Machine Learning for Fish Oil Anaylsis

No Author Given

No Institute Given

Abstract. In fish processing worldwide, 60% of fish is not served. This fish wastage can be repurposed into other products such as fish oils and fish feed. It is difficult and requires domain expertise to effectively repurpose fish biomass. Gas chromatography (GC) can be used to identify tissue samples in fish processing. The existing analytical chemistry techniques for processing GC data are manual and time-consuming. To reduce biomass waste we must incentivise biomass re-use, we reduce this barrier of entry through automation. Here, we explore classification algorithms for fish oil data that automate and significantly reduce the time required to process GC data. Visualisation is used to explore the interpretability of the models such that their efficacy can be verified for use in a factory setting. The fish oil data is high-dimensional and low sample size.

Keywords: feature selection · classification · gas chromatography · support vector machine · particle swarm optimization · high-dimensional data

1 Introduction

Omega-3 supplements are a high-value product made by repurposing fish waste in food processing. The waste is further refined into fish oil capsules commonly bought and sold at pharmacies. Certain fish species and parts are richer in omega-3 fatty acids, this makes these samples high value. We can identify valuable fish oils by classifying their species and part.

Gas chromatography (GC) is a chemistry technique used to analyze fish oils [3]. It analyzes the chemical structure of a fish oil sample [17]. The technique is used for quality assurance in food processing. To repurpose fish waste effectively we must know what is in it. Chemists manually compare GC samples to reference data to determine their class.

Classification automates this process by training a model on an existing dataset manually labelled by chemists. Given a fish oil sample, we can identify the fish species (i.e. Bluecod, Tarakihi), and part (Head, Fin). Many classifiers [4,5,6,16,1] can used to indentify the class of an instance from its features. These models can be organized into five categories: instance-based, probabilistic, tree-based, bagging and kernel-based, respectively. In this paper, we evaluate the performance of each classifier on both datasets.

Many feature selection methods [15,2,10,7] are available to eliminate redundant features. These methods can be grouped into four categories: statistics,

information theory, similarity, and swarm intelligence, respectively. In this paper, the SVM classifier [1] is evaluated on the feature subsets selected by each FS method, on both datasets.

By only using important features for classification, it is easier to interpret what features/patterns the SVM model utilises. Interpretable models are capable of troubleshooting and diagnosis by domain expertise in real-world applications. A model that is both interpretable and accurate has the potential to be deployed in a factory setting, this eliminates the need for manual work.

2 Background

2.1 Gas Chromatorgraphy

Gas chromatography (GC) is a technique for the analysis of chemical compounds [3,17]. The process separates compounds based on their boiling point and molecular weight. A compound is injected as a liquid, then heat is applied to vaporize it into a gas. This process is referred to as a phase transition. The speed at which a compound is vaporized depends on its boiling point. The vaporized gases travel through a long coiled tube. That tube has a detector at the end, this detects the rate and intensity at which compounds reach the tube's end.

Figure 1 shows a gas chromatograph from the dataset. Chemists match the peaks from known reference samples to classify unknown samples. Analysis can identify an unlabelled sample since they share similar peaks to previously labelled data. There is noise in the peaks caused by time-shift - a known limitation of the measurement technique addressed in previous works [20,22]. GC is not a definitive technique, so it is often used in conjunction with other techniques like Mass spectrometry [17].

The existing task [3,17] of classifying chemical compounds based on a chromatograph is laborious. The spikes on the graph represent peaks. Each peak represents a resolved chemical compound. Chemists integrate the area under each peak, and compare this to a reference sample, to classify the compound. GC must be performed slowly to ensure that the peaks are not too broad. This ensures each peak resolves and represents a single compound. Once we know what compounds are present in a sample it becomes possible to identify what the sample is. For this fish oil data, we classify a sample into two categories - species and part.

2.2 Feature Selection

There are two datasets with the same features. Those features are the high dimensional GC data, the class labels are the species and part. The curse of dimensionality [11] introduces problems: (1) computationally inefficient to evaluate the entire 4800 feature problem space. (2) Classification models are likely to overfit, learning noise in the data from irrelevant features. (3) Classifier models that utilize high-dimensional feature spaces are difficult to interpret, and it is difficult for domain experts to perform diagnoses or troubleshoot.

