Transformers: roBERTs in disguise

*Rapid determination of bulk composition and quality of marine biomass in Mass Spectrometry

1st Jesse Wood
School of Engineering
Victoria University of Wellington
Wellington, New Zealand
jesse.wood@ecs.vuw.ac.nz

2nd Bach Hoai Nguyen
School of Engineering
Victoria University of Wellington
Wellington, New Zealand
hoai.bach.nguyen@ecs.vuw.ac.nz

3th Bing Xue
School of Engineering
Victoria University of Wellington
Wellington, New Zealand
bing.xue@ecs.vuw.ac.nz

4th Mengjie Zhang
School of Engineering
Victoria Unviersity of Wellington
Wellington, New Zealand
mengjie.zhang@ecs.vuw.ac.nz

5th Daniel Killeen
Science Team Leader
Plant and Food Research Limited
Nelson, New Zealand
daniel.killeen@plantandfood.co.nz

***Abstract*—Navigating mass spectrometry data analysis for marine biomass and fish demands a technologically adept approach to derive accurate and actionable insights. This research will introduce a novel AI methodology to interpret a repository of rapid evaporative ionisation mass spectrometry datasets, utilizing transformers with pre-training strategies like Next Spectra Prediction and Masked Spectra Modeling. Also, it employs decision trees and genetic programming for enhanced interpretability and correlation of spectral patterns with chemical attributes. This paper proposes a toolbox of machine-learning methods for marine biomass analysis. Two core research objectives are explored: 1) precise fish species and 2) body part identification via binary and multi-class classification, respectively.**

***Index Terms*—classification, deep learning, explainable AI, genetic programming, machine learning, machine learning, mass spectrometry, transformers**

I. Introduction

Waste utilization in the global fishing industry needs improvement. As of 2020, approximately 100 million tonnes of wild fish are caught each year, but only about 40% of these fish are processed into edible parts [1]. The remaining portions are often processed into fish oil and fish meal, or discarded. In addition, many fisheries are in decline, despite global fishing not significantly increasing in the past 30 years, making waste utilization an important focus worldwide. The fishing industry must maximize the utilization and value of every kilogram of marine biomass to preserve fish stocks and ensure there are plenty of fish in the sea for future generations to reel in.

Identification provides relevant information to profile a sample of marine biomass. These profiles include, but are not limited to, the species of the fish, and the body part from which the sample was taken. In fish processing, a fish once rendered into a minced product, paste or oil, is completely unrecognizable from when it freely swam the oceans. Therefore, chemistry techniques can be used to retrieve this lost information and identify the contents of rendered marine biomass. Useful characteristics, such as species, and body parts; help decide how best to use that marine biomass. For example, due to variations in chemical composition between species and parts, some contain larger quantities of fatty oils, they can be repurposed into Omega-3 supplements. The Cyber-Marine flex-factory [2] aims to maximize waste utilization of marine biomass. Therefore, identifying characteristics of marine biomass waste, such as its species and body part, is useful. This knowledge informs decisions on how best to reduce, reuse and recycle that waste, to maximize the value of that marine biomass.

Existing works into identification of biomass, let alone marine biomass, using rapid spectrometry are limited [3], [4]. Due to rapid evaporative ionisation mass spectrometry (REIMS) [5] being a recent technological development in chemistry, and the diffusion of innovation [6], access to the REIMS mass spectrometer, and subsequent research, and real-world applications of said technology, is sparse. The tools are cost-prohibitive for widespread adoption and use in industry. However, as part of the greater Cyber-Marine research project [2] serves as a proof-of-concept, for the adoption of REIMS for rapid analysis of marine biomass in the factory of the future - the flex-factory. This research aims will show the viability of rapid mass spectrometry in real-world applications of fish processing. Rapid spectrometry has been shown effective in detecting adulteration in biomass, [3] found beef mince that was contaminated with horse meat. Adulteration is the (often criminal) process of debasing the quality of food products, by intentionally mixing them with products of lower value, to maximize profits, and dishonestly selling them labelled as ONLY the higher value product [5]. The study, [3], showed that REIMS can detect adulteration of beef samples with cross-species contamination at levels as low as 1%, for certain horse-meat offal. Rapid spectrometry has demonstrated a use-case in marine biomass when identifying species of marine biomass for the real-world application of

Industry partner Plant and Food Research

fish fraud detection [3]. Previous works [3]–[5] demonstrate that REIMS can be used to combat fraud and adulteration in food processing. This research aims to apply this method of analysis for determining the bulk composition and quality of marine biomass.

Firstly, to apply rapid mass spectrometry methods to fish processing in New Zealand, this paper tackles the unique market of New Zealand's seafood industry. Unlike other countries, for example, Canada or the United States, New Zealand has a high variability in marine biomass. In layman's terms, when a catch comes in from a fishing vessel, there is a diverse range of species, in that catch. The catches coming from trawling vessels in Canada or the United States consist mostly of one species - a homogeneous composition of marine biomass. However, the catches coming in from New Zealand vessels, consist of a diverse range of species - a heterogeneous composition of marine biomass. This translates to a multi-class problem with many classes in machine learning. This work focuses on a binary and multi-class classification problem, with a dataset containing two fish species, and six fish parts.

Secondly, a factor unique to New Zealand's seafood industry, and due to a much smaller fishing fleet and population, is a low sample size. Large trawling vessels in international waters, or the United States or Canada, have a large volume of homogeneous marine biomass, to collect and analyze with chemistry methods. Due to New Zealand's smaller size, and isolated geographical location, there is a much smaller volume of fish to create datasets from for analysis via chemistry techniques. As demonstrated in previous works [7], fish analysis for New Zealand marine biomass, is performed on high-dimensional data with data scarcity. Mutli-class problems with limited training instances would usually hinder the use of deep learning methods. However, this work introduces transfer learning via pre-training, and data augmentation, to amortize the limited number of training instances.

II. Related Works

In previous work [7], the EC technique of particle swarm optimisation (PSO) [8] was used for feature selection in fish species and part identification, on a gas chromatography fish oil dataset. In this work, for that same task of fish species and part classification, the EC techniques of genetic programming (GP) [9], specifically, multi-tree genetic programming (MT-GP) [10], are used for feature construction and classification. Multiple class-independent feature construction method (MCIFC) [11] is presented in a novel application in this research.

In 1984, Breiman et al [12] proposed classification and regression tree (CART). Decision trees have been shown effective on mass spectrometry datasets. In the late 1970s and 1981, the environmental protection agency (EPA) funded projects for classification trees to recognize individual elements in chemical compounds by analysing their mass spectra [12]. The EPA collect water and air samples to detect toxic substances. The classification trees attempt to work out how fragment ions of the original molecule were pieced together. The decision tree found correlations in mass spectra for a variety of complex molecules. For the classification task of fish species and part identification, a classification tree is used. The classification task predicts the class label an instance is most likely to belong to. CART uses a tree-based structure of both nodes, branches and leaves. The nodes are where decisions (e.g. splits) are made, branches give the outcome of those decisions (e.g. the resulting subsets), and leaves give the predicted class label (e.g. prediction). classification and regression tree (CART) is a greedy algorithm. It first considers all possible splits and selects the split that reduces the impurity the most. A low Gini impurity means the subset is more pure. It splits at the feature that is the best splitting point. CART continues to split until a stopping rule is met, or no further best splits are available.

In Black et al. [3], REIMS data modelled with PCA-LDA was able to detect species and catch method. cross-species contamination is a more complex variation of this problem. In [3], each sample belonged to one species, however, for this problem, each sample can belong to multiple classes, e.g. a cross-species contaminated sample contains a mixture of two species. Black et al. [4] performed detection and identification beef adulteration. It can identify samples that are adulterated with offal, and specify which offal was present. This is not marine biomass, but instead machine learning analysis of rapid mass spectrometry applied to animal agriculture. Everything in their work [4] translates to this paper's research, except the animal agriculture domain. This research focuses on marine biomass, whereas their research focuses on animal agriculture. The insights from their analysis are likely universal and can be applied to marine biomass analysis. Rapid mass spectrometry measurement techniques are a relatively new and niche innovation [13], so work from different application domains, should be considered.

This section provides a summary of the limitations of the existing work, and how this paper intends to fill those gaps. In particular, the research will focus on domain knowledge, state-of-the-art, transfer learning, online learning and taxonomy.

- **Domain knowledge** - The thresholds to determine outliers are determined manually by domain experts in [3], [4]. Their expertise in chemistry is needed to choose hyperparameters for every model - time. Significant markers are analysed and identified post hoc, relying on domain expertise in chemistry and human intuition. Manual hyper-parameter tuning (e.g. # principal components, RSD threshold for outliers, mass range) can be automatically selected, or replaced by models that don't need them at all!
- **State-of-the-art** - Mature statistical techniques are used for dimensionality reduction and classification, not state-of-the-art machine learning. Basic supervised statistical models (e.g. LDA, OPLS-DA [3], [4]) were used for classification. This paper introduces novel applications

for Transformers [14], [15] and Evolutionary Computation [7], [10], [11]. Future work should consider CNNs [16], [17], generative adversarial networks (GAN)s [18], Diffusion [19], [20].

