VICTORIA UNIVERSITY OF WELLINGTON TE HERENGA WAKA

Plant & Food RESEARCH RANGAHAU AHUMARA KAI

Callaghan Innovation New Zealand's Innovation Agency

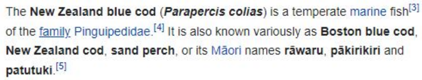
Automated Fish Classification

Using Unprocessed Fatty Acid Chromatographic Data

By Jesse Wood
Supervisors: Bach Hoai Nguyen, Bing Xue

Industry Partners: Daniel Killeen, Kevin Mitchell

Outline

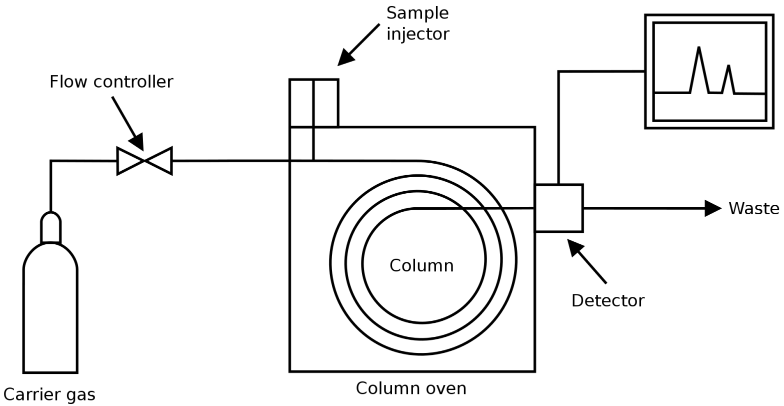


The **New Zealand blue cod** (***Parapercis colias***) is a temperate marine fish[3] of the family Pinguipedidae.[4] It is also known variously as **Boston blue cod**, **New Zealand cod**, **sand perch**, or its Māori names **rāwaru**, **pākirikiri** and **patutuki**.[5]

1. Introduction
2. Gas Chromatography (GC)
3. Preprocessing
4. Classification
5. Feature Selection
6. Why not deep learning?
7. Genetic Programming

Gas Chromatography (GC)

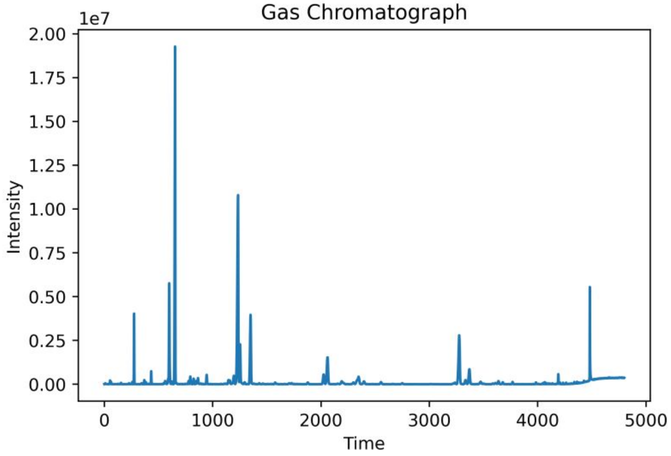
- Analysing GC (Eder 1995) is expensive and time consuming task.
- Steps:
  1. Apply heat to liquid.
  2. Evaporate into gas.
  3. Travel through long tube.
  4. Detector measures intensity.



https://en.wikipedia.org/wiki/Gas_chromatography

Gas Chromatography (GC)

- Analysing GC (Eder 1995) is expensive and time consuming task.
- Steps:
  1. Apply heat to liquid.
  2. Evaporate into gas.
  3. Travel through long tube.
  4. Detector measures intensity.
- Chromatograph is time vs. intensity.
- **Gist**: Molecules have distinct and different boiling points, these correspond to known timestamps.



