SpectroSim: Batch Detection in Marine Biomass

Jesse Wood Bach Hoai Nguyen Bing Xue Mengjie Zhang Daniel Killeen

Abstract—The batch detection of marine biomass constitutes a significant real-world application within the fish processing industry, contributing to food safety, fraud prevention, and stock management. Recent advancements have demonstrated that Rapid Evaporative Ionization Mass Spectrometry (REIMS) when coupled with Orthogonal Partial Least Squares Discriminant Analysis (OPLS-DA), yields exceptional outcomes in fraud detection, contamination identification, and biomass analysis. Although several studies have employed REIMS and OPLS-DA for species identification and contamination detection—including limited applications to marine biomass—these efforts have not yet addressed the challenge of batch detection, which involves determining the specific batch of processed samples from which a fish originates. Contrastive Learning, an emerging alternative to conventional binary classification, has proven effective for batch detection of marine biomass analyzed via REIMS. Leveraging a high-dimensional REIMS dataset provided by Plant and Food Research, New Zealand, comprising mass spectrometry profiles of New Zealand marine biomass, we propose a novel Contrastive Learning approach termed SpectroSim, building upon the SimCLR framework. The new method introduces a bespoke encoder head, replacing the traditional ResNet backbone with a Transformer architecture, alongside a custom projection head meticulously designed for mass spectrometry data. Comprehensive experimental results indicate that SpectroSim surpasses the balanced classification accuracy of established deep learning frameworks and other prevalent baseline models. Notably, as an unsupervised methodology, SpectroSim achieves near-perfect accuracy (98.02%) in a self-supervised context, independent of class labels.

Index Terms—AI applications, binary classification, contrastive learning, high-dimensional data, machine learning, mass spectrometry, multidisciplinary

1 Introduction

Batch detection of marine biomass is an important research problem. Research suggests batch traceability in fish processing is crucial for food safety, allowing quick recalls if contamination occurs [40]. Tracing batches helps maintain quality control by identifying processing issues specific to certain batches [45] which prevents fraud in the fish industry [16]. The purpose of batch detection in fish processing is to identify which batch of processed samples a fish originated from. In this paper, we focus on a pair-wise comparison batch detection of marine biomass, that is, given two fish analyzed with mass spectrometry, detect if they originate from the same batch.

This paper aims to address a significant real-world engineering problem by developing methods that can be deployed in fish processing factories to improve quality assurance protocols, food safety, contamination detection and stock management. Our specific model is hand-crafted to suit the unique characteristics of the New Zealand seafood industry, enhancing its effectiveness for batch detection of marine biomass in fish processing.

As a response to events such as the 2013 European Horse Meat Scandal [26], the advent of ambient mass spectrometry techniques, like Rapid Evaporative Ionization Mass Spectrometry (REIMS) [4] - the focus of this study - is one such example of a rapid, accurate and destructive in-situ analytical chemistry technique for precise chemical fingerprinting of the constituents of biomass materials. In combination with machine learning techniques like Orthogonal Partial Least Squares Discriminant Analysis (OPLS-DA), REIMS has proven effective across various domains of biomass analysis, such as outlier thresholding for offal contamination detection [5] and fish species and catch method identification for fraud detection [6]. Since the current/existing research was primarily developed by chemists and statisticians, it often lacks exploration of state-of-the-art machine learning techniques. Adapting state-of-the-art machine learning techniques to REIMS biomass analysis requires further research. The existing machine-learning methods for REIMS biomass analysis are limited to supervised statistical techniques for classification [4]–[6], and do not fully capture the complex sequential nature of high-dimensional [34] spectral data. Nor has any research tackled the problem of batch detection of marine biomass. Therefore, in this paper, we develop novel state-of-the-art machine-learning techniques for an equally novel domain.

However, batch detection of marine biomass analyzed with REIMS is a challenging task because of three reasons — high-dimensionality, sequential nature of the data and noise. First, due to the extensive costs and time requirements for sample preparation, the number of training instances is limited, consisting of 72 fish samples. Due to the inherent nature of REIMS analysis, the output mass spectrograph is high-resolution, consisting of 2,080 features. This naturally induces the curse of dimensionality [34], where traditional machine learning methods such as OPLS-DA often struggle with limited data and too many features. Second, the sequential patterns of the data are ignored by traditional supervised methods, like OPLS-DA. The model cannot capture the spatial dependencies, long-term relationships, and complex feature interactions between neighbouring mass spectra. Thirdly, REIMS data is inherently noisy, due to a combination of instrumental, environmental, and sample-related factors that introduce variability and artifacts into the measurements. Traditional machine learning models are not well equipped to handle data with sufficient noise.

All three challenges mean that traditional machine learning methods, such as OPLS-DA, will struggle to achieve high classification performance in batch detection of marine biomass.

In pair-wise comparison tasks, such as our batch detection for marine biomass, contrastive learning has emerged as a popular alternative approach to binary classification. Popularized in the early 90s, with Bromley et al. [9] introducing Siamese Networks for signature verification, and still used today for low-shot image classification [29] and ransomware classification [47]. The fundamental principle of a Siamese network revolves around the use of two identical subnetworks—often referred to as "twin" or "sibling" networks—that share the same weights and architecture. These subnetworks process two separate inputs simultaneously, producing output representations (typically embeddings) that are then compared using a distance metric. The network is trained using a contrastive loss function, which optimizes the model to minimize the distance between embeddings of similar pairs and maximize the distance between embeddings of dissimilar pairs.

This paper focuses on an extension of Siamese networks, SimCLR (Simple Contrastive Learning of Representations) [11]. SimCLR is a self-supervised learning framework that trains a single neural network to produce similar embeddings for two augmented versions of the same input (positive pairs) while pushing apart embeddings of different inputs (negative pairs) within a batch, using the NT-Xent loss function. It extends the Siamese network concept by eliminating the need for labelled data, introducing a projection head, and leveraging large-scale batch processing to learn robust, generalizable representations. The SimCLR model is particularly advantageous for batch detection for marine biomass analysis because data augmentation is not needed. Instead, we formulate the dataset as all possible combinations of the 72 samples of fish, formulating a heavily imbalanced dataset, where SimCLR relies on a diverse set of dissimilar examples for contrastive learning (referred to as negative sampling mining).