- **Transfer learning** - There is a large body of existing mass-spectrometry data [21], [22]. Knowledge from these datasets is not incorporated. Potential for transfer learning (incorporate previously existing data) to improve performance for few-shot classification tasks. This work proposes two new transfer learning methods, masked spectra modelling (MSM) and next spectra prediction (NSP).
- **Online learning** - Many AI models completely collapse when presented with new data, whether that be out-of-distribution anomalies [18], or conceptual drift [23], [24] where the underlying probability distribution changes over time - for example seasonal variation in composition of Hoki [25]. A flex factory needs robust models, that can be updated with new information, and an online learning scenario, where edge cases are fed back as training data, to make them more robust.
- **Taxonomy** - Previous works [3]–[5], [13] in this field have been from a purely chemistry and statistics background. This work aims to introduce this problem to machine learning practitioners. The glossary in this paper works towards providing a taxonomy to foster future multi-disciplinary collaboration. It breaks down the esoteric technical jargon from chemistry, fish processing, machine learning, and statistics.

III. Theory

A. Classifiers

Classification, within the context of machine learning and statistics, refers to the task of categorizing input data into predefined classes or categories based on their characteristics or features. The goal of classification is to learn a mapping from input features to class labels, enabling the model to predict the class of new, unseen instances accurately. In a classification problem, the input data is typically represented as feature vectors, where each feature corresponds to a measurable aspect or attribute of the data. These features serve as the basis for making predictions, with the model learning patterns and relationships in the data that differentiate one class from another. Classification algorithms employ various techniques, such as decision trees, support vector machines, and neural networks, to learn the underlying patterns in the data and make predictions about the class labels of new instances. Classification finds widespread applications across domains such as image recognition, text categorization, spam detection, medical diagnosis, and customer segmentation, among others, facilitating automated decision-making and pattern recognition tasks.

An ensemble of 7 standard classifiers is chosen for the two classification tasks of fish species and fish parts identification. Additionally, there is an ensemble voting classifier that combines those classifiers too. Specifically, those classifiers and their ensemble are:

1) Random forest (RF) [26]
2) K-nearest neighbours (KNN) [27]
3) Decision tree (DT) [12]
4) Naive bayes (NB) [28]
5) Logistic regression (LR) [29]
6) Support Vector Machine (SVM) [30]
7) Linear discriminant analysis (LDA) [31]
8) Ensemble voting classifier [12], [26]–[29], [29]–[31]

These classifiers are used as benchmarks to evaluate the results, and compared to Transformers and multi-tree genetic programming (MT-GP) for their balanced classification accuracy on fish species and fish part, binary and multi-class classification, respectively. All classifiers use their default configurations from sklearn [32], except, the SVM [30] uses a linear kernel, and the LR [29] has max iterations set to 2,000. The Ensemble voting classifier combines RF, DT, NB, LR, SVM, LDA with hard voting.

B. Transformer

In 2017, Vaswani et al [14] proposed the transformer, for machine text translation. This paper uses the same encoder-decoder architecture as Vaswani et al for marine biomass analysis.

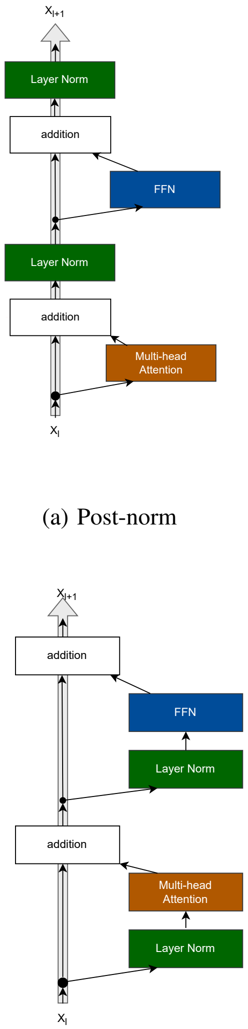
This model revolutionized the field of natural language process. A stacked encoder and decoder architecture, with residual connections between [33] layers. Since then not much has changed; different strategies for weight initialization, such as Xavier, Kaiming, and Orthogonal [34]–[36] have proven useful, and the pre-norm formulation [37], [38] for layer normalization [39].

The transformer model presented in this paper utilized the AdamW [40], addressing the issue of Adam [41], by decoupling the weight decay from the learning rate.

Dropout [42], [43] provides an efficient approximation to a bagged ensemble of many neural networks [44]. Dropout trains that ensemble of all possible subnetworks that can be made by removing non-output units from the base network. According to a dropout probability $p(d) = 0.2$, dropout randomly turns off sub-networks of neurons, by multiplying their output with zero. This forces the network to learn many sub-representations that capture the same pattern in the data. This leads to better generalization of unseen data.

Label smoothing adds noise to the class labels, as seen in fig. 2. Label smoothing introduces a penalty to the model's confidence, resulting in more conservative probability estimates. As a result the activations in the penultimate layer, which feed into the final softmax layer, may show a more diffuse, or smooth, distribution with label smoothing enabled. Label smoothing replaces one-hot encoded label vector y_hot with a mixture of y_hot and the uniform distribution. Label smoothing is done by the following formula:

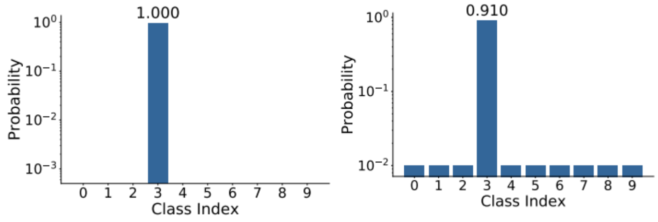
$$y_{ls} = (1 - \alpha) * \hat{y} + \frac{\alpha}{K}$$



(a) Post-norm

(b) Pre-norm

Fig. 1: Post-norm (left) Pre-norm (right) formulation



(a) Hard Label

(b) LS

Fig. 2: Label smoothing

where α is a hyperparameter that determines the amount of smoothing, $\alpha = 0$ is no smoothing with the default one-hot encoded labels. $\alpha = 1$ is a uniform distribution with the one hot encoded label information lost. This work uses $\alpha = 0.1$ as the default label smoothing value.

The GELU [45] activation function is used for the transformer network. GELU is a smooth activation function, which means it has continuous derivatives everywhere. This property aids smoother convergence with gradient descent. GELU introduces a non-linearity to the network, allowing it to model more complex relationships between inputs and outputs. GELU can be defined mathematically as:

$$GELU(x) = 0.5 * x * (1 + Tanh(\sqrt{2/\pi} * (x + 0.044715 * x^3)))$$

Empirical studies [45]–[47] have shown that GELU tends to perform well in various deep learning tasks, including image classification, natural language processing, and speech recognition. It has been observed to yield faster convergence and better generalization compared to traditional activation functions like ReLU [48], [49]. GELU helps alleviate the vanishing gradient problem, which can occur during the training of deep neural networks, e.g. the transformer presented here. By providing non-zero gradients for all inputs, GELU helps propagate gradient information effectively through the network, facilitating better learning of deep representations.

Data augmentation is explored to artificially inflate the volume of training instances [44]. Noise taken from a normal distribution $\mathcal{N}(\mu, \sigma)$, is randomly added to 5 copies of each training instance. This increases the training set fivefold. The validation and test set are not augmented, only the training set. Injecting noise into the input of a neural network is a form of data augmentation [50], [51].

In this paper, two novel methods of unsupervised pre-training are proposed. They draw inspiration from similar techniques in the field of natural language processing, that were used to pre-train the BERT large language model [15]. The first is a variation of Masked Language Modelling (MLM). In BERT, MSM would randomly mask words in a sentence, and pre-training would look to predict those missing words. An example of this can be seen here in fig. 3. This paper presents a variation of this technique for mass spectrometry, Masked Spectra Modelling. Where rather than sentences, the pre-training deals with mass spectra. Randomly masking mass-to-charge ratios, where pre-training predicts the missing spectra. This is a regression task where Mean Square Error is used to evaluate the performance, on a training and validation set, with early stopping.

The second unsupervised pre-training method is based on Next Sentence Prediction (NSP). Next sentence prediction is a pair-wise comparison task where the model must predict if two sentences follow each other. If the model thinks the sentences are related, it returns a match, otherwise it returns no match. An example of this can be seen in fig. 4. The variation of NSP proposed in this work is Next Spectra Prediction (NSP), where spectra are spliced in half, given the left and right-hand

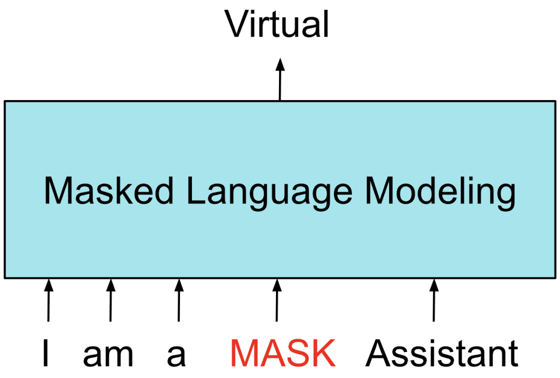


Fig. 3: Masked language modelling

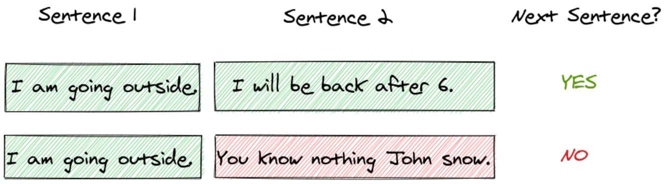


Fig. 4: Next sentence prediction

sides of two spectra, to predict whether they belong to the same original or different spectra. This is a binary classification task where categorical cross entropy is used as the loss function, evaluated on a training and validation set, with early stopping.