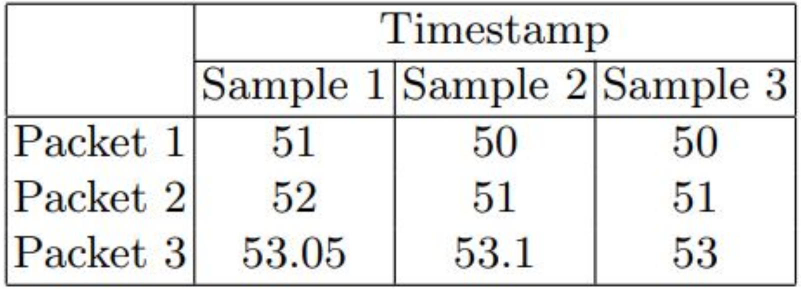
The **Australasian snapper** (*Chrysophrys auratus*) or **silver seabream** is a species of porgie found in coastal waters of Australia, Philippines, Indonesia, mainland China, Taiwan, Japan and New Zealand. Its distribution areas in the Northern and Southern Hemispheres are disjunct.[2] Although it is almost universally known in Australia and New Zealand as **snapper**, it does not belong to the snapper family, Lutjanidae. It is highly prized as an edible fish, with a sweet sea taste and a firm texture.

Preprocessing

Prionotinae is a subfamily of demersal, marine ray-finned fishes, part of the family Triglidae. The fishes in this subfamily are called **sea robins** and are found in the Western Atlantic and Eastern Pacific Oceans, the other two Triglid subfamilies are called gurnards.

- Instrumental drift, an artifact of the gas chromatography technique, leads to missing packets in the data.
- Important to align GC data (fill in missing packets) for better results.
- Steps:
  1. Find missing timestamps,
  2. impute with zero filling
- Existing works, (Tomasi 2004, Zhang 2008), provide more advance techniques to handle instrumental drift.

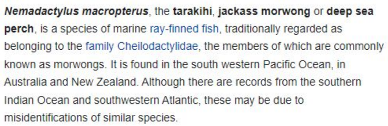
Table 1: Inconsistent Timestamps



	Timestamp		
	Sample 1	Sample 2	Sample 3
Packet 1	51	50	50
Packet 2	52	51	51
Packet 3	53.05	53.1	53

Classification: Motivations

- Reduce - byproduct in fish processing.
- Reuse - identify high-value fish oil.
- Recycle - refine fish oil, rich in omega-3, into supplements.
- Replace - Automate expensive/time-consuming of analyzing GC fish oil data.



Nemadactylus macropterus, the **tarakihi**, **jackass morwong** or **deep sea perch**, is a species of marine ray-finned fish, traditionally regarded as belonging to the family Cheilodactylidae, the members of which are commonly known as morwongs. It is found in the south western Pacific Ocean, in Australia and New Zealand. Although there are records from the southern Indian Ocean and southwestern Atlantic, these may be due to misidentifications of similar species.

https://static.countdown.co.nz/assets/product-images/zoom/9400097038961.jpg

https://www.nutraingredients.com/Article/2016/03/10/Small-fish-oil-doses-enough-to-lower-blood-pressure-RCT

Classification: Experiment

- Two datasets: fish species, body parts.
- 5 classifiers:
  1. KNN (Fix 1989)
  2. RF (Ho 1995)
  3. DT (Loh 2011)
  4. NB (Hand 2001)
  5. SVM (Cortes 1995)
- Run each classifier 30 independent runs, 10-fold cross-validation, balanced accuracy.
- Body parts more difficult to classify than fish species.
- SVM with Linear Kernel performs best for both datasets. Near perfect (98.33%) accuracy for fish species.