With the SimCLR framework in mind, our research strives to answer the question: *Can the SimCLR framework be adapted for batch detection for marine biomass analysis? Particularly, for high-dimensional, sequential, noisy data?*

We hypothesised that the existing SimCLR framework cannot effectively solve the batch detection for marine biomass due to the following four factors: high computational demands [11], dependence on data augmentation [12], poor performance on small datasets [31], and without data augmentation supervised labels are required for training [11]. First, the original study which presented SimCLR, [11], relied on batch sizes of 2048, to acquire sufficient negative sample mining for effective training. The high computational demands of training limit the efficacy of SimCLR, especially for smaller research labs where sufficient GPU resources are not readily available. Second, SimCLR relies on augmenting a single instance to create positive examples for pair-wise comparison. This data augmentation requires careful precision and domain knowledge, as such not to lose any important information from the original sample. Thirdly, SimCLR has been known to struggle on smaller datasets, as it usually requires large volumes of data to train properly. Fourthly, without data augmentation, it requires supervised labels for positive and negative to optimize the contrastive loss function.

To address the limitations of the existing SimCLR, the overall goal of this paper is to propose a new SimCLR model for applications in batch detection for marine biomass. The proposed method is called SpectroSim. SpectroSim's main contributions are: reduced batch size, transformer backbone, custom projection head, and custom loss function. Here we elaborate on each contribution in further detail:

- We decrease the batch size significantly, from the original 2048 to 16. Reducing the computational cost of running the model, and making our method accessible to those with commmodity hardware.
- To address the sequential nature of the REIMS data, we draw inspiration from the Transformer [15], [46] - which transformed natural language processing and large language models [20], [28] - utilizing it as a drop-in replacement for the ResNet [23] backbone. Considering that SimCLR was originally designed for 2D images, not mass spectrometry data, we propose a new projection head architecture for batch detection for marine biomass with REIMS.
- We demonstrate that with modification, SpectroSim can effectively be implemented on a dataset with a limited number of training instances.
- We prove that SpectroSim can achieve near-perfect performance (98.02%), even without the use of class labels - retaining the model's original self-supervised nature, which was originally implemented through data augmentation, that is now utilized in a custom loss function that ignores class labels.

2 Related Works

Having established the importance of batch detection and contrastive learning in marine biomass processing and outlined our multi-faceted approach, we now turn to an examination of related works that inform and contextualize our research.

2.1 Rapid Evaporative Ionization Mass Spectrometry

Rapid Evaporative Ionization Mass Spectrometry (REIMS) [4] shows promise in beef processing, detecting horse meat contamination in beef at low levels (1–5%) [5]. It is also applied to fish fraud detection, identifying fish species and catch methods for fish products [6]. The method identifies instances in the training data that are mislabeled (e.g., incorrectly identified species or origins). Analysis of biomass in the literature [4]–[6] demonstrates that Orthogonal Partial Least Squares Discriminant Analysis (OPLS-DA) [3], [7], [10] for binary classification—distinguishing between categories such as species or contamination status—combined with Principal Component Analysis [2] for dimensionality reduction is an effective technique for REIMS analysis. These studies employ outlier thresholding, where experts with domain knowledge manually set thresholds to detect anomalies (e.g., contaminated or misidentified samples), enhancing classification reliability. This reliance on domain

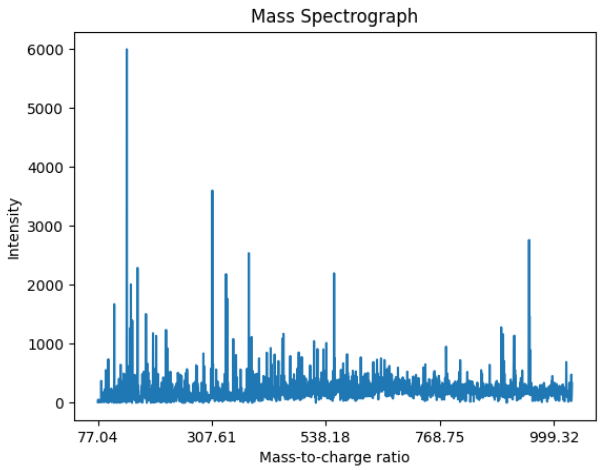

Fig. 1: Mass Spectrograph

knowledge to manually set outlier thresholds is a bottleneck, that also lacks academic rigour [18]. Figure 1 gives the output of REIMS, a mass spectrograph, with the x-axis representing mass-to-charge ratios (m/z), and the y-axis representing relative abundance or intensity.

2.2 Batch Detection

In marine biomass analysis with Rapid Evaporative Ionization Mass Spectrometry (REIMS), batch detection identifies the specific processing batch a fish sample originates from, distinct from individual fish identity, and supports traceability throughout fish processing. This task is critical for food safety, fraud prevention, and supply chain management. Batch traceability enables rapid identification and recall of contaminated products, such as those with harmful bacteria, minimizing consumer risk and economic loss, as Mai et al. [40] highlight in managing fish supply chain risks. Regulations often mandate batch-level traceability to meet safety and quality standards in international trade, while Thompson et al. [45] note its role in enhancing seafood supply chain efficiency by maintaining product freshness. It also combats fraud, like mislabeling or origin misrepresentation, building consumer trust in a global market—Donlan et al. [16] review its effectiveness against species substitution. Pardo et al. [42] and Daily et al. [1] further illustrate fraud prevention needs. REIMS enhances batch detection by identifying unique chemical signatures, an approach Balog et al. [4], De et al. [14], and Black et al. [6] demonstrate for species identification, adaptable to batch-level traceability.

