Transformers: roBERTs in disguise

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

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

Abstract. This research applies advanced AI techniques to rapid evaporative ionisation mass spectrometry datasets in marine biomass analysis. It introduces transformers with pre-training strategies such as next spectra prediction and masked spectra modelling, alongside decision trees and genetic programming for interpretable insights into spectral patterns and chemical attributes. The paper proposes a comprehensive toolbox of machine-learning methods tailored for precise fish speciation and body part identification, through binary and multi-class classification tasks.

Keywords: classification · deep learning · explainable AI · genetic programming · machine learning · mass spectrometry · transformers

1 Introduction

Efforts to maximise waste utilisation in the global fishing industry are critical, given that only 40% of the approximately 100 million tonnes of wild fish caught annually are processed into edible parts [8]. The rest is often used for fish oil, fish meal, or discarded. Techniques that can identify high-value fish parts help drive decisions on repurposing fish waste into valuable products, such as Omega 3 supplements. Accurate fish speciation methods are essential to combat the significant mislabelling issues in seafood processing, as highlighted by a meta-analysis showing a 30% average mislabelling rate in the global seafood industry [23]. These methods ensure transparency in the seafood supply chain, mitigate fraud and support sustainable fishing practices by enabling proper resource management and conservation efforts.

This research aims to demonstrate the practicality of rapid evaporative ionisation mass spectrometry (REIMS) in fish processing applications, including its effectiveness in fish speciation [4] and detecting body parts [3, 17]. Previous studies on adulteration [17] have shown REIMS capable of detecting body parts in food samples, such as identifying horse meat adulteration in beef with as little as 1% cross-species material. This underscores that REIMS can be used for combating fraud, mislabelling, and quality assurance in fish processing.

Additionally, New Zealand's smaller fishing fleet and population result in limited sample sizes for marine biomass analysis, necessitating innovative approaches like transfer learning and data augmentation, to amortise the limited number of training instances.

2 Related Works

Breiman et al. (1984) introduced CART (Classification and Regression Trees), which has been effective in analysing mass spectrometry datasets and was notably applied in EPA-funded projects in the late 1970s and early 1980s to detect toxic substances in water and air samples by correlating fragment ions of molecules [5]. Now this study employs decision trees for classifying fish speciation and fish body parts, leveraging CART's greedy algorithm to iteratively split data based on criteria that minimise Gini impurity, to enhance subset purity until a stopping criterion is met, or further splits yield no improvement.

Black et al. [4] used REIMS with PCA-LDA to classify seabass and seabream fish speciation and catch methods in a binary classification task, applied here to classify Hoki and Mackerel speciation. Additionally, this paper expands into the multi-class classification of fish parts, distinguishing among six categories. In a related study, [3] applied REIMS to detect beef adulteration, highlighting its sensitivity in identifying trace amounts of horse offal cross-species contamination.

In prior research [32], the evolutionary computation (EC) approach using particle swarm optimisation (PSO) was employed for feature selection in fish speciation and fish part identification with gas chromatography. Now this paper applies another EC technique, multi-tree genetic programming (MT-GP), for both feature construction and classification, for fish speciation and fish parts classification. This work introduces the novel application of MCIFC [30] in marine biomass analysis using REIMS.

3 Limitations

Existing methods [3,4] often rely on manual thresholds and hyperparameter tuning by domain experts in chemistry. This paper proposes automated approaches or models that require minimal tuning. Additionally, while previous studies primarily use basic statistical models like LDA and OPLS-DA [3, 4], this research introduces advanced techniques such as transformers [7, 31], evolutionary computation [29, 30, 32], and ensembles [12]. It also emphasises the introduction of machine learning techniques to the field traditionally dominated by chemistry and statistics [2–4, 17], providing the machine learning terminology and jargon to facilitate interdisciplinary collaboration.