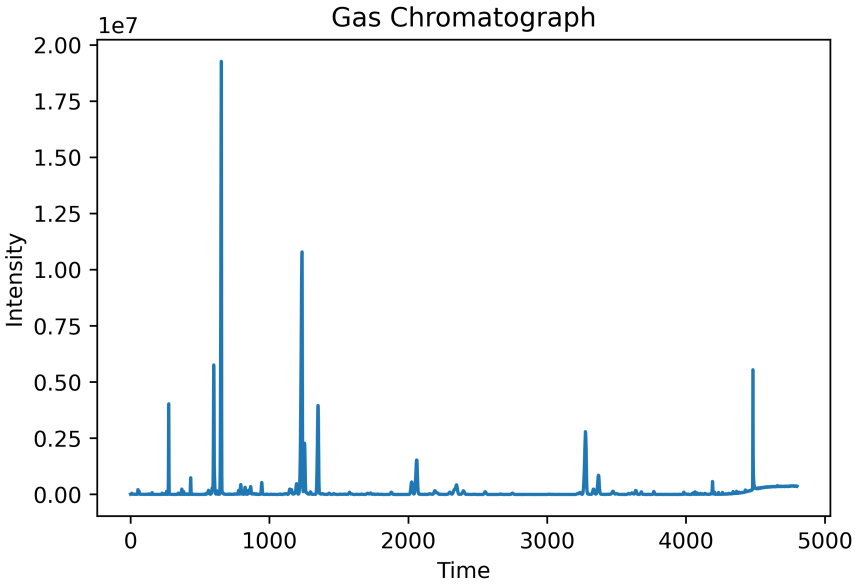


Fig. 1: A Gas Chromatograph is the artifact of the GC method. This chromatograph is taken from the fish oil dataset. The detection is used to visualize intensity (y axis) and time (x axis) on a chromotrogaph. The distinguishable peaks are unique signatures where distinct chemical compounds can likely to be identified.

Figure 1 shows at most several points of interest (peaks), and many negligible features, for an average chromatograph from the dataset. Figure 2 shows the correlation between the features. These figures illustrate the redundancy within the dataset, showing many irrelevant features. Feature selection reduces the size of the feature space by eliminating redundant and correlated attributes. To address these issues of high-dimensionality, we employ feature selection to select a subset of relevant features. Feature selection can (1) improve computational efficiency with a reduced dataset, (2) increase classification performance, and (3) lets classifiers produce simpler models which are in turn easier to troubleshoot. We employ heuristic-based algorithms to address the combinatorial explosion of searching the possible feature subsets. The algorithms used in this paper are introduced in further detail.

Liu et al. proposed Chi2 [15], a feature selection method via discretization. The algorithm is a generalized version of ChiMerge [8] that determines a good χ^2 threshold automatically from the data.

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}} \tag{1}$$

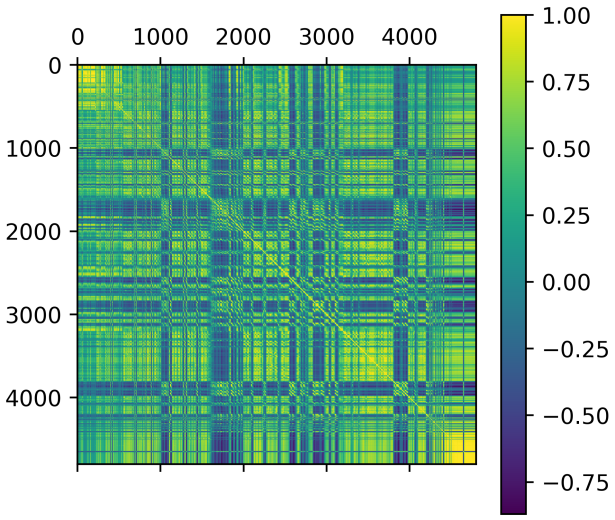


Fig. 2: Fish Oil Data pairwise correlation for the feature set using Pearson's correlation coefficient. There are many redundant features

The formula for calculating the χ^2 statistic is given by eq. (1). The technique performs both feature selection and discretization, making it ideal for continuous numeric fish oil data. The method can increase predictive performance, and efficiency (time/memory) and simplify models.

