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 focuses on machine learning techniques for marine biomass analysis. It analyses data collected from rapid evaporative ionisation mass spectrometry (REIMS). The paper introduces two novel unsupervised pre-training strategies for transformers. Masked spectra modelling (MSM) and next spectra prediction (NSP). Additionally, the paper investigates decision trees (DT) and genetic programming (GP). Techniques are known for interpretable insights into spectral patterns and chemical attributes. The paper proposes a comprehensive toolbox of machine-learning methods. Models tailored for precise fish speciation and body part identification. Models that perform 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. The global fishing industry catches 100 million tonnes of wild fish each year. Fish processing converts only 40% of that catch into edible parts [9]. The rest is often used for fish oil, fish meal, or discarded. Machine learning techniques can identify high-value fish parts. This helps drive decisions on repurposing fish waste into valuable products. Omega 3 supplements are an example of a high-value product from marine biomass. In 2016, there was a meta-analysis [25] of the global fishing industry. The study found an average mislabelling rate of 30% in the global fishing industry. This emphasizes the need for machine learning tools to prevent mislabelling. Such tools ensure transparency in the seafood supply chain. They help to mitigate fraud and mislabelling. They ensure transparency to support sustainable fishing practices.

2 Related Works

Breiman et al. [6] introduced CART (Classification and Regression Trees). CART has proven effective in analysing mass spectrometry datasets. Notably, It was

applied in EPA-funded projects in the late 1970s and early 1980s. CART was able to detect toxic substances in water and air samples. It did this by correlating fragment ions of molecules in mass spectrometry datasets. Black et al. [5] used REIMS with principal component analysis for fish speciation. They used linear discriminant analysis (PCA-LDA) to classify fish speciation and catch methods. The work differentiates the fish species of seabass and seabream. In a related study, [4] applied REIMS to detect beef adulteration. The study found REIMS was able to identify trace amounts of horse offal adulteration in beef. Wood et al. [34] performed fish speciation and body part classification in marine biomass. This study utilized a gas chromatography dataset. Their work employed the evolutionary computation (EC) approach of particle swarm optimisation (PSO). This paper applies another EC technique. We use multiple class-independent feature construction (MCIFC) [31, 32]. This is a variant of multi-tree genetic programming (MT-GP) [22].

3 Limitations

The rapid evaporative ionisation mass spectrometry dataset introduces three challenges. (1) high dimensionality, (2) few training instances, and (3) time-consuming manual analysis.

- **High-dimensionality** - The dataset provides high-resolution mass spectrometry (HRMS). It has 1023 features. Deep learning, evolutionary computation, and ensembles are well suited towards high-dimensional datasets.
- **Few training instances** - There is a limited number of samples in the dataset. - 108 instances for fish speciation and 30 for fish body part classification. This constructs a few-shot classification task. Unsupervised pre-training, and data augmentation techniques, amortise the limited number of training instances.
- **Time-consuming manual analysis** requires domain expertise in chemistry. Existing methods [4, 5] often rely on manual thresholds. Also, they need hyperparameter tuning by domain experts in chemistry and statistics. This paper automates that analysis. It removes the need for domain expertise in chemistry. Furthermore, it yields rapid inference to match the rapid evaporative ionisation mass spectrometry. Rapid analysis is necessary for REIMS analysis for applications in fish processing.
- **Jargon** - REIMS biomass analysis is a field traditionally dominated by chemistry and statistics. This paper introduces machine learning techniques and their respective terminology to the field. The machine learning terminology facilitates future interdisciplinary collaboration.

4 Theory

4.1 Classifiers

We apply seven standard classifiers to the tasks of fish speciation and body parts classification. Random forest (RF) [18], k-nearest neighbour (KNN) [10], decision

trees (DT) [6], naive bayes (NB) [12], logistic regression (LR) [21], support vector machines (SVM) [7], linear discriminant analysis (LDA) [2]. With an ensemble voting classifier [13] combining them all. [2, 6, 7, 10, 12, 18, 21]. They use default settings from sklearn [26], 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. [33] introduced the transformer architecture. Originally, transformers were proposed for machine translation. This paper adapts (see fig. 1) transformers for marine biomass analysis. The transformer's design includes stacked encoder-decoder layers with residual connections [15]. Experiments try different weight initialisation methods. These include Xavier, Kaiming and orthogonal [11, 14, 27]. Experiments explore post-norm and pre-norm layer normalization [35]. The pre-norm approach performs layer normalisation before the attention and feedforward layers. See fig. 2.