4 Theory

4.1 Classifiers

Seven standard classifiers; random forest (RF) [16], k-nearest neighbour (KNN) [9], decision trees (DT) [5], naive bayes (NB) [11], logistic regression (LR) [19], support vector machines (SVM) [6], linear discriminant analysis (LDA) [1]; are selected for the tasks of fish speciation and fish parts identification, along with an ensemble voting classifier [12], that combines them all [1, 5, 6, 9, 11, 16, 19]. They use default settings from sklearn [24], except SVM with a linear kernel and LR set to 2,000 max iterations. The ensemble voting classifier uses hard voting. More advanced classification methods are explored below.

4.2 Transformer

In 2017, Vaswani et al. [31] introduced the transformer architecture, originally for machine translation, which this paper adapts for marine biomass analysis (see fig. 1). The transformer's design includes stacked encoder-decoder layers with residual connections [14], and advancements like various weight initialisation methods [10, 13, 25] and pre-norm layer normalization [33], where layer normalisation is performed before the attention and feedforward layers (see fig. 2).

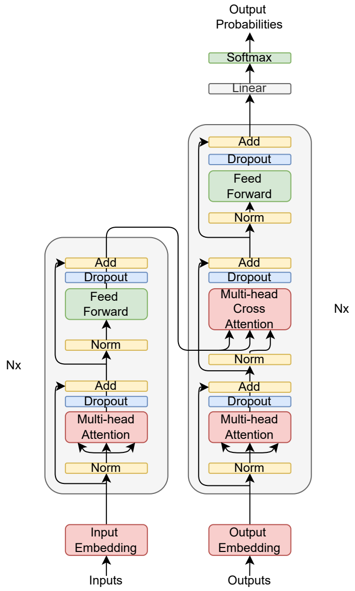
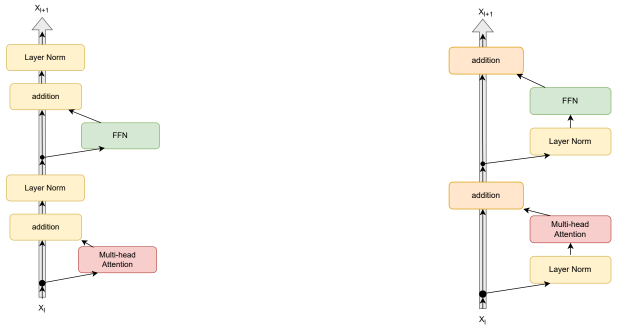


Fig. 1: Transformer architecture

This paper introduces two novel unsupervised pre-training methods for mass spectrometry, inspired by BERT [7]. The first method, masked spectra modelling (MSM), adapts masked language modelling by randomly masking mass spectra



(a) Post-norm

(b) Pre-norm

Fig. 2: Post-norm (left) Pre-norm (right) formulation

and predicting them using mean square error (MSE) with early stopping. The second method, next spectra prediction (NSP), adapts next sentence prediction by predicting if two halves of spectra belong to the same or different spectra, using categorical cross-entropy and early stopping.

The transformer model in this paper employed the AdamW optimiser [21] to address issues with the Adam optimizer [18] by decoupling weight decay from the learning rate. Dropout [27] approximates a bagged ensemble of neural networks efficiently. For regularization, label smoothing [28] softens class label targets using a combination of one-hot encodings and a uniform distribution. The transformer network utilizes the GELU activation function [15]. Data augmentation inflates the number of training instances by duplicating each instance five times and injecting noise [26], effectively expanding the training set five-fold. Early stopping [22] saves model parameters whenever validation loss improves, effectively tuning hyperparameters by controlling the number of epochs.

Table 1 outlines the transformer configuration used.

Table 1: Transformer parameter settings

Learning rate	1E-5
Epochs	100
Dropout	0.2
Label smoothing	0.1
Early stopping patience	5
Optimiser	AdamW
Loss: MSM	Mean Squared Error
Loss: NSP & Speciation	Categorical Cross Entropy
Input dimensions	1023
Hidden dimensions	128
Output dimensions: MSM	1023
Output dimensions: NSP & Speciation	2
Output dimensions: Part	6
Number of layers (Nx)	3
Number of heads	3

4.3 Mutli-tree Genetic Programming

Multi-tree genetic programming (MT-GP) is used for binary classification for fish speciation and multi-class classification for fish parts identification, offering interpretable outputs compared to transformer models and other deep neural networks. This paper utilises multiple class-independent feature construction (MCIFC) from [30] for novel applications in marine biomass analysis.