Early stopping [52] stores a copy of the model parameters every time the loss on the validation set improves. At the end of the training, the model parameters with the lowest loss are returned as the final model. The algorithm ends when no parameters have improved over the best validation error, for a pre-specified number of iterations, called patience, where $p = 5$. It is a very common regularization technique in deep learning. Early stopping can be considered hyper-parameter tuning over the number of epochs [44].

TABLE I: Parameter settings

Learning rate	1E-5
Epochs	100
Dropout	0.2
Label smoothing	0.1
Early stopping Patience	5
Optimizer	AdamW
Loss: MSM	Mean Squared Error
Loss: NSP	Categorical Cross Entropy
Loss: fish & species	Categorical Cross Entropy
Input dimensions	1023
Hidden dimensions	128
Output dimensions: MSM	1023
Output dimensions: NSP	2
Output dimensions: species	2
Output dimensions: part	6
Number of layers	3
Number of heads	3

Table I delineates the parameter configurations of the transformer methodology applied in the experiments, characterized by an encoder-decoder architecture. The input dimension corresponds to the number of features, set at 1023. For pre-training, the output dimensions are 1023 and 2, for the masked spectra modelling (MSM) and next spectra prediction (NSP) tasks, respectively. Subsequently, for fish species binary classification and fish parts multi-class classification, the output dimensions are 2 and 6, respectively. The transformer model is composed of stacked encoder and decoder blocks, containing 3 encoders and 3 decoders. The encoder incorporates multi-head attention mechanisms with 3 heads, comprising a self-attention layer followed by a simple feedforward neural network layer. Similarly, the decoder integrates a self-attention layer and a feedforward neural network layer, supplemented by a cross-attention layer, facilitating the infusion of information from the input sequence into the decoder layers.

C. Mutli-tree Genetic Programming

Multi-tree genetic programming (MT-GP) is explored for binary classification of fish species, and multi-class classification of fish parts. This technique provides a model with explainable output that can easily be interpreted, unlike the aforementioned transformer models, or other deep neural networks. This paper introduced multiple class-independent feature construction method (MCIFC) from [11] for novel applications in marine biomass analysis.

The representation of the candidate solutions is multiple trees, with one subtree for each class. This method provides both feature construction and classification. The $argmax$ of a multi-tree individual provides the prediction for the classification task. For fish species, there are two classes: Hoki and Mackerel - two trees, one for each class. For fish parts, there are six classes: Fillet, Heads, Livers, Skins and Guts - with six subtrees, one for each class. This feature construction reduces the feature space from 1023 to 2 or 6, respectively.

In multiple class-independent feature construction method (MCIFC) there are two genetic operators, crossover and mutation. These both behave slightly differently [11] than in convention genetic programming (GP) [9]. Since there is one subtree for each class, crossover only occurs between trees of the same class. The crossover operator randomly selects two trees of the same class to perform crossover, with a chance of 80% $p = 0.8$. The mutation operator randomly selects one subtree, and performs mutation, with a chance of 20% $p = 0.2$. Using VarAnd, the GP, each generation can perform crossover and/or mutation, i.e. one or the other, or both. The algorithm uses tournament selection with a tournament size of 7.

The fitness is evaluated as the accuracy metric and a distance regularization term. For accuracy, it measures the balanced accuracy score of the constructed features. The distance regularization term introduces a penalty for intraclass distance, and rewards interclass distance. The distance measure is simply the Euclidian distance between pairs of points, given by

TABLE II: Parameter settings

Function Set	$+, -, \times, \cos, \sin, \tan, -1*$
Terminal Set	$x_1, x_2, ..., x_n$
Maximum Tree Depth	6
Population size	1 * 1023 (= 1× #features)
Initial Population	Ramped Half and Half
Generations	400
Crossover	0.8
Mutation	0.2
Elitism	0.1
Selection	Tournament
Tournament Size	7
Construction ratio	1
Fitness weighting α	0.8

$$d(i,j) = \sqrt{\sum_{k=1}^{k} (i_k - j_k)^2}$$

The interclass distance measures the distance between pairs of different classes. The larger the distance, the more the reward in the fitness function. This relies on the assumption that two instances from different classes should be further away from each other than two instances of the same class. The interclass distance is given by:

$$\frac{1}{|S|} \sum_i \sum_j d(S_i, S_j) \quad \forall \quad i \neq j \quad \text{and} \quad class(i) \neq class(j)$$

Conversely, the intraclass distance measures the distance between pairs of the same class. The smaller the distance, the more the reward in the fitness function. This relies on the assumption that two instances from the same class should be closer to each other, than two instances from different classes. The intraclass distance is given by:

$$\frac{1}{|S|} \sum_i \sum_j d(S_i, S_j) \quad \forall \quad i \neq j \quad \text{and} \quad class(i) = class(j)$$

Table II describes the parameter settings of the GP-based method used in the experiments. The function set has standard arithmetic operators $+, -, \times$, trigonometry operators $\sin, \cos, \tan$ and the unary neg operator reverses the sign. The feature set, A population of 1023 individuals, the same as the number of features, is used for all experiments, with 400 generations. The construction ratio r used to determine the number of features constructed is experimentally chosen as 1. The fitness weight α is set to 0.8 to bias fitness values towards accuracy.

IV. Dataset

Rapid evaporative ionisation mass spectrometry (REIMS) is one of the newest forms of ambient mass spectrometry (AMS) and, as is the case with many analytical innovations was created for medical research purposes. It operates using an electro-surgical knife, bipolar forceps or laser which creates an aerosol (smoke) when cutting into a tissue sample. The aerosol is evacuated from the sample through a transfer line into the ionisation source of a mass spectrometer where a heated collision surface is situated and the ionisation process occurs. This description is an excerpt from [3]. Figure 5 gives the mass spectrums for the entire REIMS dataset. This gives an intuition for the range and variability across these measurements.

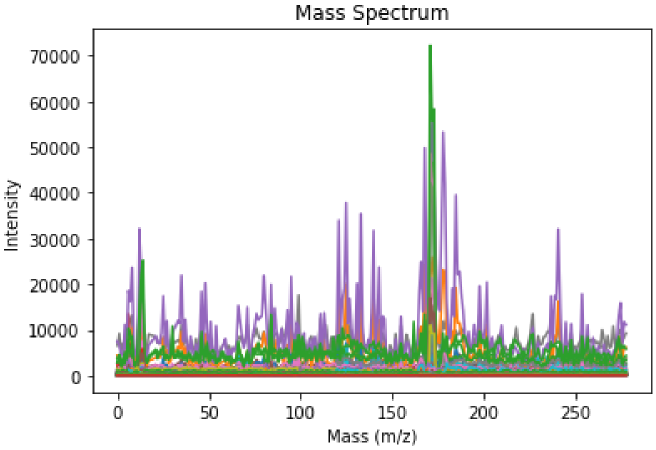


Fig. 5: Mass Spectrums

A mass spectrum measures mass charge versus intensity, where the **charge ratio** or m/z ratio is on the x-axis, where m is the **mass** - the amount of matter in an object, z is the **charge** of the ion. The mass charge ratio m/z is useful, as it allows us to differentiate between molecules of the same mass, but different charges, or the same charge but different masses. The **intensity** on the y-axis refers to the relative abundance of ions in a mass spectrum, and the intensity peak in a mass spectrum represents the number of ions with a particular mass-to-charge ratio that is detected by the mass spectrometer.

