A rapid machine-learning approach for detecting fish species and body parts using rapid evaporative ionisation mass spectrometry

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 content, and cross-species contamination as model systems representing diverse biochemical profiles, we 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% accuracy in oil detection, and 86% accuracy in cross-species contamination detection. The transformer-based model consistently outperformed traditional machine learning and other deep learning approaches across all tasks. 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 optimization, 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 methods - such as Gas Chromatography Mass Spectroscopy and Genomic profiling - for analyzing marine biomass composition are often time-consuming, labour-intensive, and require significant domain expertise [2, 17, 21]. Rapid Evaporative Ionisation Mass Spectrometry (REIMS) has emerged as a promising technique for rapid and accurate analysis of biological samples [17, 23].

By leveraging the power of machine learning, particularly transformer-based models, we seek to develop a fast, accurate, and interpretable approach for real-time marine biomass compositional analysis. This research has significant implications for quality control, product optimization, and food safety in marine-based industries.

(a) Mackerel (b) Hoki

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

2 Dataset

Following our introduction to the challenges and potential of REIMS-based analysis, we now turn our attention to the critical foundation of our study: the dataset. The dataset used in this study consists of REIMS spectra collected from various fish species, body parts, and samples with different oil contents. Additionally, we prepared samples to simulate cross-species contamination scenarios. The data collection process involved:

1. Fish speciation: Samples from multiple fish species were collected and analyzed using REIMS.
2. Fish body parts: Various parts of fish (e.g., muscle, skin, organs) were isolated and analyzed.
3. Oil content: Samples with varying levels of oil content were prepared and analyzed.
4. Cross-species contamination: Samples were prepared by intentionally mixing tissues from different fish species to simulate contamination scenarios.

The REIMS spectra were preprocessed to remove noise and normalize the data. The dataset was then split into training and testing sets for each classification task.

3 Classification

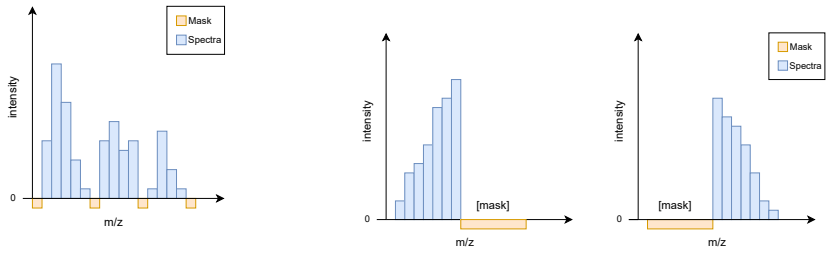
With our dataset established, we move on to the heart of our analytical approach: the classification methodologies employed to extract meaningful insights from the REIMS spectra. We employed a diverse range of machine learning techniques to classify the REIMS spectra:

1. Traditional machine learning methods: Random Forest (RF) [8], K-Nearest Neighbors (KNN) [5], Decision Trees (DT) [3], Naive Bayes (NB) [6], Logistic Regression (LR) [11], Support Vector Machines (SVM) [4], and Linear Discriminant Analysis (LDA) [1].
2. Ensemble method [7]: A combination of the above traditional methods.

3. Deep learning methods: Transformer, Long Short-Term Memory (LSTM) [9], Variational Autoencoder (VAE) [10], Kolmogorov-arnold networks (KAN) [16], Convolutional Neural Network (CNN) [12–15], and Multiple Class Independent Feature Construction (MCIFC) [19, 20]

They use default settings from sklearn [18], 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 methods of transformers and genetic programming are detailed below.

In previous work [22] introduced two novel unsupervised pre-training methods for mass spectrometry data, inspired by BERT: 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. 2: Masked spectra modelling (left) Next spectra prediction (right)

4 Results and Discussions

Having outlined our classification strategies, we now present and interpret the outcomes of applying these various machine learning techniques to the REIMS datasets. Table 1 gives the results of the classifiers, giving the average over 30 independent runs, with the best-performing model on the test set given in **bold**.

