Instance-recognition for detecting batches of marine biomass using rapid evaporative ionisation mass spectrometry

using rapid evaporative ionisation mass spectrometry

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

[1]School of Engineering and Computer Science — Te Kura Mātai Pūkaha, Pūrorohiko
Victoria University of Wellington — Te Herenga Waka
[2]New Zealand Institute for Plant and Food Research Limited, Nelson, New Zealand

ECRG, VUW, Wellington, New Zealand

https://ecs.wgtn.ac.nz/Groups/ECRG/WebHome

Plant and Food Research, Nelson, New Zealand

https://www.plantandfood.com/en-nz/

Island Bay, Wellington, New Zealand

Topics

1. Introduction
2. Related Work
   - REIMS
   - Siamese Networks
   - Deep Learning
3. Dataset
4. Methods
   - Binary Classification
   - Siamese Networks
   - Distance Measures
   - Loss functions
5. Results
6. Conclusion

Introduction

Introduction

- Identifying marine biomass batches ensures quality control and regulatory compliance.
- Traditional methods are error-prone, time-consuming, expensive, and require domain expertise in chemistry; machine learning offers a new real-time solution.
- This research compares traditional ML, evolutionary techniques, and deep learning models for instance recognition for batch detection.
- A key approach is using contrastive learning with Siamese networks for nuanced batch comparison.

Related Work

REIMS

Rapid Evaporative Ionization Mass Spectrometry (REIMS) [1]

- **Applications:** Beef adulteration [2], Fish fraud [3]
- OPLS-DA and PCA are traditional methods for REIMS analysis [4, 5, 6]. This method relies on hyperparameters, such as the outlier thresholds, and the number of principal components; domain experts in chemistry manually set these.
- *Deep learning surpasses OPLS-DA* in REIMS analysis.
- This work introduces contrastive learning via Siamese networks and models with learnable parameters.

Siamese Networks

- Siamese networks, introduced by LeCun (1993), are used for object detection, segmentation, and ransomware classification [7, 8, 9].
- They are a type of contrastive learning, which learns similarity by contrasting inputs.
- This work updates Siamese networks with modern deep learning and evolutionary computation methods.
- Evaluates various models (CNNs, Transformers, LSTM, VAE, KAN, Mamba); results show improved performance for instance recognition for fish batch detection using contrastive learning.

Siamese Networks - signature verfication

Transformer

Transformers excel in NLP tasks, capturing long-range dependencies [10, 11].

Figure 1: The Transformer - model architecture.

Transformer architecture

CNN

CNNs are key for spatial feature extraction in computer vision [12, 13, 14, 15].

CNN architecture

LSTM

LSTMs handle time-series data effectively for tasks like speech recognition [16].

LSTM architecture

VAE

VAEs learn robust latent representations for tasks like image generation [17].

VAE architecture

Mamba

Mamba is a state space model architecture designed as an alternative to transformers, offering faster and more efficient sequential data processing [18].

Mamba architecture

KAN

Kolmogorov-Arnold Networks (KAN) decompose multivariate functions into simpler univariate functions for efficient modelling [19].

KAN architecture

Dataset

Mass Spectrometry [20, 3, 2] ≈ Chemical Fingerprint

Mass Spectrometry: Steps

1. Laser Pen
2. Vacuum
3. Electromagnetic Field (EMF)
4. Detector

Mass Spectrometry: Steps

1. Laser Pen
2. Vacuum
3. Electromagnetic Field (EMF)
4. Detector

Mass Spectrometry: Steps

1. Laser Pen
2. Vacuum
3. Electromagnetic Field (EMF)
4. Detector

Mass Spectrometry: Steps

1. Laser Pen
2. Vacuum
3. Electromagnetic Field (EMF)
4. Detector

Mass Spectrometry: Steps

1. Laser Pen
2. Vacuum
3. Electromagnetic Field (EMF)
4. Detector

Fish Species

Hoki

Mackerel

Fish Body Parts

Instance Recognition for batch detection

Unlike general object recognition, which classifies objects into broad categories (e.g., identifying a "cat" in an image), instance recognition aims to identify specific occurrences or examples of objects (e.g., recognizing a particular cat among many).

- REIMS dataset: 72 fish samples from 24 batches
- Task: Instance recognition of batches using sample pairs
- Challenges: High dimensionality (1023 features) and class imbalance (more different pairs than same)
- Approach: 50 pairs per sample, pairs chosen at random, 50-50 train-test split, evaluated with balanced accuracy over 30 independent runs.

Batch identification in fish processing

Methods

Binary Classification

Binary classification for instance recognition

Siamese Networks

Flowchart of Contrastive Learning

Distance Measure

D1, D2 are Euclidean Distances

θ1, θ2 cosine similarity angles

Euclidian Distance

Euclidean distance measures the straight-line distance between two points in Euclidean space. It is defined as:

$$d = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \tag{1}$$

Where:

- a_i and b_i are the ith components of vectors **A** and **B**

Properties:

- Range: $[0, \infty)$
- 0 indicates identical vectors
- Increases as vectors become more different

Cosine Similarity

Cosine similarity measures the cosine of the angle between two vectors in a multi-dimensional space. It is defined as:

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} \tag{2}$$

Where:

- $\mathbf{A} \cdot \mathbf{B}$ is the dot product of vectors $\mathbf{A}$ and $\mathbf{B}$
- $\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ are the magnitudes (Euclidean norms) of vectors $\mathbf{A}$ and $\mathbf{B}$

Properties:

- Range: $[-1, 1]$ for general vectors, $[0, 1]$ for non-negative vectors
- 1 indicates identical orientation
- 0 indicates orthogonality (perpendicular vectors)
- -1 indicates opposite orientation

Balanced Accuracy

Balanced Accuracy is a metric used to evaluate the performance of a binary classifier, especially when dealing with imbalanced datasets. It is defined as the average of sensitivity (true positive rate) and specificity (true negative rate).