Many alternative state-of-the-art chemistry techniques could be considered for the task. The alternative chemistry techniques that could be considered were:

- **Light-based** - One approach is to use analytical techniques based on light e.g. UV or fluorescence spectrophotometry, or vibrational spectroscopy (infrared, near-infrared or Raman spectroscopies). These techniques have been applied in combination with genetic programming to nutrient assessment in horticultural products [53], [54].
- **DNA Sequencing** - This is limited due to extremely low sample size, and very high-dimensional data, e.g. the average human genome contains 3 billion base pairs and 30,000 genes. The dimensionality, and consequently the computation required to process it, rules out genomics data for real-time fish contamination detection. DNA identification methods were examined in a meta-analysis which revealed an average mislabelling rate of 30% in seafood processing [55]. DNA methods are limited, as they only differentiate between species, and are not useful for determining different body parts from the same species, or non-organic matter (e.g. engine oil) [3].
- **Gas-chromatography mass-spectrometry** - Previous work [7] demonstrated that gas-chromatography mass-

Method	Train	Test
MTGP	0.9997 ± 0.0015	0.9472 ± 0.1025
Transformer	**1.0000 ± 0.0000**	**0.9958 ± 0.0131**
Random Forest	1.0000 ± 0.0000	0.9588 ± 0.0447
KNN	0.9324 ± 0.0243	0.8369 ± 0.0691
DT	1.0000 ± 0.0000	0.9913 ± 0.0172
NB	0.9340 ± 0.0699	0.8797 ± 0.0957
LR	1.0000 ± 0.0000	0.9672 ± 0.0475
SVM	1.0000 ± 0.0000	0.9597 ± 0.0506
LDA	0.9867 ± 0.0077	0.9647 ± 0.0367
Ensemble	1.0000 ± 0.0000	0.9816 ± 0.0300

TABLE III: Species

Method	Train	Test
MTGP	0.9793 ± 0.0159	0.5583 ± 0.1897
Transformer	**1.0000 ± 0.0000**	**0.6333 ± 0.2459**
Random Forest	1.0000 ± 0.0000	0.4000 ± 0.1527
KNN	0.4288 ± 0.0537	0.3166 ± 0.1449
DT	1.0000 ± 0.0000	0.2722 ± 0.1325
NB	1.0000 ± 0.0000	0.4500 ± 0.1560
LR	1.0000 ± 0.0000	0.5666 ± 0.1527
SVM	1.0000 ± 0.0000	0.5611 ± 0.1458
LDA	0.7561 ± 0.0320	0.4555 ± 0.1606
Ensemble	1.0000 ± 0.0000	0.5166 ± 0.1572

TABLE IV: Part

spectrometry (GC-MS) can identify fish species with high accuracy. However, GC-MS techniques require significant time and domain expertise to prepare and analyze samples. This is not applicable for real-time fish contamination detection.

V. Results

Table III gives the results for the binary classification task using REIMS on fish species. Each model is run 30 independent times, and the average performance across those runs is reported here. The model with the best performance on the test set is given in **bold**.

Table IV gives the results for the multi-class classification task using REIMS on fish parts. Each model is run 30 independent times, and the average performance across those runs is reported here. The model with the best performance on the test set is given in **bold**.

Figure 6 gives the loss curve for the transformer on the fish species classification task. The transformer achieved 100% accuracy on the training data, and 99.58% accuracy on the test set when averaged over 30 independent runs. The convergence curve is very smooth, for both the training and validation set. This shows the model effectively fits the training data and generalizes well on unseen data.

Figure 7 gives the loss curve for the transformer on the fish part multi-class classification task. The best performance on the validation set is approximately 80 epochs with 100% training accuracy, and 63.33% test accuracy, averaged over 30 independent runs. The transformer model overfits the training data and fails to generalize on the test set. However, out of all the evaluated methods, a transformer model with 66.33% on the test set performs best on unseen data.

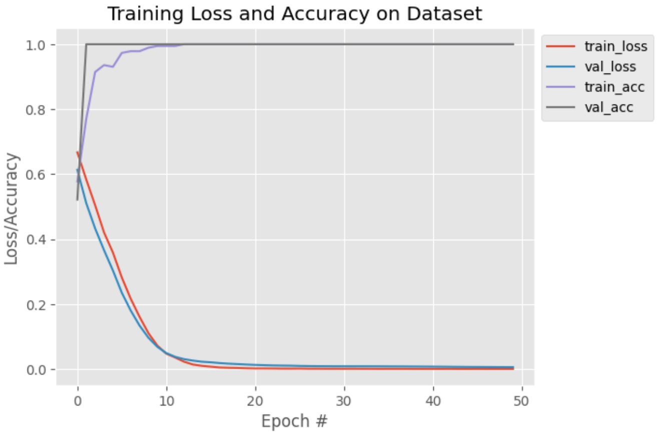


Fig. 6: Fish species: Transformer loss curve

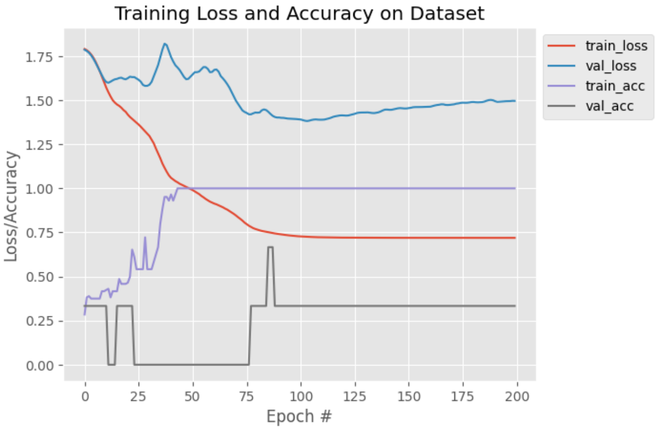


Fig. 7: Fish part: Transformer loss curve

VI. Ablation Studies

Different weight initialization strategies were experimented with on the transformer model. These include Xavier, Kaiming and Orthogonal [34]–[36]. None of these weight initialization methods improved the convergence of the transformer. Compare the smooth loss curve of the transformer in fig. 6, to Xavier in fig. 8, Kaiming in fig. 9, and Orthogonal in fig. 10. Orthogonal and Xavier weight initializations have smoother convergence curves than the Kaiming. However, for all three weight initialization strategies, the model fails to reach convergence in under 50 training epochs. The default weight initialization with pre-training works best.

Figure 11 gives the loss and accuracy curve, for both training and validation, for the pre-layer-normalization (pre-LN) and post-layer-normalization (post-LN) transformer variants, on the fish species dataset. Post-LN [14] has an accuracy of 0.9833 ± 0.0204 on the fish species dataset. Pre-LN [37], [38] has an accuracy of 0.9916 ± 0.0166 on the fish species dataset. Pre-LN has a higher mean accuracy and less variance than post-LN. Furthermore, post-LN fails to converge in 50 training epochs, whereas Pre-LN converges in under 15 epochs. The