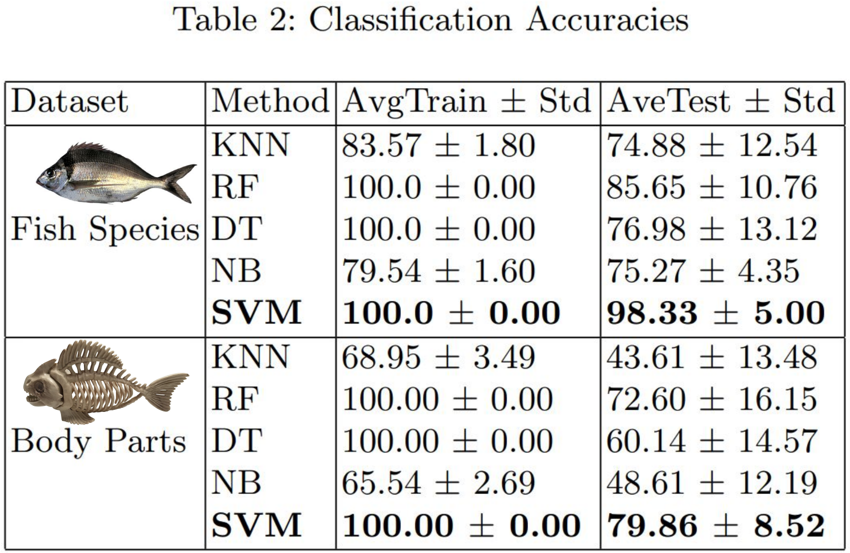


Table 2: Classification Accuracies

Dataset	Method	AvgTrain ± Std	AveTest ± Std
Fish Species	KNN	83.57 ± 1.80	74.88 ± 12.54
	RF	100.0 ± 0.00	85.65 ± 10.76
	DT	100.0 ± 0.00	76.98 ± 13.12
	NB	79.54 ± 1.60	75.27 ± 4.35
	SVM	**100.0 ± 0.00**	**98.33 ± 5.00**
Body Parts	KNN	68.95 ± 3.49	43.61 ± 13.48
	RF	100.00 ± 0.00	72.60 ± 16.15
	DT	100.00 ± 0.00	60.14 ± 14.57
	NB	65.54 ± 2.69	48.61 ± 12.19
	SVM	**100.00 ± 0.00**	**79.86 ± 8.52**

Classification: Results

- RF, DT have 100% training accuracies, but poor test accuracy. Low sample size → DT, RF memorize rather than learn, overfit training set, fail to generalize to test.
- KNN performs poorly → distance-based classification algorithm suffers the most from the large number of features.
- NB performs poorly, since it assumes conditional independence between features, that may not be true in the fish datasets.
- The SVM has best test performance, with 98.33% and 79.86% for fish species and body parts, respectively. Linear SVM well suited to GC fish oil data.
- Classifying body parts is harder. Perhaps, more similarities between fish of same species, than body parts from different species.

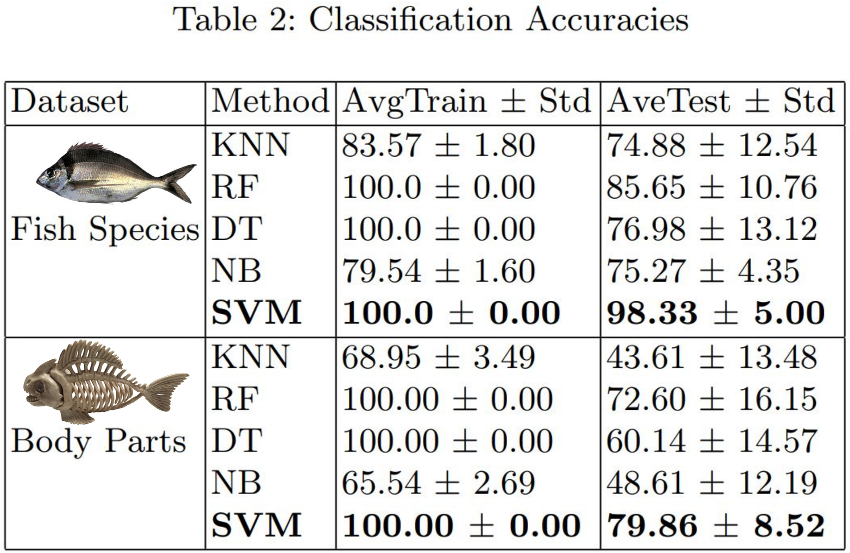
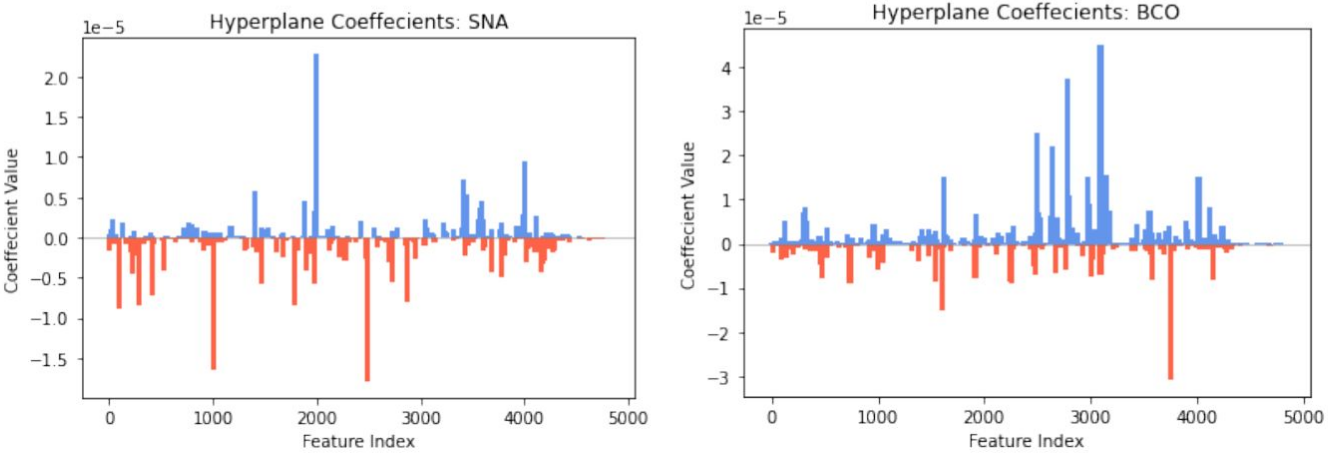