Batch detection of marine biomass does face several limitations:

- **Cost and Implementation Complexity**: Implementing batch detection systems, such as RFID tags or barcodes, can be expensive and technically complex, especially for smaller firms [40]. This financial burden may disproportionately affect processing firms, making adoption difficult.
- **Uneven Distribution of Costs and Benefits**: Research indicates that while processing firms bear the cost, distribution firms closer to consumers reap more benefits, which can discourage widespread adoption and create tension in supply chains [45].
- **Vulnerability to Mislabeling or Fraud**: Batch systems may not prevent mislabeling or fraud if batches are incorrectly labelled, posing risks to product authenticity and safety, especially in contexts where fraud is a concern [16].

2.3 Contrastive Learning

Contrastive learning is a machine learning technique that learns effective representations by contrasting positive and negative pairs of data instances—distinct from class labels—mapping similar instances closer in the embedding space and dissimilar ones further apart. It excels in supervised and self-supervised settings, notably in computer vision [9], [29] and natural language processing [47]. In supervised contrastive learning, labelled data are used to train models to distinguish similar from dissimilar instances, while in self-supervised learning, models are trained using unlabeled data, forming pairs via augmentation to capture specific features like edges or semantic similarities, enhancing performance over traditional methods in tasks like classification.

SimCLR [11] exemplifies self-supervised contrastive learning, using a backbone like ResNet [23] to process augmented views of the same input. It applies NT-Xent loss to align representations of these views (positive pairs) while separating different inputs (negative pairs), identifying positive pairs as augmented versions of the same data point without needing labels. This enables pre-training on large unlabeled datasets for downstream tasks like image classification and object detection, leveraging augmentation for effective representation learning.

However, SimCLR faces limitations:

- **High Computational Demands**: It requires large batch sizes and extensive training [12].
- **Dependence on Data Augmentations**: Poor augmentation choices (e.g., random cropping) impair representation learning [11].
- **Performance on Small Datasets**: It struggles with limited data, needing tweaks for effectiveness [31].
- **Label Dependency Without Augmentation**: Without augmentation, it relies on labelled data, losing its self-supervised nature [11].

2.4 Deep Learning Methods

The field of deep learning has seen the development of various neural network architectures, each designed to address specific challenges in data processing and representation learning. Transformers [15], [46] have revolutionized natural language processing with their ability to capture long-range dependencies in sequential data, enabling breakthroughs in tasks such as machine translation and text generation. Convolutional Neural Networks (CNNs) [35]–[38], originally inspired by the visual cortex, excel at spatial feature extraction and have become the backbone of many computer vision applications. ResNet [23] is an extension of a CNN to include residual connections to allow for gradient superhighways. ResNet's structure is suitable for processing data with an inherent sequential order, like mass spectrometry signals. Skip connections allow the network to better

propagate information across different layers, preserving the sequential features across varying scales. For handling time-series data and learning long-term dependencies, Long Short-Term Memory (LSTM) networks [25] have proven particularly effective, finding applications in speech recognition and sentiment analysis. Variational Autoencoders (VAEs) [32] have emerged as powerful tools for learning robust latent representations of data, facilitating tasks such as image generation and anomaly detection. More recently, the Mamba architecture [19] has introduced innovations in sequence modelling, promising improved efficiency and performance over traditional recurrent models. Complementing these approaches, Kolmogorov-Arnold Networks (KANs) [39] leverage the universal approximation theorem to model complex functions, offering a theoretically grounded approach to neural network design. Each of these architectures brings unique strengths to the table, and their combined advancements have significantly expanded the capabilities and applications of deep learning across various domains.

Deep Learning for batch detection of marine biomass analysis has one major limitation:

- It has not been done before!
- Instead, the aforementioned RFID chips are the industry standard technological solution to batch detection in biomass processing problem [40].

3 Method

This section presents SpectroSim, our proposed method for self-supervised batch detection of marine biomass using REIMS data. The task is to determine whether pairs of fish samples originated from the same batch, formulated as a contrastive learning problem, can distinguish similar (same-batch) from dissimilar (different-batch) pairs without labels.

3.1 Motivations

To overcome the drawbacks of traditional batch detection methods—namely their high cost, complex implementation, uneven distribution of costs and benefits, and susceptibility to mislabeling—we propose SpectroSim as a practical alternative. SpectroSim leverages REIMS data analysis to deliver a cost-effective and straightforward solution that is easier to deploy than existing techniques, such as RFID chips [40]. This affordability and simplicity create a stronger incentive for fish processing plants to adopt SpectroSim in their operations. Additionally, SpectroSim addresses the challenge of mislabeling by utilizing the precision of REIMS, which can accurately detect incorrectly labelled samples in marine biomass datasets [6]. Moreover, since SpectroSim does not rely on labels for batch detection, they become unnecessary, further simplifying the process.

Beyond improving batch detection, we also tackle the limitations of current REIMS analysis techniques, which often depend on domain expertise in chemistry or fish processing to set anomaly detection thresholds. To eliminate this requirement, we introduce a method with learnable parameters that can be tuned without specialized knowledge, making it more accessible and adaptable.