Ding and Peng proposed Maximum Redundancy Maximum Relevance (MRMR) [2] as a feature selection method for gene microarray data. MRMR is a filter-based multi-objective method,

$$I(X;Y) = D_{KL}(P_{(X,Y)}||P_X \otimes P_X) \tag{2}$$

These are both measured in terms of mutual information, given by eq. (2), Let (X, Y) be a pair of random variables, take the KL divergence [13] between their join distribution $P_{(X,Y)}$ and the product of their maginal distribution $P_X \otimes P_X$.

$$f^{mRMR} = I(Y;X_i) - \frac{1}{|S|}\sum_{X_s \in S} I(X_s;X_i) \tag{3}$$

The MRMR feature importance score is given in eq. (3), for full derivation see [15,23]. It balances both, the relevance for predicting outcome $I(Y;X_i)$, and the redundancy within features $I(X_s;X_i)$, scaled by the size of the feature subset $|S|$. It builds a set of features, adding the X_i with the maximum feature importance to the selected subset.

Robnik proposed Relief-F [10], a feature selection method based on k nearest neighbours. The algorithm extends Relief [9], the extension is noise-tolerant, handles incomplete data and multi-class problems. Relief estimates feature importance based on their ability to separate other nearby instances. With intuition, a good feature can distinguish between classes. Relief-F is noise tolerant by averaging the contribution of the k nearest neighbours. Missing values of attributes are treated probabilistically[18]. Relief-F handles multi-class problems by taking the weighted average of near misses to all classes.

$$W[A] := W[A] - \frac{1}{m}\left(diff(A,R,H) + \sum_{C \neq class(R)} P(C) \times diff(A,R,M(C))\right) \tag{4}$$

The attributes are estimated with the weight update function given by eq. (4).

Kennedy and Eberhart proposed PSO [7] swarm intelligence method for nonlinear optimization. This imitates social behaviour, e.g. birds flocking or fish schooling. Their synchronous movement was modelled as a function of each fish maintaining equal distance to its neighbours. To give intuition, we imagine a simple model of a schooling fish in search of food. With p_{best} an individual fish's best spot, and g_{best} the best spot the school has found. Each fish updates its movement in equal parts towards p_{best} and g_{best}, both multiplied by a stochasticity factor. The stochasticity introduces randomness, which makes the fish overshoot their target about half the time. The overshooting has fundamental to the success of the model, it allows the school to explore uncharted waters (unknown regions of the problem domain). The social model was simplified into a particle swarm, which allowed collisions where two particles share the same location.

Comparisons from these algorithms.

- Talk about the normalization term in MRMR $\frac{1}{|S|}$ and ReliefF $\frac{1}{M}$, make noise-tolerant
- chi2 does not have a normalization term. Also, the $(.)^2$ operator, makes the χ^2 metric especially sensitive to outliers with large positive or negative values.
- Compare χ^2 threshold to reliefF normalization term m, both parameters tuned by the algorithm.

3 Data processing

- Why the raw data is not applicable to existing classification algorithms?
- Extracting datasets that are ready for classification algorithms:
  - Sum up the intensity.
  - Aligning missing packets.
- Overview of extracted data.

4 Classification

We measure the predictive ability of classifiers on both the fish species and part dataset. We are looking for the model with the highest accuracy. As a result, we start broadly by exploring a variety of models from the different families of AI, and then we narrow and refine the search.

For each of the following experiments, the same experiment setup is used. We use stratified cross-validation ($k = 10$) to measure the classification accuracy. Each method has its performance recorded on the same cross-folds. Then we average over 30 independent runs. This experimental setup evaluates performance on both the fish species and part datasets.

4.1 Classification Algorithms

We examine 5 classification models:

1. K-Nearest Neighbors [4]
2. Random Forest [6]
3. Naive Bayes [5]
4. Decision Tree [16]
5. Support Vector Machine [1]

Table 1: Accuracy for different classification techniques. Accuracy is given as the stratified k-fold cross validation over 30 independent runs.