Table 1: Fish speciation and fish part 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%	32.5% ± 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%**	**86%**
LSTM	96 %	58%	31%	0.83%
VAE	98%	50%	35%	77%
KAN	97%	60%	21%	69%
CNN	97%	50%	**42%**	**86%**
MCIFC	94%	55%	38%	72%

5 Conclusion

After thoroughly examining the results of our classification experiments, we can now draw comprehensive conclusions about the effectiveness of our REIMS-based machine learning approach 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 content detection, and cross-species contamination detection.

The high accuracy achieved in fish speciation (99.58%) and cross-species contamination detection (86%) showcases the potential of this approach for quality control and food safety applications in the fishing industry. While the performance on fish body part classification (63.33%) and oil content detection (42%) was lower, it still represents a significant improvement over traditional analysis methods in terms of speed and automation.

The success of the transformer model highlights the importance of leveraging advanced deep learning techniques in processing complex spectral data. Its ability to capture long-range dependencies and learn hierarchical features from the REIMS spectra contributes to its superior performance.

6 Future Work

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

- Expand the dataset: Collect REIMS spectra from a wider variety of fish species, marine organisms, and contaminants to improve the model's generalizability.
- Improve oil content detection: Investigate additional feature engineering techniques or alternative machine learning architectures to enhance the accuracy of oil content classification.
- Real-time analysis: Develop a system for real-time REIMS data acquisition and analysis, allowing for immediate classification results in industrial settings.
- Transfer learning: Explore the potential of transfer learning to adapt the trained models to new, related tasks or similar spectral data from other mass spectrometry techniques.
- Interpretability: Enhance the interpretability of the transformer model's decisions, potentially through attention visualization or other explainable AI techniques.
- Multi-task learning: Investigate the potential of multi-task learning approaches to simultaneously perform multiple classification tasks, potentially improving overall performance.
- Quantitative analysis: Extend the approach to perform quantitative analysis of specific compounds or contaminants in marine biomass samples.
- Integration with other data sources: Combine REIMS data with other analytical techniques or metadata to create more comprehensive and accurate classification models.
- Industrial validation: Conduct large-scale industrial trials to validate the performance and practicality of the REIMS-ML approach in real-world settings.
- 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. Breiman, L.: Classification and regression trees. Routledge (2017)
4. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3), 273–297 (1995)
5. Fix, E., Hodges, J.L.: Discriminatory analysis. nonparametric discrimination: Consistency properties. International Statistical Review/Revue Internationale de Statistique **57**(3), 238–247 (1989)
6. Hand, D.J., Yu, K.: Idiot's bayes—not so stupid after all? International statistical review **69**(3), 385–398 (2001)

7. Hansen, L.K., Salamon, P.: Neural network ensembles. IEEE transactions on pattern analysis and machine intelligence **12**(10), 993–1001 (1990)
8. Ho, T.K.: Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition. vol. 1, pp. 278–282. IEEE (1995)
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
10. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
11. Kleinbaum, D.G., Dietz, K., Gail, M., Klein, M., Klein, M.: Logistic regression. Springer (2002)
12. 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)
13. 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)
14. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
15. LeCun, Y., et al.: Generalization and network design strategies. Connectionism in perspective **19**(143-155), 18 (1989)
16. 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)
17. Pardo, M.Á., Jiménez, E., Pérez-Villarreal, B.: Misdescription incidents in seafood sector. Food Control **62**, 277–283 (2016)
18. 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)
19. 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)
20. Tran, B., Xue, B., Zhang, M.: Genetic programming for multiple-feature construction on high-dimensional classification. Pattern Recognition **93**, 404–417 (2019)
21. 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)
22. 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)
23. Zhang, L., Agrawala, M.: Adding conditional control to text-to-image diffusion models. arXiv preprint arXiv:2302.05543 (2023)