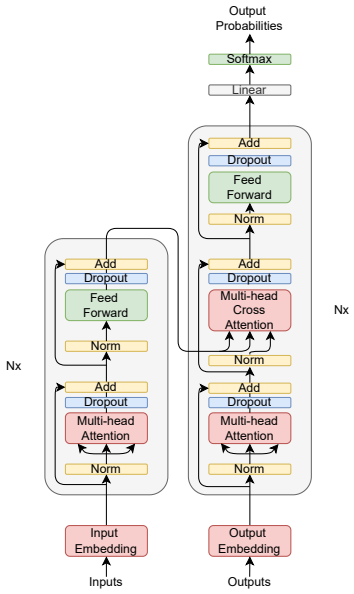
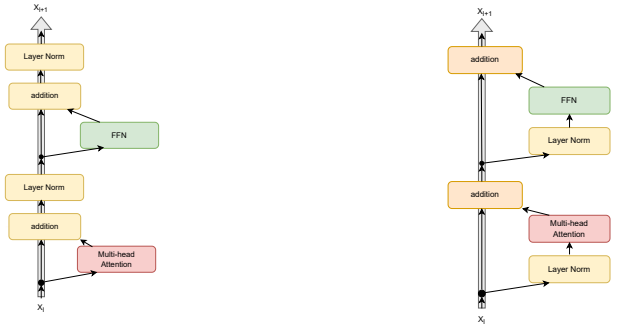


Fig. 1: Transformer architecture

This paper introduces two novel unsupervised pre-training methods for mass spectrometry. BERT [8] inspires these pre-training methods. The first method



(a) Post-norm (b) Pre-norm

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

is masked spectra modelling (MSM). This adapts masked language modelling (MSM) by randomly masking mass spectra. The model predicts those masked spectra. It uses mean square error (MSE) with early stopping. The second method is next spectra prediction (NSP). This adapts next sentence prediction (NSP), it predicts if two halved spectra belong to the same or different spectra. It uses categorical cross-entropy and early stopping. The transformer model in this paper employed the AdamW optimiser [23]. This improves the Adam [20] by decoupling weight decay from the learning rate. Dropout [17, 29] approximates a bagged ensemble of neural networks efficiently. For regularization, label smoothing [30] softens class label targets. It combines one-hot encodings with a uniform distribution. The transformer network utilizes the Gaussian error linear unit (GELU) activations [16]. Used in place of rectified linear unit (ReLU) [1]. Data augmentation inflates the number of training instances. It duplicates each instance five times and injects noise [28]. Effectively expanding the training set five-fold Early stopping [24] saves model parameters whenever validation loss improves. Effectively tuning the hyperparameters of epochs. Table 1 outlines the transformer configuration used.

4.3 Mutli-tree Genetic Programming

The paper uses multiple class independent feature construction (MCIFC) [32]. A novel application of MCIFC for marine biomass analysis using REIMS. MCIFC represents candidate solutions as multiple trees. Given one subtree which corresponds to each class. This approach serves both feature construction and classification. The class prediction is the index of the largest output of the subtree. Otherwise known as a winner-takes-all strategy. For fish speciation (Hoki and Mackerel) use two trees. Fish parts (Fillet, Heads, Livers, Skins, and Guts) use six subtrees. This reduces the feature space from 1023 to 2 or 6 dimensions, respectively. The genetic operators in MCIFC include crossover and mutation. These operators are adapted from conventional genetic programming (GP). Crossover

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

operates between trees of the same class. There is an 80% probability of one-point crossover. Mutation randomly alters one subtree with a 20% probability. The algorithm employs the VarAnd strategy. This means each generation can 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:

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

`inter`class distance measures distance between instance pairs from *different classes*. It aims for greater distances to enhance fitness:

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

`intra`class distance measures distance between instance pairs from the *same class*. It aims for smaller distances to improve fitness:

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

where $d(i,j)$ enumerates over all interclass and intraclass pairs respectively. $|S|$ is the total number of pairs. Together they yield a regularisation term of 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 of the MCIFC method. 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 a form of ambient mass spectrometry (AMS). REIMS was 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. Those tools produce an aerosol (smoke) from tissue samples. This aerosol is then directed into a mass spectrometer's ionisation source. Here ions are formed on a heated collision surface [19]. The data is normalised 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.

