methods and levels of dehydration. Journal of Proteomics **272**, 104771 (2023)