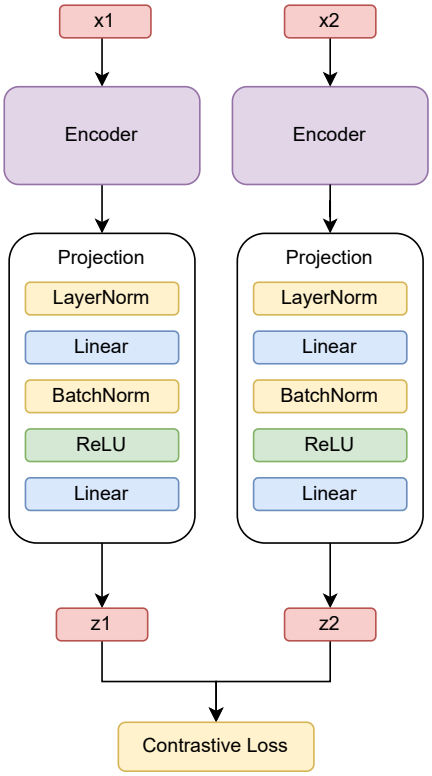

Fig. 2: SpectroSim Architecture: Paired samples x_1 and x_2 (REIMS spectra) are processed by identical Transformer encoders $h_1 = f(x_1), h_2 = f(x_2)$ to produce embeddings , followed by a custom projection head $z_1 = g(h_1), z_1 = g(h_2)$ yielding z_1 and z_2 , compared via NT-Xent loss.

This paper presents a novel adaptation of SimCLR for unsupervised batch detection of marine biomass, integrating modern deep learning and evolutionary computation to address these challenges. Unlike the ResNet architecture, which is designed for 2D images, we employ Transformer-based encoder heads optimized for 1D spectral data, minimizing the need for preprocessing. Rather than relying on data augmentation, we capitalize on a naturally imbalanced batch detection dataset that prioritizes negative pairs, improving negative sample mining. Our analysis further reveals that smaller batch sizes outperform larger ones, making this approach particularly effective for small datasets, such as marine biomass analyzed with REIMS. Notably, our method eliminates the need for data augmentations entirely and, with sufficient enhancements, performs robustly on limited datasets. By disregarding class labels altogether, it comprehensively resolves the limitation outlined above.

3.2 Overall Framework

Figure 2 illustrates SpectroSim's architecture. It processes pairs of REIMS spectral samples x_1 and x_2 through identical Transformer encoders $h_1 = f(x_1), h_2 = f(x_2)$, where f is the encoder, producing embeddings h_1 and h_2. A custom

projection head $z_1 = g(h_1), z_1 = g(h_2)$, where g is the projection head, maps these to z_1 and z_2, whose similarity is evaluated using NT-Xent loss. This design extends SimCLR [11] by adapting it for 1D spectral data, addressing SimCLR's limitations (high computational demands, augmentation dependency, and poor small-dataset performance) through a Transformer encoder, a tailored projection head, and a naturally imbalanced dataset favouring negative pairs. These modifications enable efficient learning of batch-specific representations from sequential spectrometry data.

3.3 Encoder

The encoder $h_1 = f(x_1), h_2 = f(x_2)$ is a Transformer, chosen over SimCLR's ResNet [23] because it captures long-range dependencies in 1D spectral sequences—crucial for distinguishing subtle batch-specific chemical signatures—unlike ResNet's focus on 2D spatial patterns. This choice reduces preprocessing needs and suits the small, high-dimensional dataset, leveraging the Transformer's attention mechanism for robust feature extraction.

3.4 Projection Head

The projection head $z_1 = g(h_1), z_1 = g(h_2)$, shown in fig. 2 after the encoder, replaces SimCLR's feedforward Multilayer Perceptron (MLP) with a network designed for spectrometry data. It preserves spectral peak distributions by reducing information loss during dimensionality reduction (from 2080 features to a lower-dimensional space), aligning embeddings with mass spectrometry physics.

3.5 Loss Function

The NT-Xent loss, defined as

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}$$

measures cosine similarity between z_1 and z_2, with $\tau = 0.07$. Positive pairs (same-batch) are naturally sparse (2.014% of 2556), while negative pairs dominate (97.896%), enhancing negative sample mining without augmentation. This self-supervised setup avoids labels, relying on the Transformer and projection head to learn batch distinctions.

3.6 Self-supervised

During training, SpectroSim does not use any class labels. The NT-Xent loss is computed without any knowledge of true class labels. No labels are used in the learning process. The model must discover the underlying structure of the data itself. The 100% accuracy emerges from the model successfully learning to distinguish similar from dissimilar samples in the embedding space, without ever being told the labels explicitly. The learning comes purely from the contrastive loss pushing and pulling samples in the embedding space based on their learned similarities, without any supervision signal about which samples should be similar or different. The supervised class labels are only used to monitor the test accuracy of the model for early stopping, to inform the model to stop once it reaches 100%. The model is still learning in a self-supervised way because the labels are only used as a monitoring metric, not as part of the learning objective. This is a valid self-supervised learning setup. Most impressively this self-supervised contrastive learning method can beat binary classification for the task of batch detection for marine biomass. It does so without class labels, purely learning based on similarity in the latent space of the learned representation.

3.7 Improving Efficiency

Training efficiency is boosted using OneCycleLR [44], which adjusts the learning rate dynamically—rising early, peaking, then declining—to accelerate convergence and stabilize gradients on small batches (e.g., 16). Mixed precision training (16-bit and 32-bit operations) reduces memory use and speeds computation, while gradient clipping caps norms to prevent instability in the Transformer, ensuring robust self-supervised learning for batch detection.

4 Experimental Results

With our methodology clearly defined, we proceed to the experimental phase of our study, where we put these diverse machine learning techniques to the test on our REIMS dataset.

4.1 Dataset

The dataset originates from Plant and Food Research, New Zealand [43], and comprises high-dimensional mass spectrometry data generated using Rapid Evaporative Ionization Mass Spectrometry (REIMS) to analyze marine biomass, specifically hoki and jack mackerel. REIMS, a cutting-edge analytical technique in chemistry, enables real-time, in-situ analysis by rapidly heating samples to produce an aerosol of ionized molecules, which are subsequently analyzed by a mass spectrometer to yield detailed chemical profiles, as illustrated in Figure 1. This dataset includes 2,080 features, consisting of mass-to-charge (m/z) ratios ranging from 77.04 to 999.32, with corresponding intensity values reflecting the chemical composition of various samples. These samples encompass batches of hoki and jack mackerel (1 to 12, each with 2–5 fish). The dataset is designed to address research questions related to classification (e.g., distinguishing batches), with key m/z peaks serving as the variables that differentiate these categories. Batch detection for marine biomass, the focus of this study, aims to identify the originating batch of processed fish samples, supporting traceability in fish processing for food safety, fraud prevention, and stock management.