In MCIFC, candidate solutions are represented as multiple trees, containing a subtree corresponding to each class. This approach serves both feature construction and classification, where the class prediction is the index of the largest output of subtree, otherwise known as a winner-takes-all strategy. For fish speciation (Hoki and Mackerel), two trees are used, while for fish parts (Fillet, Heads, Livers, Skins, and Guts), six subtrees are employed, thereby reducing the feature space from 1023 to 2 or 6 dimensions, respectively [30]. The genetic operators in MCIFC include crossover and mutation, which are adapted from conventional genetic programming (GP). Crossover selectively operates between trees of the same class with an 80% probability, using one-point crossover, while mutation randomly alters one subtree with a 20% probability. The algorithm employs the VarAnd strategy, allowing each generation to perform crossover, mutation, or both, using tournament selection with a tournament size of 7.

The fitness evaluation combines accuracy with a distance regularisation term. Accuracy is calculated as the balanced accuracy score of the constructed features. The distance regularisation term penalises intraclass distances and rewards interclass distances. The Euclidean distance between pairs of points i and j is given by:

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

The `inter`class distance measures the distance between pairs of instances from *different classes*, aiming for greater distances to enhance fitness:

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

Conversely, `intra`class distance measures the distance between pairs of instances from the *same class*, aiming for smaller distances to improve fitness:

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

where $d(S_i, S_j)$ enumerates over all interclass and intraclass pairs respectively and $|S|$ is the total number of pairs to yield the regularisation term as the average distance between pairs.

The fitness weight α is fixed at 0.8 to prioritise accuracy in fitness values. β controls the balance between inter and intra-class distance, balanced evenly at 0.5. The fitness function is given as:

$$\alpha * \texttt{balanced\_accuracy} + (1-\alpha)(\beta * \texttt{inter} + (1-\beta) * (1 - \texttt{intra}))$$

Table 2 outlines the parameter settings for the GP-based MCIFC method in the experiments. The construction ratio is the number of trees per class.

Table 2: MCFIC 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
Distance weighting β	0.5

5 Dataset

Rapid Evaporative Ionisation Mass Spectrometry (REIMS) is an advanced form of ambient mass spectrometry (AMS) developed for medical research. Mass spectrometry measures mass-to-charge ratios (x-axis) and their relative abundance (y-axis). It utilises tools like electro-surgical knives, bipolar forceps, or lasers to produce an aerosol (smoke) from tissue samples. This aerosol is then directed into a mass spectrometer's ionisation source, where ions are formed on a heated collision surface [17].

The data is normalised during pre-processing between $x \in [0, 1]$. The dataset has a stratified split into training (80%), validation (10%), and test (10%) sets. Fish speciation has two classes with a distribution of 44.4% Hoki and 55.56% Mackerel. Fish part identification has six classes with a distribution of 20% fillet, 20% heads, 10% livers, 20% skins, 20% guts, and 10% frames.

The REIMS dataset introduces the challenges of high dimensionality [20], a limited number of instances in the dataset, and domain expertise is usually required for actionable insights. The machine learning methods proposed in this work are suited towards high-dimensional problems and don't require domain expertise. Data augmentation via noise injection, and, transfer learning via unsupervised pre-training, help amortise the limited instances in the dataset.