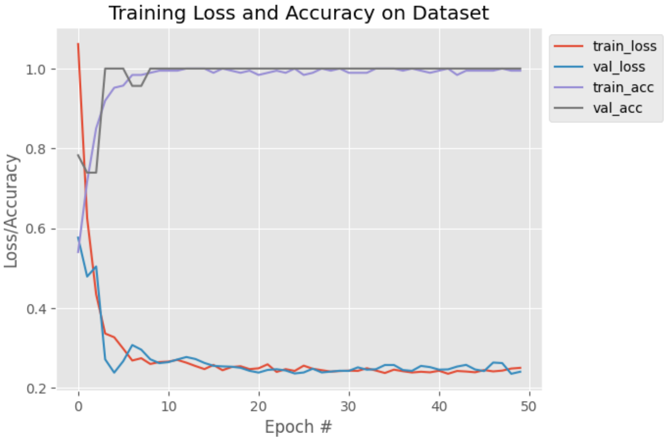


Fig. 8: Xavier weight initialization

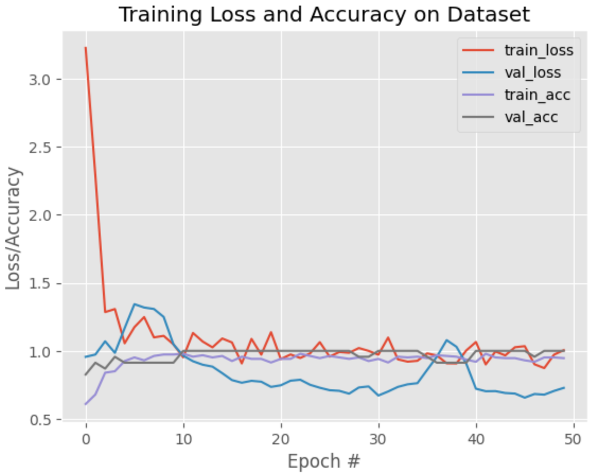


Fig. 9: Kaiming weight initialization

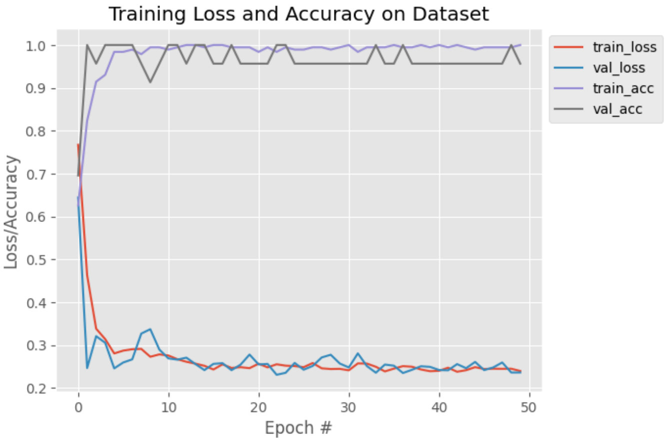
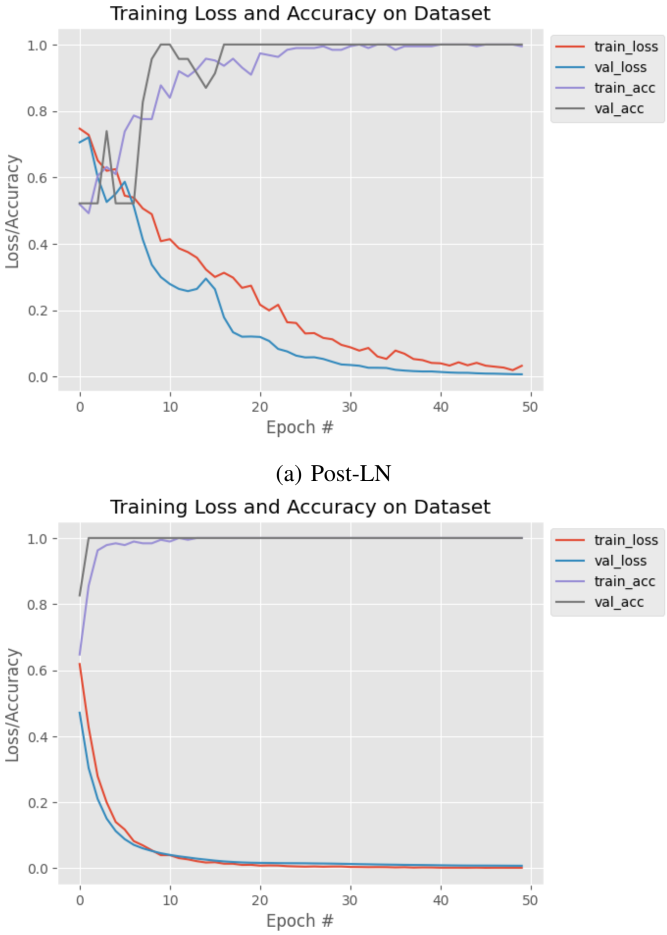


Fig. 10: Orthogonal weight initialization

pre-norm formulation for layer normalization in a transformer offers superior performance on the fish species dataset.

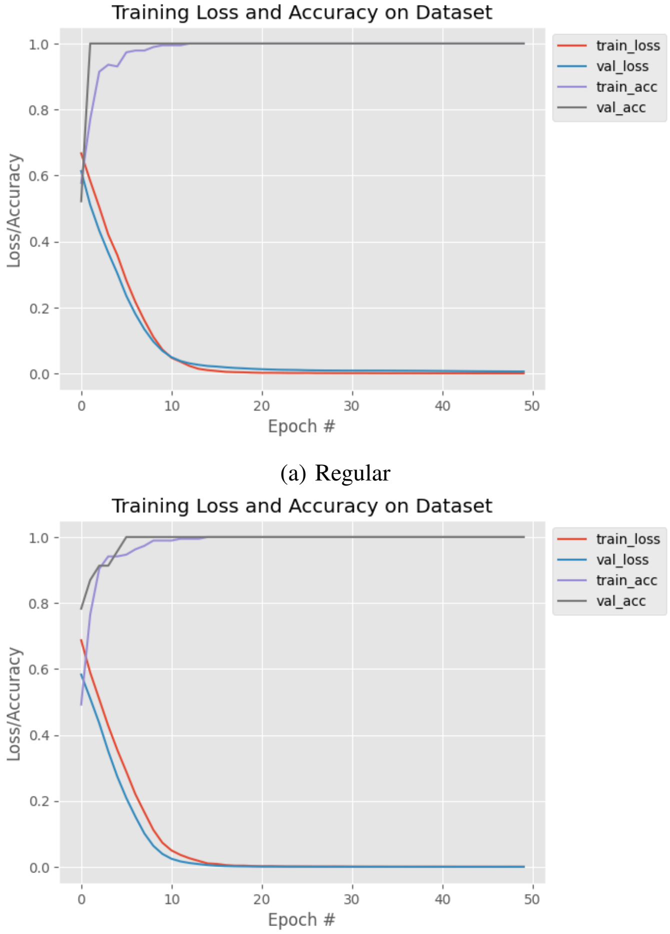


(a) Post-LN

(b) Pre-LN

Fig. 11: Post-LN (left) Pre-LN (right) transformer

Label smoothing [56] is a regularization technique for categorical cross-entropy. It softens the class label targets, by combining one-hot encodings with a uniform distribution. Adding noise to the target labels can be considered a form of regularization [44]. With label smoothing enabled, with an alpha $\alpha = 0.1$, the transformer gets 0.9958 ± 0.0131, compared to 0.9916 ± 0.0166 without label smoothing, on the fish species dataset. Figure 12 gives the loss and accuracy curve, for both training and validation, for the transformer variants, with and without label smoothing, on the fish species dataset.



(a) Regular

(b) Label Smoothing

Fig. 12: Regular (left) label smoothing (right) transformer

VII. Interpretability

A. Decision Tree

By examining the results in the previous section, the decision tree [12], [57] outperforms nearly all other methods (except for the transformer) for fish species identification. It achieves 100% accuracy on the training set, and 99.2% ± 1.64 on the test set, over an average of 30 runs. This remarkable performance deserves an explanation, lucky enough, not only are decision trees accurate, but their underlying representation is also rather straightforward to understand.

Figure 13 shows the interpretability of a decision tree. The internal nodes, refer to the species mass-to-charge ratio, and the intensity threshold needed to make that decision. The mass-to-charge ratio is the x-axis, and the intensity is the y-axis, respectively, on a mass spectrograph. Internal nodes are those nodes which decide how to split the dataset into smaller groups.

Here that mass-to-charge ratio of *110.122786584657* is chosen as the best splitting criterion, whose initial superset had a Gini impurity of 0.496. The decision to split the data into two subsets is made by checking if the intensity on the y-axis, of the feature at *110.122786584657* on the x-axis, is greater than or equal to *19.426*. This split separates 101 Hoki, from 86 other instances, note that those 86 instances only contain one Hoki, and 86 Mackerel, with a Gini impurity of 0.023. The best split has *nearly* classified the entire dataset in one decision.

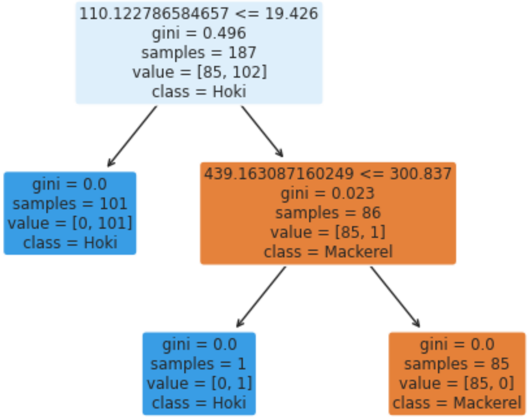


Fig. 13: Fish species: decision tree

The second node, with mass-to-charge ratio *439.163087160249*, separates the remaining instances into two pure subsets, each with a Gini impurity of 0.0. The decision is made by checking if the intensity on the y-axis, of the feature *439.163087160249* on the x-axis, is greater than or equal to *300.837*. The classification tree has separated all the classes into pure subsets (e.g. leaves), so the CART algorithm is finished.

The results would suggest that the molecules at mass-to-charge ratio *110.122786584657* and *439.163087160249* are highly correlated, both with each other and the fish species Mackerel. Given the large threshold, greater than or equal to *300.837*, a significant amount of that molecule is needed to give the Mackerel prediction. And the absence, or trace amounts, of either molecule, likely indicate Hoki.

B. Transformer: Attention maps

The attention map visualizes the attention mechanism between the first ten mass-to-charge ratios. Attention maps illustrate the attention weights that are assigned to each input token and how they attend to one another. Figure 14 gives the attention map for the first self-attention layer of the encoder.

The final self-attention layer (fig. 14) of the decoder is the penultimate layer of the neural network, it is the second to last layer, one layer before the final linear layer that computes the cross entropy loss criterion. Note the penultimate layer is the target layer for the label smoothing technique.