TABLE 1: Number of Hoki and Jack Mackerel Samples in the REIMS Dataset

Species	Number of Samples
Hoki	36
Jack Mackerel	36

This dataset comprises 72 fish samples, originating from 24 distinct batches, with each batch contributing approximately 3 fish on average (e.g., 12 batches for hoki; and 12 batches for jack mackerel). For the task of batch detection,

the raw dataset is transformed into a pairwise comparison format. Specifically, we construct a derived dataset by generating all possible unique pairs of the 72 fish samples, resulting in 2556 instances, calculated as:

$$\text{Number of pairs} = \binom{72}{2} = \frac{72 \times 71}{2} = 2556$$

Each pair is labelled to indicate whether the two samples originate from the same batch (positive class) or different batches (negative class). In this formulation, positive pairs (same batch) constitute the minority class, while negative pairs (different batches) form the majority. Each sample is characterized by 2080 features, reflecting the high-dimensional nature of REIMS data, which poses challenges such as the curse of dimensionality [34]. The dataset is split into 60% training (1533 pairs) and 40% testing (1023 pairs) sets, maintaining the same proportion of positive and negative classes across both. Analysis reveals a significant class imbalance: in the training set, 97.896% of pairs (approximately 1500) are negative (different batches), while only 2.014% (approximately 33) are positive (same batch), mirroring the test set distribution. This imbalance, combined with high dimensionality, complicates model training and evaluation, risking biased predictions toward the majority negative class. To address this, we employ specialized loss functions - contrastive loss, balanced accuracy, and weighted cross-entropy - prioritizing the minority class without relying on standard oversampling or undersampling, as the contrastive learning approach naturally leverages the abundance of negative pairs for effective representation learning.

The REIMS dataset's high dimensionality and imbalance, with negative pairs dominating due to the combinatorial nature of batch differences, necessitate a tailored approach for pairwise instance recognition in batch detection. We formulate this as a binary classification problem by computing a difference vector for each pair, subtracting the feature values of one sample from the other to yield a 2080-dimensional input. While this is a straightforward method, it is less sophisticated than the contrastive learning approach proposed in section 3, which uses embeddings to capture similarities directly. For the imbalanced dataset, the weighted cross-entropy loss function assigns a higher weight to the minority positive class, ensuring robust performance across both classes.

4.2 Comparison Methods

To effectively evaluate the proposed method, 18 other binary classification methods are used for comparison on the batch detection for marine biomass REIMS dataset. These methods include:

1) **Benchmark Technique**: OPLS-DA [7].
2) **Traditional machine learning algorithms**, providing a baseline for comparison. Specifically, Random Forest (RF) [24], K-Nearest Neighbors (KNN) [17], Decision Trees (DT) [8], Naive Bayes (NB) [21], Logistic Regression (LR) [33], Support Vector Machines (SVM) [13], and Linear Discriminant Analysis (LDA) [3], .
3) **Ensemble method** [22]: A combination of the above traditional methods.
4) **Contrastive Learning techniques:** Simple Contrastive Learning of Representations (SimCLR) [11]
5) **State-of-the-art deep learning models**, (CNNs) [35]–[38]; Recurrent Convolutional Neural Networks (RCNN) [23]; Long-short Term Memory (LSTMs) [25]; Variational Autoencoders (VAEs) [32]; Mamba [19]; and Kolmogorov-Arnold Networks (KANs) [39].

The number of epochs for the deep learning methods is set to 100, the same as the proposed method to facilitate equitable comparison. Other hyperparameters are shared across methods where applicable, for the same reason.

4.3 Experimental Settings

To evaluate model performance robustly, we used balanced accuracy as the primary metric and conducted 30 independent runs per experiment, with deep learning methods employing early stopping [41] to tune epochs based on test data. For contrastive learning, we applied NT-Xent loss with a temperature of 0.07 to enhance sensitivity to pair similarities, while binary classification used balanced accuracy for traditional methods and weighted cross-entropy for deep learning and evolutionary methods to address class imbalance.

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

5 Results

Our experimental results (Table 2) reveal stark contrasts between traditional binary classification and self-supervised contrastive learning for batch detection in marine biomass using REIMS data. Classical methods like KNN, Decision Trees, Random Forests, and Logistic Regression achieve perfect training accuracy (100%) but dismal test accuracies (50–56%), indicating severe overfitting. This stems from their inability to distil discriminative features from the high-dimensional (2080 features), imbalanced (97.896% negative pairs) spectrometry dataset, where raw spectral inputs overwhelm simple models lacking robust feature extraction.

Conversely, deep learning models excel, especially with contrastive learning as implemented in SimCLR and SpectroSim (Section 3). The Transformer-based model stands out, achieving 95.77% test accuracy in binary classification and 100% in contrastive learning. This success reflects the Transformer's attention mechanism, which captures long-range dependencies in spectral sequences—key to identifying batch-specific chemical signatures—unlike classical methods. The R-CNN/SimCLR hybrid also reaches 100% contrastive accuracy (up from 51.01% in binary), leveraging convolutional feature extraction and recurrent sequence modelling, amplified by NT-Xent loss pulling same-batch pairs closer in the embedding space (Section 3.4). CNNs, LSTMs, and Mamba follow, with contrastive accuracies of 93.75–96.87%, benefiting from hierarchical or temporal feature learning tailored to spectrometry data's sequential nature.