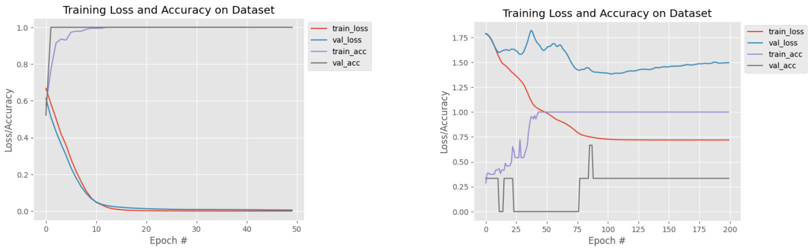
6 Results

Table 3 gives the results of the classifiers, averaged over 30 independent runs, with the best-performing model on the test set given in **bold**.

Table 3: Fish speciation and fish part classification results

	Fish Speciation		Fish Part	
Method	Train	Test	Train	Test
MTGP	0.9997 ± 0.0015	0.9472 ± 0.1025	0.9793 ± 0.0159	0.5583 ± 0.1897
Transformer	**1.0000 ± 0.0000**	**0.9958 ± 0.0131**	**1.0000 ± 0.0000**	**0.6333 ± 0.2459**
Random Forest	1.0000 ± 0.0000	0.9588 ± 0.0447	1.0000 ± 0.0000	0.4000 ± 0.1527
KNN	0.9324 ± 0.0243	0.8369 ± 0.0691	0.4288 ± 0.0537	0.3166 ± 0.1449
DT	1.0000 ± 0.0000	0.9913 ± 0.0172	1.0000 ± 0.0000	0.2722 ± 0.1325
NB	0.9340 ± 0.0699	0.8797 ± 0.0957	1.0000 ± 0.0000	0.4500 ± 0.1560
LR	1.0000 ± 0.0000	0.9672 ± 0.0475	1.0000 ± 0.0000	0.5666 ± 0.1527
SVM	1.0000 ± 0.0000	0.9597 ± 0.0506	1.0000 ± 0.0000	0.5611 ± 0.1458
LDA	0.9867 ± 0.0077	0.9647 ± 0.0367	0.7561 ± 0.0320	0.4555 ± 0.1606
Ensemble	1.0000 ± 0.0000	0.9816 ± 0.0300	1.0000 ± 0.0000	0.5166 ± 0.1572

Figure 3 depicts the transformer's superior performance for fish speciation and fish parts classification. The transformer's attention mechanism models and captures the spatial connectivity of mass spectrometry data by attending to feature interactions between mass-to-charge ratios. Decision trees offer second best near-perfect performance for fish speciation, able to capture the correlation between important features and their class boundaries. The ensemble model performs third best for fish species, showing an ensemble of diverse and independent models can be aggregated to increase the generalisation ability on unseen data. KNN is poorly suited to noisy and/or high-dimensional mass spectrometry datasets.

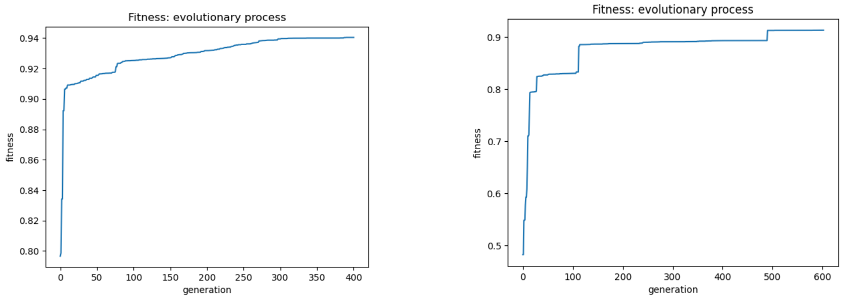


(a) Fish speciation

(b) Fish part

Fig. 3: Fish speciation (left) Fish part (right) loss curve

Figure 4 displays the evolutionary process, for fish speciation and fish part classification, respectively. The fitness function on both graphs reaches a plateau where the regularization term of intra and inter-class distance continually improves with diminishing returns for balanced accuracy.