Dataset	Method	AvgTrain ± Std	AveTest ± Std
Species	KNN	83.57 ± 1.80	74.88 ± 12.54
	RF	1.00 ± 0.00	85.65 ± 10.76
	DT	1.00 ± 0.00	76.98 ± 13.12
	NB	79.54 ± 1.60	75.27 ± 4.35
	SVM	**1.00 ± 0.00**	**98.33 ± 5.00**
Part	KNN	68.95 ± 3.49	43.61 ± 13.48
	RF	1.00 ± 0.00	72.60 ± 16.15
	DT	1.00 ± 0.00	60.14 ± 14.57
	NB	65.54 ± 2.69	48.61 ± 12.19
	SVM	**1.00 ± 0.00**	**87.14 ± 8.52**

Table 1 shows for the random forest, decision tree and support vector machine have perfect training accuracy. The decision tree and random forest overfit the training data. Only the SVM achieves similar performance on the test data. The SVM classifier outperforms the other classifiers. It does so for the test set for both the species and part datasets.

4.2 Discussion

We evaluated an ensemble of classification techniques. Naive Bayes performed poorly. This is likely due to the assumption of conditional independence between features. KNN also performed poorly. This is likely due to the high dimensionality of the data. Points drawn from high-dimensional spaces tend to never be close together. SVM provided the best results. This model can identify fish species from gas chromatography data with near-perfect accuracy. This prompted further investigation into this technique.

Classification accuracy for all models was better for the fish species than the part. This suggests tissue samples for different species may have distinct chemical compositions. Yet, different fish parts may have fewer underlying structural differences. For GC data the intra-class variation between species provides a larger signal than part variation. For example, we expect there to be more difference between a tarakihi and a bluecod, than there is a similarity between two livers from different species.

4.3 Weight Analysis

Visualisation Two heuristics are optimized when selecting a suitable model: interpretability and accuracy. Interpretability is important for verification in a safety-critical environment. We intend to employ the chosen model in a factory setting. Accuracy is preferable, but not at the expense of interpretability. The efficacy of the model must be explainable through domain knowledge. Or else it is difficult to ensure reliability. The focus on interpretability ensures the model can be used in the real world.

Model interpretability is explored through visualisation. We aim to uncover learnt patterns that can be verified with domain knowledge. The desired algorithm should strike a balance between predictive performance and semantically meaningful features.

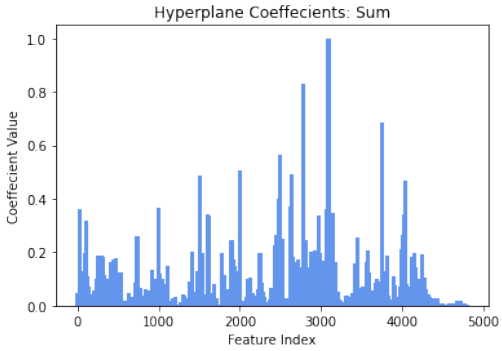
What constitutes semantic meaning varies from one domain to another. It is easy to build intuition for semantic meaning in computer vision and natural language processes, they correspond to recognisable images and structured text. In the domains of gas chromatography and fish processing, our meaning is derived from performance on the classification task(s) and similarity to underlying chemical compounds. We expect models that generate knowledge that can be verified with domain expertise. For example, important features will correspond to timestamps of important chemicals in the GC data.

SVM Hyperplane Coefficients Cortes and Vapnik proposed the Support Vector Machine (SVM) [1]. This model creates a hyperplane that can draw distinct class boundaries between classes. We call these class boundaries the support vectors. We are performing multi-class classification, so it used a one-vs-all approach [19]. This creates a divide between one class and the rest, then repeats for the other classes.

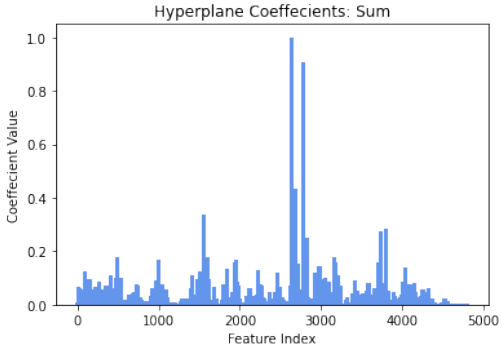
The l1 regularization term leads to sparse models. So, they include fewer features - making them easier to interpret. Eq 5 defines the total hyperplane as

$$\beta_{\mathrm{t}} = \mathrm{minmax}(\sum_{\mathrm{c}\in\mathrm{C}} |\beta_{\mathrm{c}}|) \tag{5}$$

where there is the number of classes ($c \in C$) sets of hyperplane coefficients. β_{t} coefficient as the sum of hyperplane coefficients magnitude for each class β_{c}. We normalize the coefficients with a min-max feature scaling.