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \tag{3}$$

Where:

- TP is the number of true positives
- TN is the number of true negatives
- FP is the number of false positives
- FN is the number of false negatives

Cross Entropy

Cross Entropy Loss, also known as Log Loss, measures the performance of a classification model whose output is a probability value between 0 and 1. For binary classification, it is defined as:

$$\mathrm{CE}(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{4}$$

Where:

- N is the number of samples
- y_i is the true label (0 or 1) for the i-th sample
- $\hat{y}_i$ is the predicted probability of the positive class for the i-th sample

Weighted Cross Entropy

For binary classification, the Weighted Cross Entropy loss is defined as:

$$\mathrm{WCE}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}[w_1 y_i \log(\hat{y}_i) + w_0(1 - y_i)\log(1 - \hat{y}_i)] \tag{5}$$

Where:

- N is the number of samples
- y_i is the true label (0 or 1) for the i-th sample
- $\hat{y}_i$ is the predicted probability of the positive class for the i-th sample
- w_1 is the weight for the positive class (typically the minority class)
- w_0 is the weight for the negative class (typically the majority class)

The weights w_1 and w_0 are often set inversely proportional to the class frequencies in the training data. A common approach is to use:

Contrastive Loss

Let z_1 and z_2 be the encoded representations of two input samples, and y be the label indicating whether the pair is similar ($y = 1$) or dissimilar ($y = 0$). The cosine similarity is defined as:

$$\text{cos_sim}(z_1, z_2) = \frac{z_1 \cdot z_2}{\|z_1\| \|z_2\|} \tag{6}$$

The Contrastive Loss function is then defined as:

$$\begin{aligned} \text{CL}(z_1, z_2, y) = & y \cdot (1 - \text{cos_sim}(z_1, z_2))^2 + \\ & (1 - y) \cdot \max(0, \text{cos_sim}(z_1, z_2) - \text{margin})^2 \end{aligned} \tag{7}$$

Where:

- margin is a hyperparameter (often set to 0.1) that determines the minimum distance between dissimilar pairs
- cos_sim is the cosine similarity distance metric between pairs
- y is the label (1 for similar pairs, 0 for dissimilar pairs)
- z_1 and z_2 are the encoded representations of the input pairs

Results

Results

Method	Binary Classification		Contrastive Learning	
	Train	Test	Train	Test
KNN	52.86% ± 1.30%	49.88% ± 0.09%		
DT	**100.0% ± 0.00%**	52.28% ± 3.32%		
LDA	57.53% ± 0.99%	54.07% ± 2.64%		
NB	63.00% ± 1.24%	53.94% ± 3.67%		
RF	*99.94% ± 0.17%*	51.48% ± 1.59%		
SVM	52.84% ± 1.27%	51.82% ± 2.04%		
LR	53.45% ± 1.17%	51.58% ± 1.53%		
OPLS-DA	57.08% ± 1.46%	53.19% ± 2.22%		
Ensemble	58.56% ± 0.99%	51.48% ± 1.59%		
CNN	50.00% ± 0.00%	50.00% ± 0.00%	94.50 ± 1.17%	70.82% ± 6.67%
GP	76.08% ± 1.65%	65.80% ± 2.88%	85.35% ± 5.32%	61.27% ± 6.37%
KAN	98.80% ± 0.58%	59.28% ± 5.23%	94.47% ± 1.60%	*75.27% ± 3.24%*
LSTM	98.74% ± 0.86%	*68.02% ± 2.09%*	**98.65% ± 0.51%**	75.05% ± 5.63%
Mamba	99.10% ± 0.85%	59.80% ± 2.98%	95.23% ± 1.25%	**76.96% ± 4.11%**
Transformer	59.04% ± 12.70%	**70.09% ± 9.89%**	*97.83% ± 0.43%*	73.79% ± 2.60%
VAE	50.10% ± 0.28%	52.89% ± 1.81%	94.62% ± 5.23%	67.83% ± 3.01%

Binary Classification and Contrastive Learning Results

Deep Learning Methods Box Plot

Box plot of deep learning methods

Conclusion

TLDR;

Contrastive learning can perform **instance recognition** for **batch detection**, better than traditional binary classification, for **Rapid Evaporative Ionisation Mass Spectrometry**.

Download the slides & paper.

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a" siamese" time delay neural network," *Advances in neural information processing systems*, vol. 6, 1993.

[8] L. Jing, J. Zhu, and Y. LeCun, "Masked siamese convnets," *arXiv preprint arXiv:2206.07700*, 2022.

[9] J. Zhu, J. Jang-Jaccard, A. Singh, I. Welch, A.-S. Harith, and S. Camtepe, "A few-shot meta-learning based siamese neural network using entropy features for ransomware classification," *Computers & Security*, vol. 117, p. 102691, 2022.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

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

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

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

[19] 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.

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