Figure 16 gives the attention map for the final layer of the decoder - with label smoothing. This attention map shows less extreme values in weight parameters for the transformer with label smoothing than without fig. 15. Note that this heatmap has two black squares and two white squares. Many black and

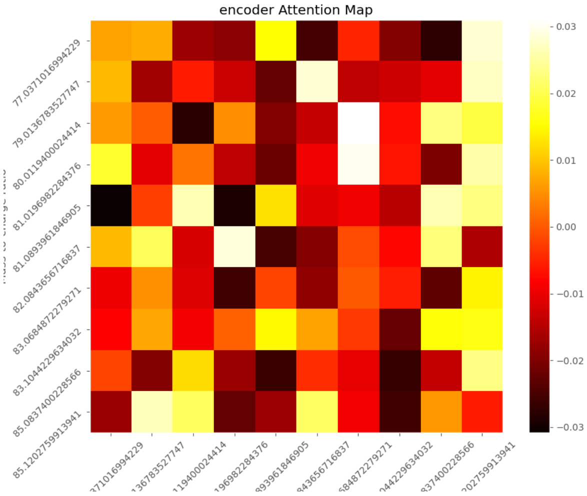


Fig. 14: Encoder: Attention map

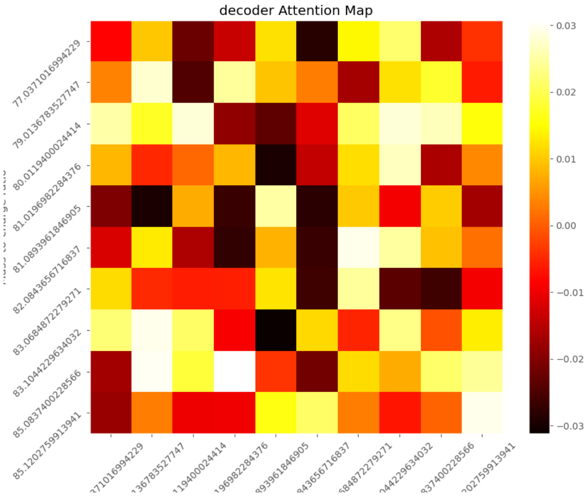


Fig. 15: Decoder: Attention map

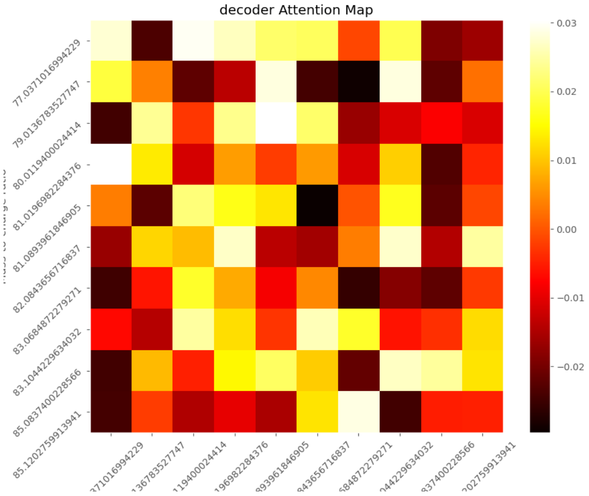


Fig. 16: Decoder: Attention map with label smoothing

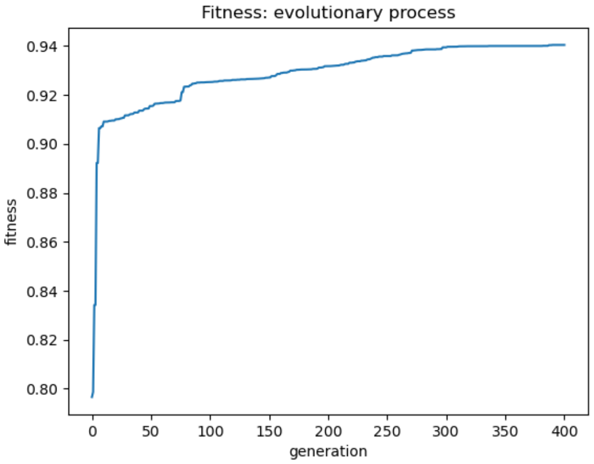


Fig. 17: Evolutionary process: fish species

white squares are on the other in fig. 15. This provides visual proof that the weights contain less extreme values with label smoothing [56].

C. Genetic Programming: Trees

Figure 17 gives the evolutionary process for the fish species classification task. The figure shows a gradual incline without any significant jumps - a steady evolutionary process. Given the high classification accuracy on both training and test data, here we see a smooth evolutionary process. The model fits the training set accurately, and performance generalizes well to unseen data in the test set. These trees achieved a balanced accuracy of 100% on the training set, and 95% on the test set, for fish species classification.

Figure 18 gives the evolutionary process for the fish part classification task. Given the high classification accuracy on the training set, and noticeably lower classification accuracy on the test set. The model overfits the fish parts dataset and doesn't generalize as well on unseen data, as it did for fish species. The evolutionary process shows a series of four large jumps in fitness, not a smooth gradual incline, as it had done for fish species. These trees achieved a balanced accuracy of 97.93% on the training set, and 55.83% on the test set, for fish part classification.

Figure 19 is an example of an output tree for detecting Hoki generated by genetic programming. This model achieved 100% classification accuracy on the training set and 95% classification accuracy on the test set. The tree performs well, it fits the training data and generalizes well on unseen data in the test set.

Figure 20 is an example of an output tree for detecting Mackerel generated by genetic programming. This subtree is taken from the same individual as above. The results on GP are

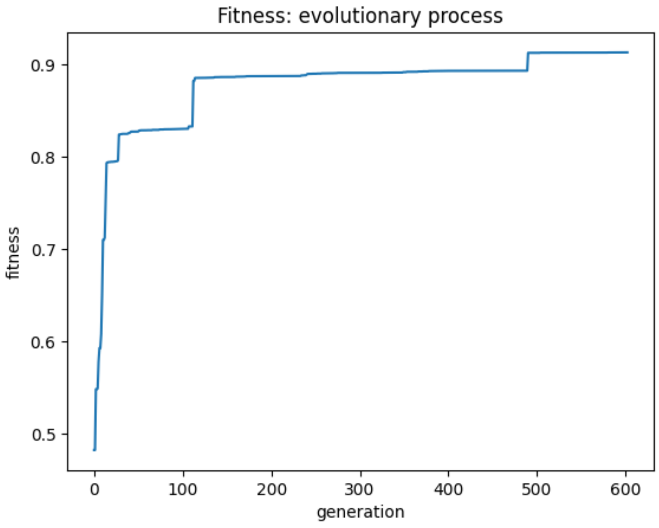


Fig. 18: Evolutionary process: fish part

easy to interpret and can be verified by experts with domain expertise.

VIII. Discussion

In contemporary machine learning research, the Transformer model stands out for its remarkable accuracy in various tasks. This paper shows that the transformer method, although proposed for natural language processing, can be applied to mass spectrometry on marine biomass. However, its inherent complexity often renders it an inscrutable black box, impeding a comprehensive understanding of its decision-making processes. Conversely, genetic programming, while exhibiting lesser accuracy than the Transformer, offers a distinct advantage by generating interpretable results, thus enabling clearer insights into its functioning. In specific applications, such as fish species classification, Decision Trees emerge as a practical choice due to their effectiveness in delineating classification boundaries and facilitating accurate categorization of fish species.

IX. Future Work

Future work will involve three areas, those include: (1) exploring further techniques for AI explainability, (2) verifying the output of the explainable models with domain knowledge, (3) extending the existing classifiers to oil and cross-species contamination detection. For real-world use of this research in the application domain of fish processing, the models must be accurate, explainable, and able to be troubleshooter by domain experts in fish processing, biology and chemistry.

Glossary

adulteration

Food adulteration is the act of intentionally debasing the quality of food offered for sale either by the admixture or substitution of inferior substances or by the removal of some valuable ingredient [5] . 1, 2

AI

artificial intelligence. 3

AMS

ambient mass spectrometry. 6

analysis

Analysis is concerned with identifying which contaminants are present. Not to be confused with **detection**, which simply tells us if a sample is contaminated. Analysis takes this one step further and gives predictions for which contaminants are present in the fish tissue. Take for example, **cross-species contamination**, contaminant analysis predicts which species are present in a contaminated sample, e.g. detection: contaminated, analysis: Hoki and Mackerel both present . 1, 2

anomalies

Anomalies refer to out-of-distribution data that the model could not possibly expect. It is unrealistic for the model to correctly classify these instances, but a model can be built to detect such anomolies, as seen in [18]. In fish processing, an example of an anomaly would be a new species of fish, or marine biomass, that is not a labelled class or present in the training or validation data . 3

CART

classification and regression tree. 2

charge

characteristic of a unit of matter that expresses the extent to which it has more or fewer electrons than protons. Electric charge is the physical property of matter that causes it to experience a force when placed in an electromagnetic field. In the context of mass spectrometry, particularly REIMS which uses a Time-of-Flight (TOF), this uses an electric field to accelerate generated ions through the same electrical potential and then measures the time each ion takes to reach the detector. Depending on the charge of each particle, that time will vary, because the electric field applies different amounts of force to particles with different charges . 6

CNN

convolutional neural networks. 3

conceptual drift

A term from data stream mining, [23], [58], that refers to a change in the underlying distribution of the data. In fish processing, conceptual drift occurs in **seasonal variation** where the composition of fish changes between different seasons . 3

contamination

Food contamination is generally defined as foods that are spoiled or tainted because they either contain microorganisms, such as bacteria or parasites, or toxic substances that make them unfit for consump-