(a) Fish Species: Hyperplane Coefficients



(b) Fish Part: Hyperplane Coefficients

Fig. 3: Hyperplane coefficients β_t. The normalized sum of the magnitude of the coefficients for each class is given in Eq 5. (a) Coefficients for the fish species dataset. (b) Coefficients for the fish part dataset.

The total hyperplane for both datasets is given in Figure 3. We visualize the hyperplane to approximate the important features. The outliers correspond to feature timestamps that are important for drawing class boundaries. These are chemical compounds that separate the fish part and species, respectively.

5 Feature Selection

- Why feature selection on this data?
- Brief the main ideas of the feature selection algorithms that were used.

For each method, we measure classification accuracy with an SVM model [19]. It has linear kernel, l1 regularization [18] and 10,000 maximum iterations. We examine 3 feature selection methods from [14], each with default parameters. A PSO implementation described in [7]. The fitness function balances the SVM classification accuracy and the number of features selected. The population size 30, iterations 100, $[x]^n$ where $x \in [0, 1], n = 4800$, $v_{max} = 0.2$, $v_{min} = -0.2$

1. Chi^2 [15]
2. Minimum Redundancy Maximum Relevance [2]
3. Relief-F [18]
4. Particle Swarm Optimization [7]

We measure classification accuracy as a function of feature number. We compared this for several FS methods. Due to limitations, PSO optimizes feature number k automatically. So, to compare its performance, we plot the results of 30 independent runs.

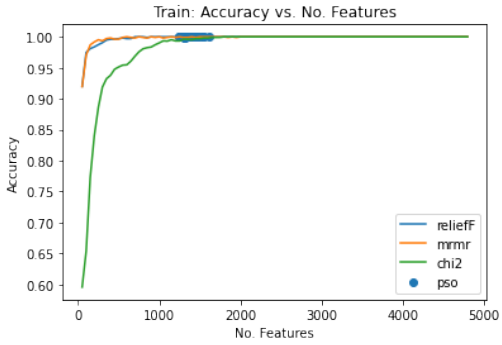
5.1 Fish Species

Figure 4a shows accuracy for fish species. We show accuracy on the training set for each feature selection method. At $k = 1050$ all feature selection methods achieve 100% accuracy on the training set. The SVM fits the training data for each method using a fraction of the full feature set. Figure 4b shows accuracy for fish species. We show test set accuracy for each feature selection method. The accuracy reaches a plateau (96% accuracy) at around $k = 1050$ features for all methods. The test performance is less than the train performance, yet the test accuracy is still very high. This suggests the model can generalize well on unseen data for the fish species.

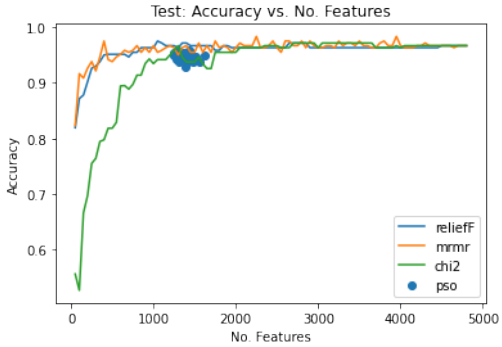
5.2 Fish Part

Figure 5 shows accuracy for part dataset. We show train accuracy for each feature selection method. All feature selection methods struggle to fit the training set for the fish part. Even with the full feature set, a perfect train accuracy is never reached. Figure 5b shows accuracy for part dataset. We show the test accuracy for each feature selection method. The classification accuracy fluctuates for all feature selection methods. At around $k = 1050$ features, it begins to decrease. The training accuracy improves, as the test does not from this point onwards. The SVM is overfitting to noise (redundant features) in the training set.

- Compare the performance of selected feautres and use all feautres.
- (Optional) analyse the selected features .