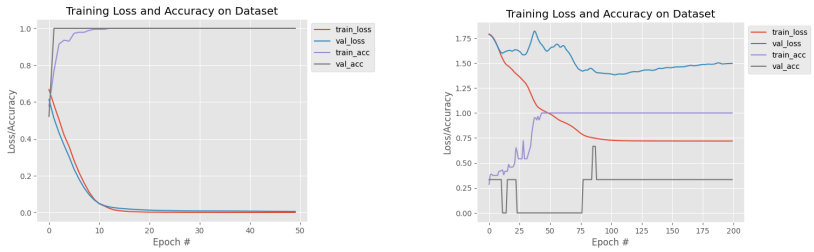
6 Results

Table 3 gives the results of the classifiers. The table gives the average results over 30 independent runs. The best-performing model on the test set is 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

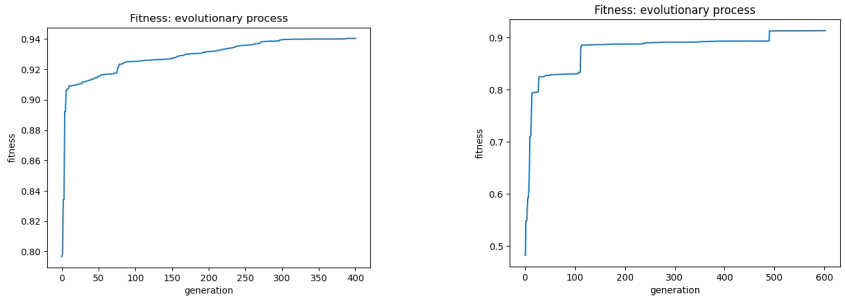
Figure 3 depicts the transformer's superior performance. It exceeds in both fish speciation and fish parts classification. The transformer's attention mechanism captures the spatial connectivity of mass spectrometry data. It attends to feature interactions between mass-to-charge ratios. Decision trees (DT) offer second best near-perfect performance for fish speciation. DT can capture correlations between important features and their class boundaries. The ensemble model performs third best for fish species. Ensembles aggregate diverse and independent models to increase the generalisation ability. 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. Given for both fish speciation and fish part classification, respectively. The fitness function on both graphs reaches a plateau. At his plateau, the regularization term (i.e. intra and inter-class distance) improves. But it has 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. None improved convergence for the transformer model. 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. It outperforms post-LN which achieves 98.33% ± 2.04% accuracy and fails to converge in under 50 epochs. Ablation studies on fish speciation compared, both with and without label smoothing. With label smoothing, the transformer achieves 99.58% ± 1.31% test accuracy. Without label smoothing, the transformer achieves show 99.16% ± 1.66% test accuracy.

8 Interpretability

8.1 Decision Tree

Figure 5 gives the decision tree. This tree 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.

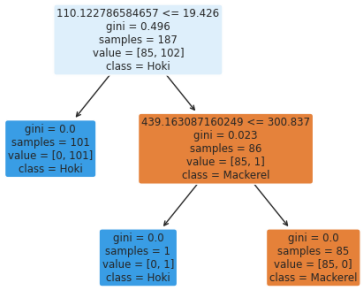


Fig. 5: Fish speciation: decision tree

Assigning compounds to high-resolution mass spectrometry (HRMS) is challenging. This is due to the enormous amounts of metabolites present in homogenised fish tissues. The feature at 110.1128 m/z is consistent with a diphenol group. The feature 439.1631 m/z is consistent with a fragmented phospholipid (1-Lauroyl-2-hydroxy-sn-glycero-3-phosphocholine). Compounds known to vary between fish species and tissues [3].

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. This suggests 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 fragment ions are present in Hoki than in Mackerel.

9 Discussion

Transformers were initially proposed for natural language processing. This paper shows them applied to ambient mass spectrometry for marine biomass analysis. However, despite its accuracy, the complexity renders these models black boxes. This hinders a comprehensive understanding of its decisions. In contrast, genetic programming and decision trees are less accurate. But provide interpretable results. These models offer clearer insights into their functioning. Decision trees detect fish species with high accuracy. They can correlate fragment ions of molecules related to fish speciation.