Table 2: Classification Accuracies

Dataset	Method	AvgTrain ± Std	AveTest ± Std
Fish Species	KNN	83.57 ± 1.80	74.88 ± 12.54
	RF	100.0 ± 0.00	85.65 ± 10.76
	DT	100.0 ± 0.00	76.98 ± 13.12
	NB	79.54 ± 1.60	75.27 ± 4.35
	SVM	**100.0 ± 0.00**	**98.33 ± 5.00**
Body Parts	KNN	68.95 ± 3.49	43.61 ± 13.48
	RF	100.00 ± 0.00	72.60 ± 16.15
	DT	100.00 ± 0.00	60.14 ± 14.57
	NB	65.54 ± 2.69	48.61 ± 12.19
	SVM	**100.00 ± 0.00**	**79.86 ± 8.52**

Classification: Interpret Linear SVM

- Interpretable to: statisticians, chemists, mathematicians, computer scientists. Requires knowledge of linear algebra, dot products, classification.
- SVM uses a one-vs-all approach for multi-class classification, it breaks into a series of binary classification tasks, with one hyperplane for each class.
- GC data is non-negative, only negative weights push toward the negative class, therefore positive weights are expected values, and the negative values are not.
- Both figures show most features have small weights, this suggests not all the 4800 features are needed to classify the fish data.



(a) Snapper

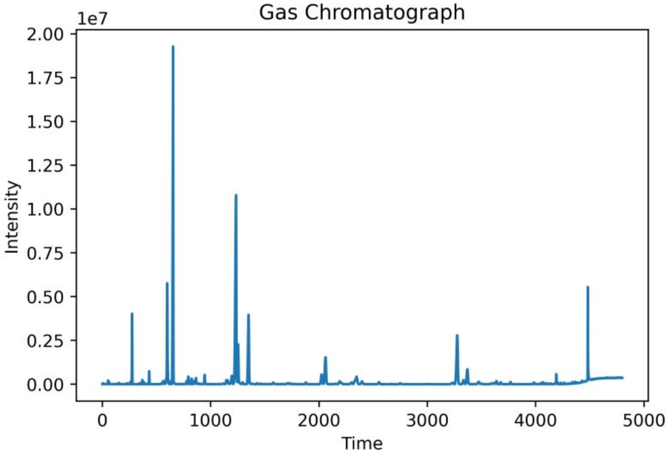
(b) Blue cod

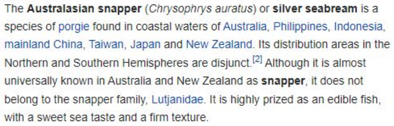
The **Australasian snapper** (*Chrysophrys auratus*) or **silver seabream** is a species of porgie found in coastal waters of Australia, Philippines, Indonesia, mainland China, Taiwan, Japan and New Zealand. Its distribution areas in the Northern and Southern Hemispheres are disjunct.[2] Although it is almost universally known in Australia and New Zealand as **snapper**, it does not belong to the snapper family, Lutjanidae. It is highly prized as an edible fish, with a sweet sea taste and a firm texture.

The **New Zealand blue cod** (***Parapercis colias***) is a temperate marine fish[3] of the family Pinguipedidae.[4] It is also known variously as **Boston blue cod**, **New Zealand cod**, **sand perch**, or its Māori names **rāwaru**, **pākirikiri** and **patutuki**.[5]

Feature Selection: Motivations

- High-dimensionality (4800 features).
- Redundant and correlated features hinder classification accuracy.
- Performance - less features improves compute and (potentially) performance.
- Interpretability - troubleshooting/diagnosis in a factory settings.