(a) Species: Training set



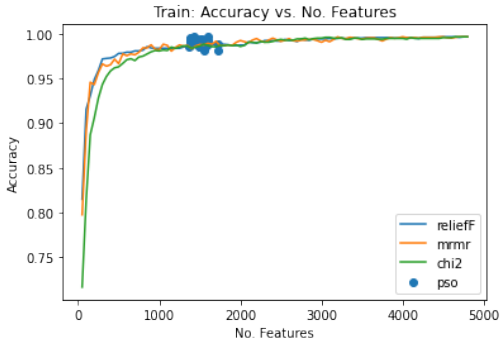
(b) Species: Test set

Fig. 4: Fish species dataset: Classification accuracy for feature selection methods for a given k. We measure the balanced accuracy of k-fold cross-validation. We compare Relief-F, maximum relevance - minimum redundancy (MRMR), chi^2, and particle swarm optimisation (PSO).

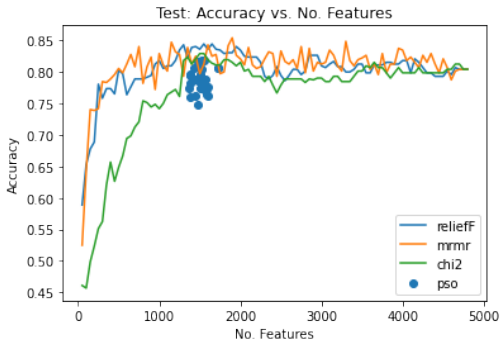
5.3 Disucssion

Feature selection methods helped reduce dimensionality. We evaluated performance with an SVM classifier. Which, Relief-F and PSO were best for fish species and part, respectively. Relief-F can identify conditional dependencies between features when providing feature rankings. Relief-F algorithms are robust and noise-tolerant, which explains their superior performance. PSO provides a combination of global and local searches. A search through a near-infinite combinatorial space of possible feature subsets. This stochastic method is computationally expensive but can offer effective solutions.

For both general and specific cases, and across all methods, the fish species have lower variance than the fish part in classification accuracy. The classification results support this, they also show higher test accuracy for fish species, than for fish parts. They suggest different fish parts may have fewer underlying structural differences.



(a) Part: Training set



(b) Part: Test set

Fig. 5: Fish part dataset: classification accuracy for feature selection methods for a given k. We measure the balanced accuracy of k-fold cross-validation. We compare Relief-F, maximum relevance - minimum redundancy (MRMR), chi^2, and particle swarm optimisation (PSO).

For the general case and both datasets, a lot of interesting behaviour happens at $k = 1050$. The fish species reach a plateau, but the fish part accuracy begins to decrease. Accuracy comparable to or better than the full dataset is possible with 21.8% of its features.

For the fish species dataset, we see high accuracy with very few features. Relief-F and MRMR can achieve above 90% classification accuracy with $k = 50$. Chi^2 is not able to mirror this performance. This shows that Relief-F and MRMR are very effective feature selection methods for this task.

The PSO may not have a hyperparameter for feature number k. Instead, it automates the selection of this parameter. Yet, it achieves comparable results to other state-of-the-art methods. This automation may prove useful for automating the classification task for online learning. In a factory, we may want to train a model as new data arrives. PSO requires less human intervention, yet still, provides competitive performance.

MRMR and Relief-F both have high accuracy with very few features on the fish species dataset. This suggests that few features are required to construct a reasonable representation of a fish tissue sample. This is a good indication that the fish species dataset contains less noise.

6 Conclusions and Future Work

This paper has demonstrated an interpretable and effective method for fish oil analysis. The method can be understood, domain experts can understand the important features in the decision-making. Not only have we found effective classification and feature selection techniques, but we have also tried to explain their performance with visualisation and analytical results. We can draw many conclusions from the analytical results and visualisations, but here we recall the most important:

1. Fish species are easier to predict than fish parts - there is more intra-class variation within fish species than there is a similarity between the same part from different fish.
2. The Linear SVM classifier performs better for both classification tasks - the fish oil data is linearly separable on a hyperplane.
3. This model can achieve 90% fish species accuracy using only 1% (k = 50) of its features, enabling efficient training, and showing many redundant and correlated features.
4. Weight analysis of the SVM hyperplane serves as a useful approximation to gain insight into the model, useful for troubleshooting/diagnosis from domain experts in biochemistry.

LinearSVC applied to GC fish oil data can help to reduce waste in fish processing. This ensures more sustainable eco-friendly practices in fish processing. Waste reduction and repurposing is a rising tide that lifts all boats. Sustainable practices will leave plenty of fish in the sea, preserving resources for future generations.