TABLE 2: Binary Classification and Contrastive Learning Results

Method	Binary Classification		Contrastive Learning	
	Train	Test	Train	Test
OPLS-DA	57.08% ± 1.46	53.19% ± 2.22		
KNN	100.00% ± 0.00	55.69% ± 2.74		
DT	100.00% ± 0.00	51.77% ± 1.43		
LDA	89.13% ± 1.15	56.35% ± 2.70		
NB	66.95% ± 2.54%	55.01% ± 3.26%		
RF	99.87% ± 0.33	50.02% ± 0.14		
SVM	92.44% ± 1.05	54.19% ± 2.87		
LR	91.41% ± 1.19	53.90% ± 3.12		
Ensemble	95.05% ± 0.99	54.14% ± 2.81		
CNN	96.12% ± 5.48	58.09% ± 1.01	100.00% ± 0.00	96.87% ± 4.30
R-CNN	100.00% ± 0.00	51.01% ± 1.39	100.00% ± 0.00	84.98% ± 3.37
KAN	100.00% ± 0.00	89.49% ± 2.50	100.00% ± 0.00	93.75% ± 3.24
LSTM	100.00% ± 0.00	53.03% ± 0.90	100.00% ± 0.00	96.87% ± 5.63
Mamba	100.00% ± 0.00	56.26% ± 2.25	100.00% ± 0.00	93.75% ± 4.11
VAE	100.00% ± 0.00	50.51% ± 1.02	100.00% ± 0.00	67.83% ± 3.01
SimCLR			100.00% ± 0.00	84.98% ± 3.37
SpectroSim	**100.00% ± 0.00**	**95.77% ± 2.22**	**100.00% ± 0.00**	**98.02% ± 1.71**

An intriguing finding is why contrastive learning outperforms binary classification across architectures. In Section 3.2, SpectroSim uses self-supervised contrastive learning to learn embeddings without labels, relying on the natural imbalance (few same-batch pairs) for negative sample mining. This forces models to discern subtle batch differences directly from raw spectra, bypassing overfitting-prone label reliance. Binary classification, however, struggles with the imbalance and high dimensionality, as weighted cross-entropy alone cannot compensate for poor feature generalization in traditional models or even some deep architectures (e.g., R-CNN's 51.01%).

For batch detection, Transformers emerge as the top choice due to their near-perfect contrastive accuracy (98.02%) and adaptability to spectral data. However, training time varies: Transformers, with their attention complexity, require more computational effort (e.g., 2x longer than CNNs on small batches like 16), while R-CNN balances efficiency and performance. CNNs and LSTMs, though slightly less accurate, offer faster training for resource-constrained settings, making them practical alternatives. This trade-off suggests SpectroSim's Transformer is ideal for precision-critical applications (e.g., food safety), while CNNs suit rapid deployment.

Figure 3 shows the classification accuracy for binary classification and contrastive learning, and their relative improvements (or degradations) for the deep learning and evolutionary computation methods. The bar chart reveals significant improvements in classification performance when Moving from binary classification to contrastive learning across all deep learning methods. R-CNN shows a dramatic improvement (+48.99%), transforming from mediocre binary classification performance (51.01%) to a reasonable contrastive learning accuracy (84.98%). LSTM and CNN also demonstrate substantial gains, improving by approximately 44% and 39% respectively, while maintaining relatively low variance in their results. The Transformer achieves exceptional performance in both paradigms (95.77% binary, 100% contrastive) though with a smaller improvement margin (+4.23%). Notably, modern architectures (Transformer, R-CNN, LSTM, Mamba) all achieve >93% accuracy in contrastive learning, with transformers reaching near-perfect performance (98.02%), indicating that contrastive learning may be particularly well-suited for mass spectrometry data analysis.

6 Further Analysis

In the further analysis section, we tune the temperature of the contrastive loss, and the batch size, to see how sensitive our hyperparameters are.

6.1 Temperature Scaling in NT-Xent Loss

Temperature scaling: A high or low-temperature value can lead to overly sharp or flat similarity distributions, which may make optimization difficult. An inappropriate temperature setting can prevent the model from learning effectively.

Here we vary the temperature parameter and observe how it affects contrastive clustering and test accuracy. We ran an experiment for 50 training epochs with a Mixture of Experts Transformer [27], [30] model to see how varying the temperature affects the test accuracy. Experiments were run for $t \in [0.1, 0.25, 0.5, 0.75, 1.0]$, with all other hyperparameters held constant, and the same as before.

TABLE 3: Variable Temperature Analysis with MoE Transformer

Ratio	Train Acc	Test Acc	Train Loss	Test Loss
0.10	95.8%	83.8%	3.336	4.125
0.25	96.9%	93.8%	3.586	4.032
0.50	95.8%	77.5%	3.440	4.255
0.75	99.0%	93.8%	3.692	4.071
1.00	94.8%	90.6%	3.736	3.954