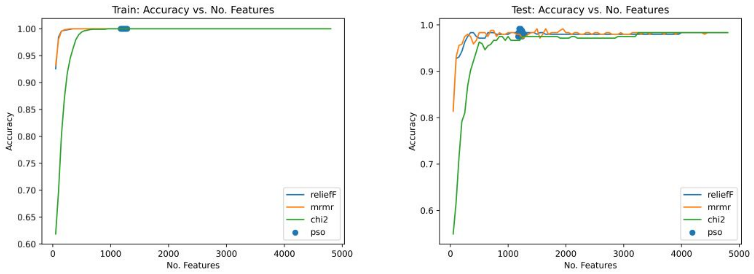




The **Australasian snapper** (*Chrysophrys auratus*) or **silver seabream** is a species of porgie found in coastal waters of Australia, Philippines, Indonesia, mainland China, Taiwan, Japan and New Zealand. Its distribution areas in the Northern and Southern Hemispheres are disjunct.[2] Although it is almost universally known in Australia and New Zealand as **snapper**, it does not belong to the snapper family, Lutjanidae. It is highly prized as an edible fish, with a sweet sea taste and a firm texture.

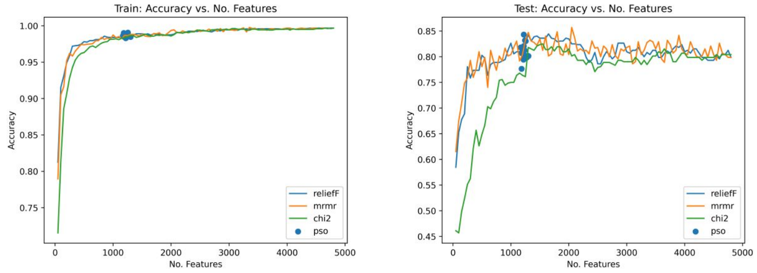
Feature Selection: Experiment

- Two datasets: fish species, body parts
- 4 feature selection methods:
  1. mRMR (Ding 2005)
  2. Relief-F
  3. Chi2 (Liu 1995)
  4. Wrapper PSO (Kennedy 1995, Nguyen 2020)
- 10-fold cross-validation, balanced accuracy.
- Run 1-3 for a range of k in [0,4800] in increments of 50.
- PSO automatically selects k, evaluate with 30 independent runs. PSO wrapped in SVM classifier.



(a) Species: Training set

(b) Species: Test set



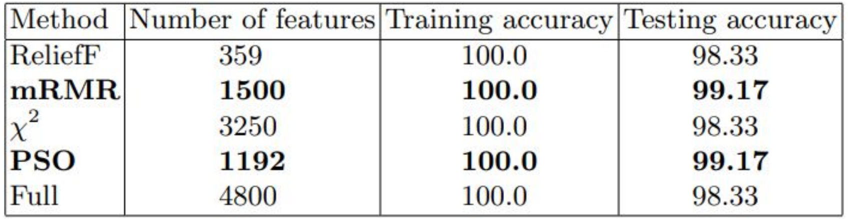
(a) Part: Training set

(b) Part: Test set

Feature Selection: Results

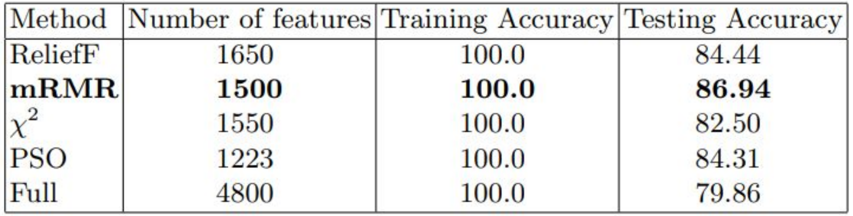
- In general, feature selection can significantly reduce the number of required packets/features while the classification performance is mostly maintained.
- For classifying the fish species, 75% of packets can be removed. For classifying the body parts, 60% of packets can be removed. Overall classification system can be up to 4 times faster.
- Classifying the body parts is harder than classifying fish species. Explains why body parts requires more features.
- PSO automatically selects k. PSO achieves good classification performance, except for some signs of overfitting on body parts.