It is worth noting that the classification and feature selection methods presented in this paper could be extended to potentially improve performance. This is particularly useful for the lower-accuracy fish part dataset. Here we give a non-exhaustive list of possible extensions: (1) Further investigate feature selection by evaluating the FQC MRMR-variant proposed in [23]. (2) Extend classifier to S3VM [21], a semi-supervised SVM that can utilize unlabelled GC data. This uses unlabelled data to ensure the decision boundaries are drawn through low-density areas. (3) NIST has published a Gas Chromatographic Retention dataset [12]. NIST could be applied to S3VM described, as well as pre-training/transfer learning approaches.

References

1. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3), 273–297 (1995)

2. Ding, C., Peng, H.: Minimum redundancy feature selection from microarray gene expression data. Journal of bioinformatics and computational biology **3**(02), 185–205 (2005)
3. Eder, K.: Gas chromatographic analysis of fatty acid methyl esters. Journal of Chromatography B: Biomedical Sciences and Applications **671**(1-2), 113–131 (1995)
4. Fix, E., Hodges, J.L.: Discriminatory analysis. nonparametric discrimination: Consistency properties. International Statistical Review/Revue Internationale de Statistique **57**(3), 238–247 (1989)
5. Hand, D.J., Yu, K.: Idiot's bayes—not so stupid after all? International statistical review **69**(3), 385–398 (2001)
6. Ho, T.K.: Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition. vol. 1, pp. 278–282. IEEE (1995)
7. Kennedy, J., Eberhart, R.C.: Particle swarm optimization. In: Proceedings of ICNN'95-international conference on neural networks. vol. 4, pp. 1942–1948. IEEE (1995)
8. Kerber, R.: Chimerge: Discretization of numeric attributes. In: Proceedings of the tenth national conference on Artificial intelligence. pp. 123–128 (1992)
9. Kira, K., Rendell, L.A.: A practical approach to feature selection. In: Machine learning proceedings 1992, pp. 249–256. Elsevier (1992)
10. Kononenko, I.: Estimating attributes: Analysis and extensions of relief. In: European conference on machine learning. pp. 171–182. Springer (1994)
11. Köppen, M.: The curse of dimensionality. In: 5th Online World Conference on Soft Computing in Industrial Applications (WSC5). vol. 1, pp. 4–8 (2000)
12. Kovats, E.: Gas chromatographic characterization of organic compounds. i. retention indexes of aliphatic halides, alcohols, aldehydes, and ketones. Helv. Chim. Acta **41**, 1915 (1958)
13. Kullback, S., Leibler, R.A.: On information and sufficiency. The annals of mathematical statistics **22**(1), 79–86 (1951)
14. Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., Liu, H.: Feature selection: A data perspective. ACM Computing Surveys (CSUR) **50**(6), 94 (2018)
15. Liu, H., Setiono, R.: Chi2: Feature selection and discretization of numeric attributes. In: Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence. pp. 388–391. IEEE (1995)
16. Loh, W.Y.: Classification and regression trees. Wiley interdisciplinary reviews: data mining and knowledge discovery **1**(1), 14–23 (2011)
17. Restek: High-resolution gc analyses of fatty acid methyl esters (fames)
18. Robnik-Šikonja, M., Kononenko, I.: Theoretical and empirical analysis of relieff and rrelieff. Machine learning **53**(1), 23–69 (2003)
19. Sklearn: 1.13. feature selection
20. Tomasi, G., Van Den Berg, F., Andersson, C.: Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. Journal of Chemometrics: A Journal of the Chemometrics Society **18**(5), 231–241 (2004)
21. Zemmal, N., Azizi, N., Dey, N., Sellami, M.: Adaptative s3vm semi supervised learning with features cooperation for breast cancer classification. Journal of Medical Imaging and Health Informatics **6**(4), 957–967 (2016)
22. Zhang, D., Huang, X., Regnier, F.E., Zhang, M.: Two-dimensional correlation optimized warping algorithm for aligning gc× gc- ms data. Analytical Chemistry **80**(8), 2664–2671 (2008)

23. Zhao, Z., Anand, R., Wang, M.: Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform. In: 2019 IEEE international conference on data science and advanced analytics (DSAA). pp. 442–452. IEEE (2019)