(a) Fish speciation

(b) Fish part

Fig. 4: Fish speciation (left) Fish part (right) evolutionary process

7 Ablation Studies

The default weight initialization with pre-training proved to be the most effective. Different weight initialization strategies (Xavier, Kaiming, and Orthogonal) were tested on the transformer model, but none improved convergence. Ablation studies on the fish speciation dataset compared pre-LN and post-LN transformer variants. Pre-LN offered superior performance than post-LN. Pre-LN achieves

99.16% ± 1.66% accuracy and converges in 15 epochs, outperforming post-LN which achieves 98.33% ± 2.04% accuracy and fails to converge in under 50 epochs. With label smoothing, the transformer achieves 99.58% ± 1.31% test accuracy, compared to 99.16% ± 1.66% without it on the fish speciation dataset.

8 Interpretability

8.1 Decision Tree

Figure 5 gives the decision tree, which splits data when key mass-to-charge ratios exceed a threshold. For when 110.1228 m/z exceeds intensity 19.426, and later when 439.1631 m/z exceeds intensity 300.837. The intensity threshold for 439.1631 m/z is much greater than 110.1228 m/z, suggesting a large abundance of that molecule in Mackerel. These are important molecules and their thresholds for fish speciation - indicating these molecules and their relative abundance are highly correlated with fish speciation between Hoki and Mackerel.

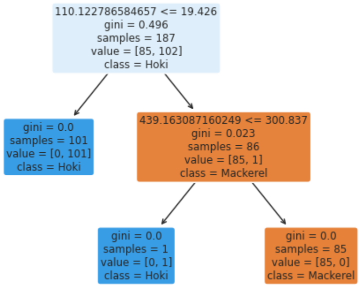


Fig. 5: Fish speciation: decision tree

8.2 Genetic Programming: Trees

For Mackerel a decision tree with seven features in its terminal set achieves perfect training accuracy and 95% test accuracy. Notably feature 5, 81.0893 m/z, is included twice, suggesting this is an important feature highly correlated with Mackerel species prediction. More features are needed for Hoki than Mackerel, with thirteen features in the terminal set, suggesting more features are highly correlated with Hoki prediction.

9 Discussion

Transformers, although initially proposed for natural language processing, can additionally be applied to ambient mass spectrometry for marine biomass analysis. However, despite its accuracy, the complexity renders these models as black

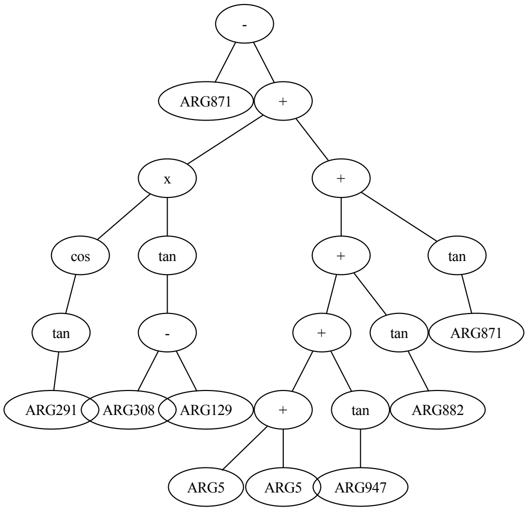


Fig. 6: Fish species: Mackerel

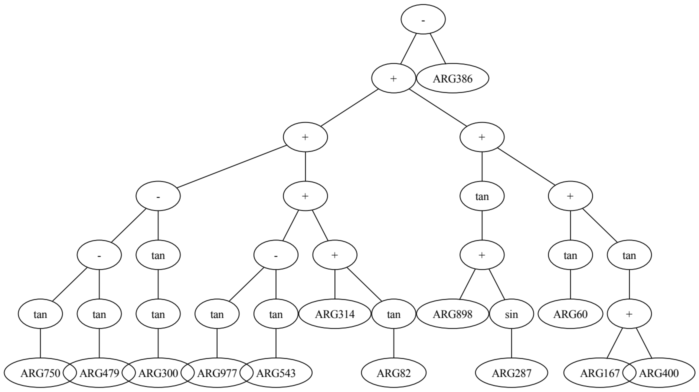


Fig. 7: Fish species: Hoki

boxes, hindering a comprehensive understanding of its decisions. In contrast, genetic programming and decision trees are less accurate but provide interpretable results, offering clearer insights into their functioning. Decision trees can detect fish species in marine biomass samples by correlating fragment ions of molecules.