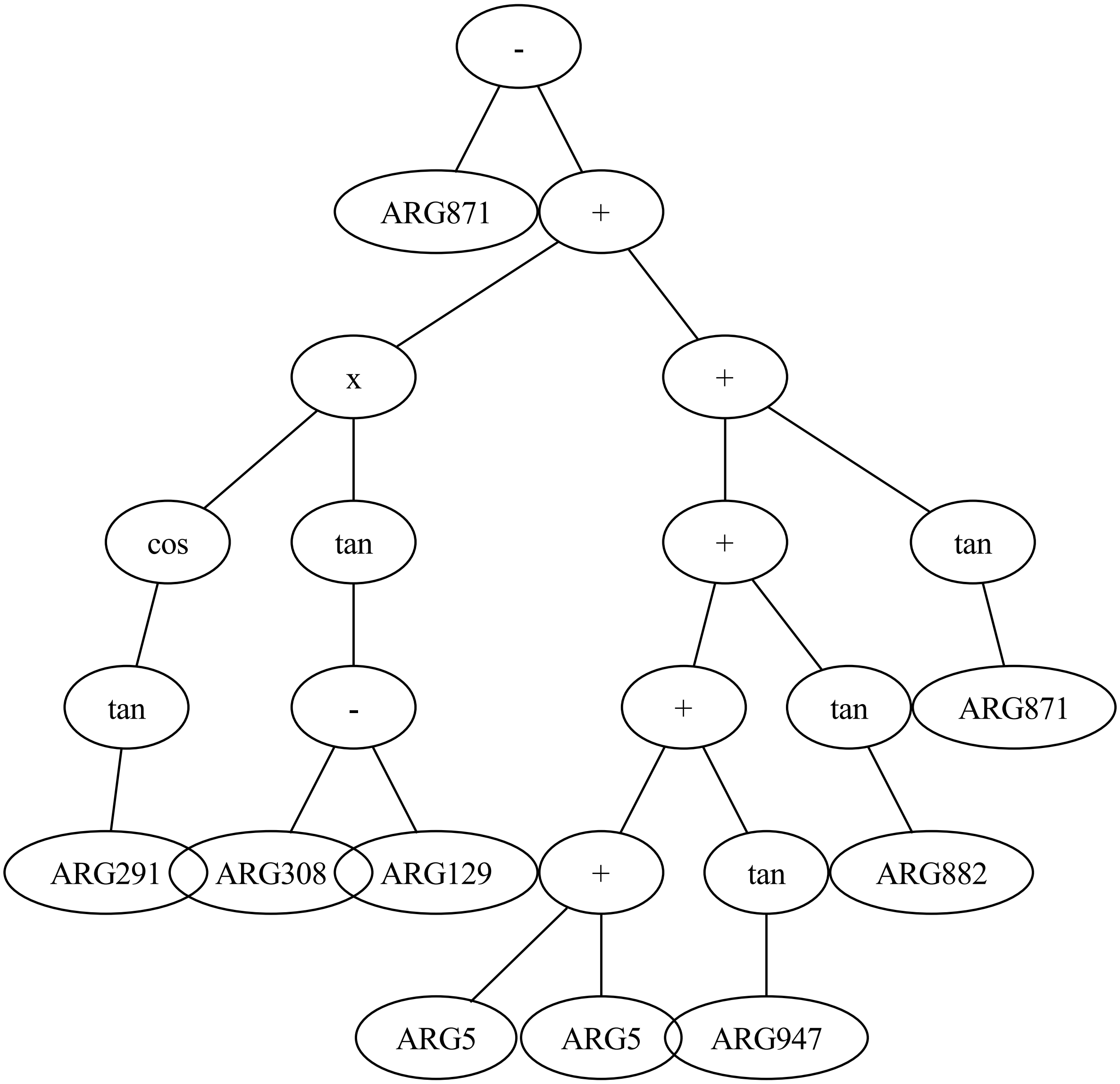


Fig. 19: Fish species: Hoki

tion. A food contaminant can be biological, chemical or physical, with the former being more common. These contaminants have several routes throughout the supply chain (farm to fork) to enter and make a food product unfit for consumption [59] . 1, 2, 6, 7

cross-species

Cross-species refers to a form of contamination, where two species are mixed together, e.g. a sample with both Hoki and Mackerel. In the mass spectrometry datasets, these species are mixed thoroughly in a blender to give a homogeneous sample with a maximum blend of the two species . 1, 2

domain knowledge

Knowledge related to the application domain. For example, biochemistry and fish processing . 2

DT

decision tree. 3

EC

evolutionary computation. 2

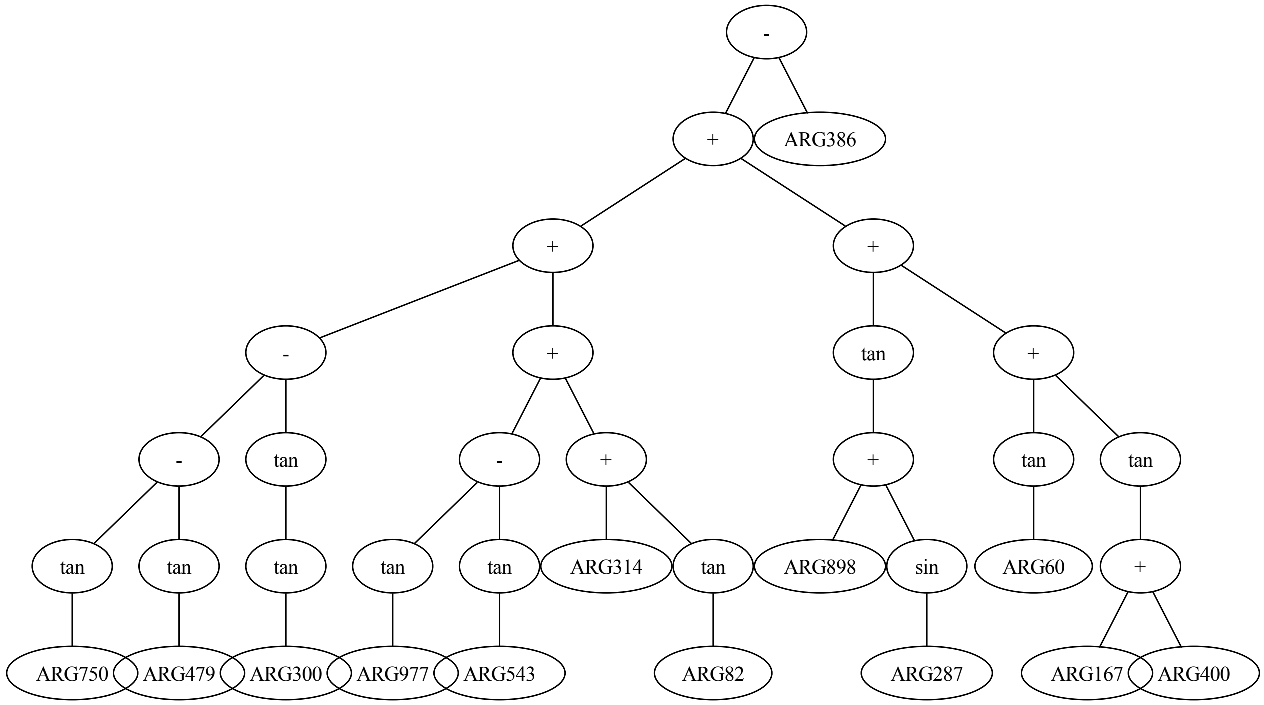


Fig. 20: Fish species: Mackerel

EPA

environmental protection agency. 2

GAN

generative adversarial networks. 3

GC-MS

gas-chromatography mass-spectrometry. 6, 7

GP

genetic programming. 2, 5, 6, 10, 11

heterogeneous

The antonym of **homogeneous**. Consisting of many different elements. In the context of fish processing, New Zealand's marine biomass, the incoming catch from trawling vessels, is heterogeneous, as it consists of many different species - a wide range of marine biomass . 2

homogeneous

This term is used heavily in chemistry. In the context of chemistry homogeneous means the same, or having a similar structure. In fish processing, the fish tissue samples are taken from a homogeneous blend of marine biomass. Also, in the Hoki season, the input to the flex-factory is predominantly one species, this may also be referred to as homogeneous. The marine biomass of Canada or the United States, the incoming catch from trawling vessels, is homogeneous, as it consists of mostly one (or few) species - a narrow range of marine biomass . 2

hyperparameter

Hyperparameter (machine learning) In machine learning, a hyperparameter is a parameter whose value is used to control the learning process. These are often manually set by the user, and are comparable to nuisance parameters from statistics, as they require tuning for models to perform well . 2

identification

Different to detection, identification involves detecting the presence of phenomena in a sample and then specifying what the phenomena were. E.g., an identification system can find **cross-species** contamination and identify both species in the contamination . 1, 2, 6

intensity

The intensity on the y-axis refers to the relative abundance of ions in a mass spectrum, the intensity peak in a mass spectrum represents the number of ions with a particular mass-to-charge ratio that are detected by the mass spectrometer . 6