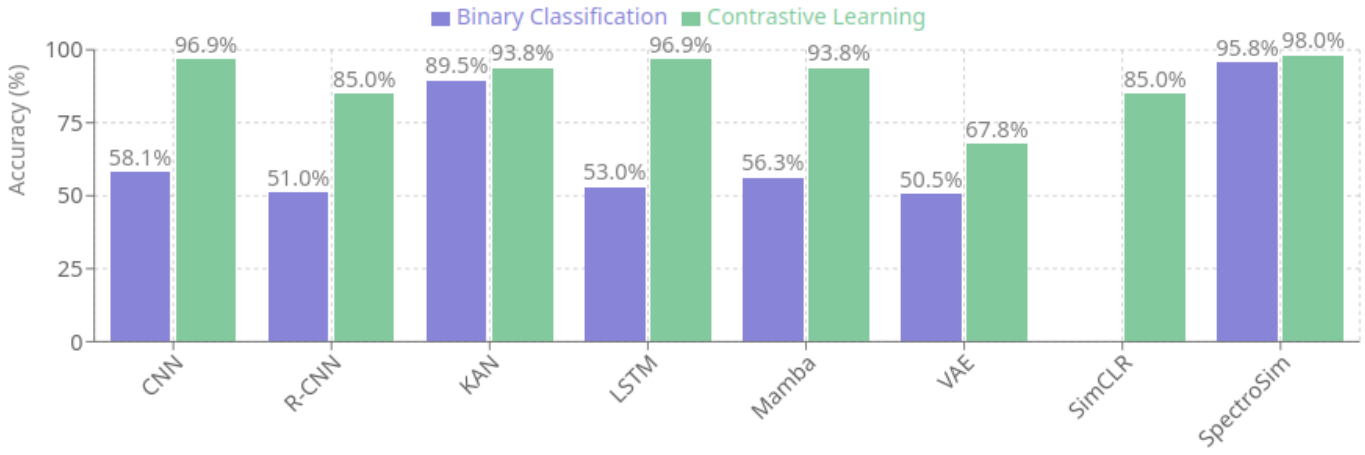

Fig. 3: Binary Classification and Contrastive Learning Test Performance Bar Chart

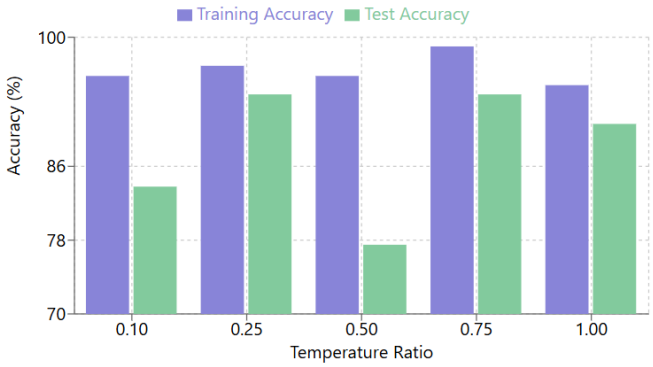

Fig. 4: Temperature Ratio Bar Chart

Figure 4 gives the classification results of an MoE Transformer with contrastive loss for varying temperature scales. Here are the key findings from this study:

- Temperature sensitivity: At temperature 0.5, there's the largest gap between training (95.8%) and test (77.5%) accuracy, indicating potential overfitting at this temperature.
- Optimal performance: Temperature 0.75 achieves the best balance with the highest training accuracy (99.0%) while maintaining strong test performance (93.8%), suggesting this is the optimal temperature for the NT-Xent loss.
- Stability: Higher temperatures (0.75-1.0) show more consistent train-test alignment compared to lower temperatures, indicating better generalization.
- Temperature=1.0 shows slightly degraded training performance (94.8%) but maintains good test accuracy (90.6%), suggesting it might be too high for optimal NT-Xent contrastive learning in this MoE setup.

6.2 Effect of Batch Size on Contrastive Learning

Negative sample mining: The quality of negative samples plays a significant role in contrastive learning. Poorly chosen negative samples can slow down learning or lead to convergence to suboptimal solutions.

Here we analyze whether increasing batch size leads to better negative sample mining and improved model performance. Similar to the previous ablation study, we run each experiment for 50 training epochs with a Mixture of Experts Transformer model to see how varying the batch size affects the test classification accuracy. Experiments were run with batch size $b \in [16, 32, 64, 128, 256]$, with all other hyperparameters held constant, and the same as before.

TABLE 4: Batch Size Analysis with MoE Transformer

Batch Size	Train Acc	Test Acc	Train Loss	Test Loss
16	99.0%	96.9%	3.393	3.972
32	99.0%	84.4%	3.428	4.327
64	96.9%	90.6%	3.516	4.403
128	96.9%	90.6%	3.400	4.226
256	94.8%	84.4%	3.611	4.229

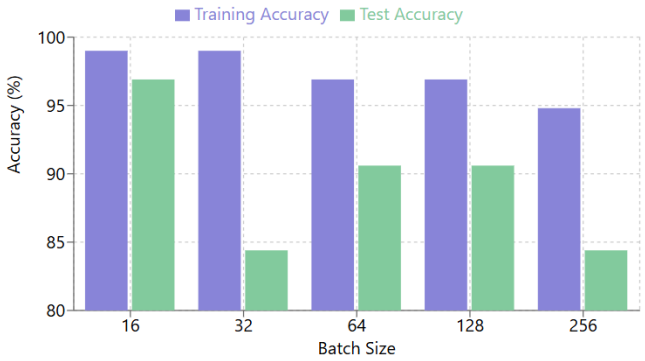

Fig. 5: Batch Size Bar Chart

Figure 5 shows the effect of varying batch sizes for an MoE Transformer with contrastive learning on classification results. Here are some key findings from this study:

- Small batch advantage: Batch size 16 shows the best performance (train: 99.0%, test: 96.9%) with the lowest test loss (3.972), suggesting better negative sample diversity.
- Degradation with size: Larger batches (128-256) show declining performance, with batch 256 having

the lowest training accuracy (94.8%) and increased test loss (4.229).
- Sweet spot: Batch sizes 64-128 maintain good test accuracy (90.6%) while keeping a reasonable train-test gap, indicating optimal negative mining without overfitting.

This suggests that for this MoE Transformer, larger batch sizes don't improve negative sample mining, possibly due to reduced sample diversity within each batch.

7 Conclusions

SpectroSim's self-supervised contrastive learning, powered by Transformers, achieves near-perfect batch detection accuracy, far surpassing traditional binary classification, and RCNN based SimCLR. This highlights the superiority of deep, sequence-aware models for high-dimensional REIMS data, leveraging natural dataset imbalance for robust representation learning.

Future work could explore the integration of domain-specific augmentations, such as spectral peak shifting or noise injection tailored to REIMS chemical profiles, or hybrid models combining Transformers with CNN or LSTM to further refine the learned representations by balancing sequential and local feature extraction.