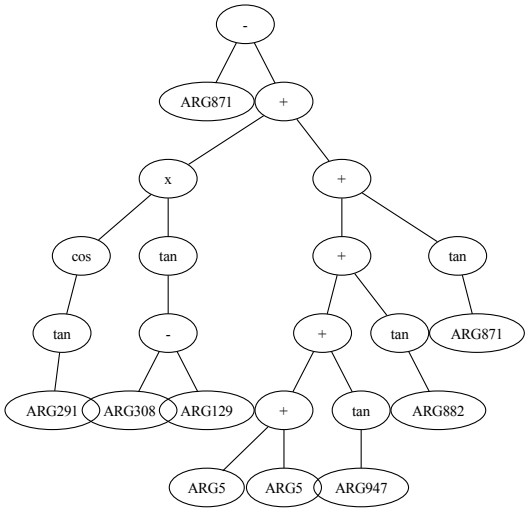


Fig. 6: Fish species: Mackerel

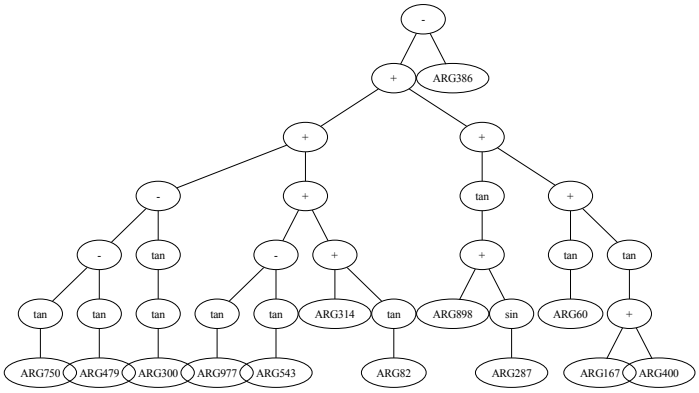


Fig. 7: Fish species: Hoki

References

1. Agarap, A.F.: Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375 (2018)
2. Balakrishnama, S., Ganapathiraju, A.: Linear discriminant analysis-a brief tutorial. Institute for Signal and information Processing **18**(1998), 1–8 (1998)
3. Bettjeman, B.I., Hofman, K.A., Burgess, E.J., Perry, N.B., Killeen, D.P.: Seafood phospholipids: extraction efficiency and phosphorous nuclear magnetic resonance spectroscopy (31p nmr) profiles. Journal of the American Oil Chemists' Society **95**(7), 779–786 (2018)
4. 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)
5. 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)
6. Breiman, L.: Classification and regression trees. Routledge (2017)
7. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3), 273–297 (1995)
8. 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)
9. FAO: The State of World Fisheries and Aquaculture, 2020. FAO (2020). https://doi.org/https://doi.org/10.4060/ca9229en
10. Fix, E., Hodges, J.L.: Discriminatory analysis. nonparametric discrimination: Consistency properties. International Statistical Review/Revue Internationale de Statistique **57**(3), 238–247 (1989)
11. 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)
12. Hand, D.J., Yu, K.: Idiot's bayes—not so stupid after all? International statistical review **69**(3), 385–398 (2001)
13. Hansen, L.K., Salamon, P.: Neural network ensembles. IEEE transactions on pattern analysis and machine intelligence **12**(10), 993–1001 (1990)
14. 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)
15. 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)
16. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
17. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580 (2012)
18. Ho, T.K.: Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition. vol. 1, pp. 278–282. IEEE (1995)

19. Jha, S.N.: Rapid detection of food adulterants and contaminants: theory and practice. Academic Press (2015)
20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
21. Kleinbaum, D.G., Dietz, K., Gail, M., Klein, M., Klein, M.: Logistic regression. Springer (2002)
22. Koza, J.R., et al.: Genetic programming II, vol. 17. MIT press Cambridge (1994)
23. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
24. Morgan, N., Bourlard, H.: Generalization and parameter estimation in feedforward nets: Some experiments. Advances in neural information processing systems **2** (1989)
25. Pardo, M.Á., Jiménez, E., Pérez-Villarreal, B.: Misdescription incidents in seafood sector. Food Control **62**, 277–283 (2016)
26. 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)
27. 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)
28. Sietsma, J., Dow, R.J.: Creating artificial neural networks that generalize. Neural networks **4**(1), 67–79 (1991)
29. 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)
30. 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)
31. 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)
32. Tran, B., Xue, B., Zhang, M.: Genetic programming for multiple-feature construction on high-dimensional classification. Pattern Recognition **93**, 404–417 (2019)
33. 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)
34. 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)
35. 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)