KNN

K-nearest neighbours. 3

LDA

linear discriminant analysis. 3

LR

logistic regression. 3

marine biomass

A fancy term for fish. To get super technical, marine biomass is a super-set, which includes fish, whales, plankton, crustaceans, marine animals and plants. A fish processing plant will deal with marine biomass from many forms of organic matter. So marine biomass is a catch-all term to refer to the incoming biological materials that enter the factory . 1

mass

The amount of matter in an object . 6

mass charge ratio

The mass charge ratio m/z is useful, as it allows us to differentiate between molecules of the same mass, but different charges, or the same charge but different masses . 6

mass spectrum

The mass spectrum is the artefact of the mass spectrometry technique. A mass spectrum measures mass charge versus intensity, where the **charge ratio** or m/z ratio is on the x-axis, where m is the **mass** - the amount of matter in an object, z is the **charge** of the ion. The mass charge ratio m/z is useful, as it allows us to differentiate between molecules of the same mass, but different charges, or the same charge but different masses. The **intensity** on the y-axis refers to the relative abundance of ions in a mass spectrum, and the intensity peak in a **mass spectrum** represents the number of ions with a particular mass-to-charge ratio that is detected by the mass spectrometer . 6

MCIFC

multiple class-independent feature construction method. 2, 5

MS

mass-spectrometry. 3

MSM

masked spectra modelling. 3, 5

MT-GP

multi-tree genetic programming. 2, 3, 5

NB

naive bayes. 3

NSP

next spectra prediction. 3, 5

online learning

Online learning refers to a model that can be updated and adapt to new instances after its initial training. Take for example the Tesla FSD training programme. The FSD edge cases are referred to as the long tail of computer vision. These edge cases are where the car demonstrates undesirable behaviour, e.g. a crash, swerve, unsafe/irregular driving, are sent back to the DOJO computing facility, and the model is retrained via Monte-Carlo simulation of that edge case, to perform the desired behaviour. This human-in-the-loop online learning is a powerful method to bootstrap algorithms for robustness. Not to be confused with **online** Online learning is the opposite of **offline learning** . 2, 3

PCA-LDA

principal component analysis - linear discriminant analysis. 2

post-LN

post-layer-normalization. 7

pre-LN

pre-layer-normalization. 7

PSO

particle swarm optimisation. 2

REIMS

rapid evaporative ionisation mass spectrometry. 1, 2, 6

RF

random forest. 3

RSD

relative standard deviation. 2

seasonal variation

The composition of **marine biomass** varies by season, a reoccurring **conceptual drift**. The temperature of the ocean, diets of fish, changes from Winter to Summer, oceans heat up, migration/spawning. For example, while spawning, Hoki changes composition, extracting their lipids, and putting them all into their eggs, after spawning adult Hoki is a mess [25] . 3

SOTA

state-of-the-art. 2

SVM

Support Vector Machine. 3

taxonomy

A taxonomy is a hierarchical classification system that organizes a set of concepts or subjects into categories and subcategories based on shared characteristics. Taxonomies are often used in fields such as biology, where they are used to classify and organize living organisms into a systematic hierarchy based on their characteristics and evolutionary relationships. They are also used in other fields, such as information science and library science, to classify and organize knowledge in a way that is easy to understand and navigate . 2, 3

tissue

See part . 6

transfer learning

Transfer learning is a machine learning technique where shared knowledge is transferred between related tasks. Take for example, the source task of riding a bike, and the target task of riding a motorcycle. Although the tasks are different, there is shared knowledge from the source task, that will be useful when performing the target task. In layman's terms, if you already can ride a bike, it will be easier to ride a motorcycle . 2, 3

References

[1] FAO, *The State of World Fisheries and Aquaculture, 2020*. FAO, 2020.

[2] Plant and F. Research, "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.

[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] 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.

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

[6] J. Kaminski, "Diffusion of innovation theory," *Canadian Journal of Nursing Informatics*, vol. 6, no. 2, pp. 1–6, 2011.

[7] J. Wood, B. H. Nguyen, B. Xue, M. Zhang, and D. Killeen, "Automated fish classification using unprocessed fatty acid chromatographic data: A machine learning approach," in *Australasian Joint Conference on Artificial Intelligence*, pp. 516–529, Springer, 2022.

[8] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95-international conference on neural networks*, vol. 4, pp. 1942–1948, IEEE, 1995.

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

[10] B. Tran, B. Xue, and M. Zhang, "Genetic programming for feature construction and selection in classification on high-dimensional data," *Memetic Computing*, vol. 8, no. 1, pp. 3–15, 2016.

[11] B. Tran, B. Xue, and M. Zhang, "Genetic programming for multiple-feature construction on high-dimensional classification," *Pattern Recognition*, vol. 93, pp. 404–417, 2019.

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

[13] 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.

[14] 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.

[15] 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.

[16] K. Bi, D. Zhang, T. Qiu, and Y. Huang, "Gc-ms fingerprints profiling using machine learning models for food flavor prediction," *Processes*, vol. 8, no. 1, p. 23, 2019.

[17] D. D. Matyushin and A. K. Buryak, "Gas chromatographic retention index prediction using multimodal machine learning," *Ieee Access*, vol. 8, pp. 223140–223155, 2020.

[18] F. Di Mattia, P. Galeone, M. De Simoni, and E. Ghelfi, "A survey on gans for anomaly detection," *arXiv preprint arXiv:1906.11632*, 2019.

[19] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[20] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[21] V. B. O'Donnell, E. A. Dennis, M. J. Wakelam, and S. Subramaniam, "Lipid maps: Serving the next generation of lipid researchers with tools, resources, data, and training," *Science signaling*, vol. 12, no. 563, p. eaaw2964, 2019.

[22] C. A. Smith, G. O'Maille, E. J. Want, C. Qin, S. A. Trauger, T. R. Brandon, D. E. Custodio, R. Abagyan, and G. Siuzdak, "Metlin: a metabolite mass spectral database," *Therapeutic drug monitoring*, vol. 27, no. 6, pp. 747–751, 2005.

[23] Y. Sun, B. Pfahringer, H. M. Gomes, and A. Bifet, "Soknl: A novel way of integrating k-nearest neighbours with adaptive random forest regression for data streams," *Data Mining and Knowledge Discovery*, vol. 36, no. 5, pp. 2006–2032, 2022.

[24] A. Bifet and R. Gavalda, "Learning from time-changing data with adaptive windowing," in *Proceedings of the 2007 SIAM international conference on data mining*, pp. 443–448, SIAM, 2007.

[25] "Hoki macruronus novazelandiae." https://openseas.org.nz/fish/hoki/, Oct 2021.

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

[27] E. Fix and J. L. Hodges, "Discriminatory analysis. nonparametric discrimination: Consistency properties," *International Statistical Review/Revue Internationale de Statistique*, vol. 57, no. 3, pp. 238–247, 1989.

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

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

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

[31] S. Balakrishnama and A. Ganapathiraju, "Linear discriminant analysis-a brief tutorial," *Institute for Signal and information Processing*, vol. 18, no. 1998, pp. 1–8, 1998.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[34] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, JMLR Workshop and Conference Proceedings, 2010.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

[36] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," *arXiv preprint arXiv:1312.6120*, 2013.

[37] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T. Liu, "On layer normalization in the transformer architecture," in *International Conference on Machine Learning*, pp. 10524–10533, PMLR, 2020.

[38] A. Karpathy, "Let's build gpt: from scratch, in code, spelled out." https://youtu.be/kCc8FmEb1nY?si=1vM4DhyqsGKUSAdV, 2023.

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

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

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

[42] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[43] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[44] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

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

[46] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, "Improving language understanding by generative pre-training," *Arxiv*, 2018.

[47] P. C. English, J. Kelleher, and J. Carson-Berndsen, "Domain-informed probing of wav2vec 2.0 embeddings for phonetic features," in *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pp. 83–91, 2022.

[48] K. Fukushima, "Visual feature extraction by a multilayered network of analog threshold elements," *IEEE Transactions on Systems Science and Cybernetics*, vol. 5, no. 4, pp. 322–333, 1969.

[49] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

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

[51] Y. Tang and C. Eliasmith, "Deep networks for robust visual recognition," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 1055–1062, 2010.

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

[53] D. Robinson, Q. Chen, B. Xue, D. Killeen, S. Fraser-Miller, K. C. Gordon, I. Oey, and M. Zhang, "Genetic algorithm for feature and latent variable selection for nutrient assessment in horticultural products," in *2021 IEEE Congress on Evolutionary Computation (CEC)*, pp. 272–279, IEEE, 2021.

[54] D. Robinson, Q. Chen, B. Xue, D. Killeen, K. C. Gordon, and M. Zhang, "A new genetic algorithm for automated spectral pre-processing in nutrient assessment," in *International Conference on the Applications of Evolutionary Computation (Part of EvoStar)*, pp. 283–298, Springer, Cham, 2022.

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

[56] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[57] W.-Y. Loh, "Classification and regression trees," *Wiley interdisciplinary reviews: data mining and knowledge discovery*, vol. 1, no. 1, pp. 14–23, 2011.

[58] H. M. Gomes, J. Montiel, S. M. Mastelini, B. Pfahringer, and A. Bifet, "On ensemble techniques for data stream regression," in *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2020.

[59] M. A. Hussain, "Food contamination: major challenges of the future," 2016.