References

1. Balakrishnama, S., Ganapathiraju, A.: Linear discriminant analysis-a brief tutorial. Institute for Signal and information Processing **18**(1998), 1–8 (1998)
2. 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)
3. Black, C., Chevallier, O.P., Cooper, K.M., Haughey, S.A., Balog, J., Takats, Z., Elliott, C.T., Cavin, C.: Rapid detection and specific identification of offals within minced beef samples utilising ambient mass spectrometry. Scientific reports **9**(1), 1–9 (2019)
4. Black, C., Chevallier, O.P., Haughey, S.A., Balog, J., Stead, S., Pringle, S.D., Riina, M.V., Martucci, F., Acutis, P.L., Morris, M., et al.: A real time metabolomic profiling approach to detecting fish fraud using rapid evaporative ionisation mass spectrometry. Metabolomics **13**(12), 1–13 (2017)
5. Breiman, L.: Classification and regression trees. Routledge (2017)
6. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3), 273–297 (1995)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. FAO: The State of World Fisheries and Aquaculture, 2020. FAO (2020). https://doi.org/https://doi.org/10.4060/ca9229en
9. Fix, E., Hodges, J.L.: Discriminatory analysis. nonparametric discrimination: Consistency properties. International Statistical Review/Revue Internationale de Statistique **57**(3), 238–247 (1989)
10. Glorot, X., Bengio, Y.: 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)
11. Hand, D.J., Yu, K.: Idiot's bayes—not so stupid after all? International statistical review **69**(3), 385–398 (2001)
12. Hansen, L.K., Salamon, P.: Neural network ensembles. IEEE transactions on pattern analysis and machine intelligence **12**(10), 993–1001 (1990)
13. He, K., Zhang, X., Ren, S., Sun, J.: 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)
14. 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)
15. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)

16. Ho, T.K.: Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition. vol. 1, pp. 278–282. IEEE (1995)
17. Jha, S.N.: Rapid detection of food adulterants and contaminants: theory and practice. Academic Press (2015)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
19. Kleinbaum, D.G., Dietz, K., Gail, M., Klein, M., Klein, M.: Logistic regression. Springer (2002)
20. Köppen, M.: The curse of dimensionality. In: 5th online world conference on soft computing in industrial applications (WSC5). vol. 1, pp. 4–8 (2000)
21. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
22. Morgan, N., Bourlard, H.: Generalization and parameter estimation in feedforward nets: Some experiments. Advances in neural information processing systems **2** (1989)
23. Pardo, M.Á., Jiménez, E., Pérez-Villarreal, B.: Misdescription incidents in seafood sector. Food Control **62**, 277–283 (2016)
24. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
25. Saxe, A.M., McClelland, J.L., Ganguli, S.: Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. arXiv preprint arXiv:1312.6120 (2013)
26. Sietsma, J., Dow, R.J.: Creating artificial neural networks that generalize. Neural networks **4**(1), 67–79 (1991)
27. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014)
28. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
29. Tran, B., Xue, B., Zhang, M.: Genetic programming for feature construction and selection in classification on high-dimensional data. Memetic Computing **8**(1), 3–15 (2016)
30. Tran, B., Xue, B., Zhang, M.: Genetic programming for multiple-feature construction on high-dimensional classification. Pattern Recognition **93**, 404–417 (2019)
31. 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)
32. Wood, J., Nguyen, B.H., Xue, B., Zhang, M., Killeen, D.: Automated fish classification using unprocessed fatty acid chromatographic data: A machine learning approach. In: Australasian Joint Conference on Artificial Intelligence. pp. 516–529. Springer (2022)
33. Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., Liu, T.: On layer normalization in the transformer architecture. In: International Conference on Machine Learning. pp. 10524–10533. PMLR (2020)