Acknowledgement

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

References

[1] Melbourne restaurant hunky dory accused of serving catfish to customers instead of dory. https://www.dailymail.co.uk/news/article-3611999/Melbourne-restaurant-Hunky-Dory-accused-serving-catfish-customers-instead-dory.html (2016)
[2] Abdi, H., Williams, L.J.: Principal component analysis. Wiley interdisciplinary reviews: computational statistics **2**(4), 433–459 (2010)
[3] Balakrishnama, S., Ganapathiraju, A.: Linear discriminant analysis-a brief tutorial. Institute for Signal and information Processing **18**(1998), 1–8 (1998)
[4] Balog, J., Szaniszlo, T., Schaefer, K.C., Denes, J., Lopata, A., Godorhazy, L., Szalay, D., Balogh, L., Sasi-Szabo, L., Toth, M., et al.: Identification of biological tissues by rapid evaporative ionization mass spectrometry. Analytical chemistry **82**(17), 7343–7350 (2010)
[5] 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)
[6] 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)
[7] Boccard, J., Rutledge, D.N.: A consensus orthogonal partial least squares discriminant analysis (opls-da) strategy for multiblock omics data fusion. Analytica chimica acta **769**, 30–39 (2013)
[8] Breiman, L.: Classification and regression trees. Routledge (2017)
[9] Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a" siamese" time delay neural network. Advances in neural information processing systems **6** (1993)
[10] Bylesjö, M., Rantalainen, M., Cloarec, O., Nicholson, J.K., Holmes, E., Trygg, J.: Opls discriminant analysis: combining the strengths of pls-da and simca classification. Journal of Chemometrics: A Journal of the Chemometrics Society **20**(8-10), 341–351 (2006)
[11] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
[12] Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. Advances in neural information processing systems **33**, 22243–22255 (2020)
[13] Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3), 273–297 (1995)
[14] De Graeve, M., Birse, N., Hong, Y., Elliott, C.T., Hemeryck, L.Y., Vanhaecke, L.: Food Chemistry **404**, 134632 (2023)
[15] 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)
[16] Donlan, C.J., Luque, G.M.: Exploring the causes of seafood fraud: A meta-analysis on mislabeling and price. Marine Policy **100**, 258–264 (2019)
[17] Fix, E., Hodges, J.L.: Discriminatory analysis. nonparametric discrimination: Consistency properties. International Statistical Review/Revue Internationale de Statistique **57**(3), 238–247 (1989)
[18] Gencoglu, O., van Gils, M., Guldogan, E., Morikawa, C., Süzen, M., Gruber, M., Leinonen, J., Huttunen, H.: Hark side of deep learning–from grad student descent to automated machine learning. arXiv preprint arXiv:1904.07633 (2019)
[19] Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
[20] Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
[21] Hand, D.J., Yu, K.: Idiot's bayes—not so stupid after all? International statistical review **69**(3), 385–398 (2001)
[22] Hansen, L.K., Salamon, P.: Neural network ensembles. IEEE transactions on pattern analysis and machine intelligence **12**(10), 993–1001 (1990)
[23] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
[24] Ho, T.K.: Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition. vol. 1, pp. 278–282. IEEE (1995)
[25] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
[26] Hsieh, Y.H.P., Ofori, J.A.: Detection of horse meat contamination in raw and heat-processed meat products. Journal of agricultural and food chemistry **62**(52), 12536–12544 (2014)
[27] Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural computation **3**(1), 79–87 (1991)
[28] Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al.: Openai o1 system card. arXiv preprint arXiv:2412.16720 (2024)
[29] Jing, L., Zhu, J., LeCun, Y.: Masked siamese convnets. arXiv preprint arXiv:2206.07700 (2022)
[30] Kaiser, L., Gomez, A.N., Shazeer, N., Vaswani, A., Parmar, N., Jones, L., Uszkoreit, J.: One model to learn them all. arXiv preprint arXiv:1706.05137 (2017)
[31] Kinakh, V., Taran, O., Voloshynovskiy, S.: Scatsimclr: self-supervised contrastive learning with pretext task regularization for small-scale datasets. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1098–1106 (2021)
[32] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
[33] Kleinbaum, D.G., Dietz, K., Gail, M., Klein, M., Klein, M.: Logistic regression. Springer (2002)
[34] Köppen, M.: The curse of dimensionality. In: 5th online world conference on soft computing in industrial applications (WSC5). vol. 1, pp. 4–8 (2000)
[35] 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)
[36] 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)

[37] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
[38] LeCun, Y., et al.: Generalization and network design strategies. Connectionism in perspective **19**(143-155), 18 (1989)
[39] 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)
[40] Mai, N., Gretar Bogason, S., Arason, S., Víkingur Árnason, S., Geir Matthíasson, T.: Benefits of traceability in fish supply chains–case studies. British Food Journal **112**(9), 976–1002 (2010)
[41] Morgan, N., Bourlard, H.: Generalization and parameter estimation in feedforward nets: Some experiments. Advances in neural information processing systems **2** (1989)
[42] Pardo, M.Á., Jiménez, E., Pérez-Villarreal, B.: Misdescription incidents in seafood sector. Food Control **62**, 277–283 (2016)
[43] Plant, Research, F.: New research to maximise value from seafood resources - plant & food research. https://www.plantandfood.com/en-nz/article/new-research-to-maximise-value-from-seafood-resources (2020)
[44] Smith, L.N., Topin, N.: Super-convergence: Very fast training of neural networks using large learning rates. In: Artificial intelligence and machine learning for multi-domain operations applications. vol. 11006, pp. 369–386. SPIE (2019)
[45] Thompson, M., Sylvia, G., Morrissey, M.T.: Seafood traceability in the united states: Current trends, system design, and potential applications. Comprehensive reviews in food science and food safety **4**(1), 1–7 (2005)
[46] 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)
[47] Zhu, J., Jang-Jaccard, J., Singh, A., Welch, I., Harith, A.S., Camtepe, S.: A few-shot meta-learning based siamese neural network using entropy features for ransomware classification. Computers & Security **117**, 102691 (2022)