Table 3: Best accuracy on Fish Species.



Method	Number of features	Training accuracy	Testing accuracy
ReliefF	359	100.0	98.33
mRMR	**1500**	**100.0**	**99.17**
χ^2	3250	100.0	98.33
PSO	**1192**	**100.0**	**99.17**
Full	4800	100.0	98.33

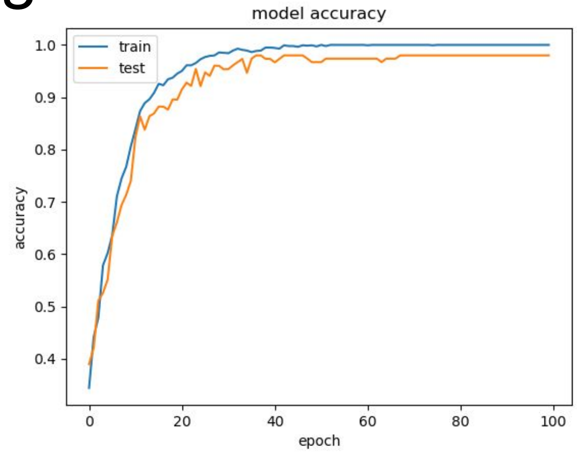
Table 4: Best Accuracy on Fish Body Parts

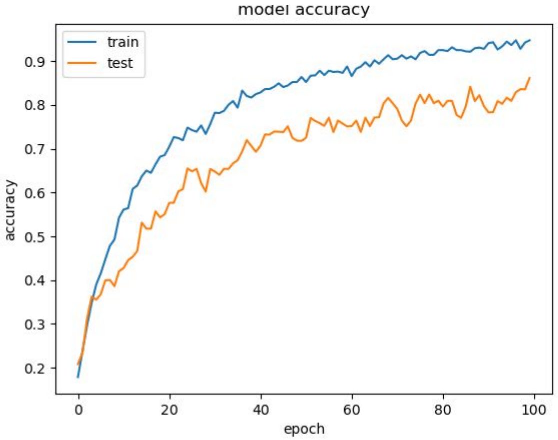


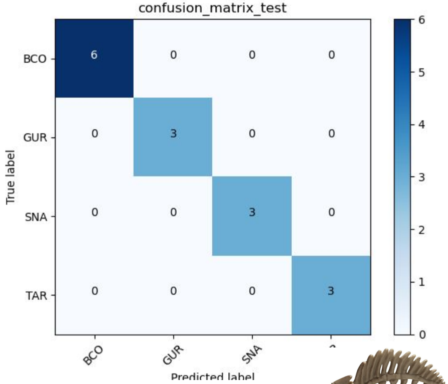
Method	Number of features	Training Accuracy	Testing Accuracy
ReliefF	1650	100.0	84.44
mRMR	**1500**	**100.0**	**86.94**
χ^2	1550	100.0	82.50
PSO	1223	100.0	84.31
Full	4800	100.0	79.86

Why not deep learning?

- Existing works, (Bi 2019, Matyushin 2020), use CNN for GC-MS data. Diffusion of innovation.
- 1D ConvNet for GC time-series data:
  - *Pooling* addresses instrumental drift / alignment in GC data. It looks for features in vicinity of a timestamp.
  - *LeakyRelu*, best activation function. Introduces nonlinearity while maintaining nice properties of a linear function (Goodfellow 2016).
  - *Dropout*, basic regularization technique, forces network into a bagged ensemble, where multiple subnetworks map a pattern (Goodfellow 2016).
- Limitations:
  - Low volume data (153 instances).
  - Trial and error.
  - Requires domain expertise in machine learning and chemistry (Black magic!)
- Future Work: Grid Search, Neural Architecture Search, EvoCNN (Wang 2018)

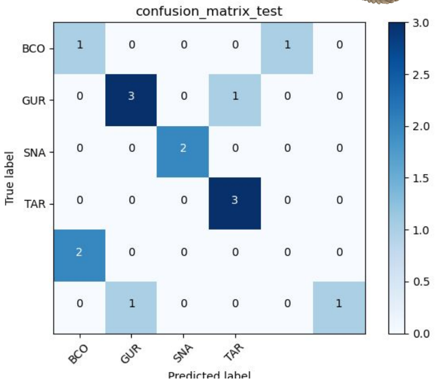






confusion_matrix_test

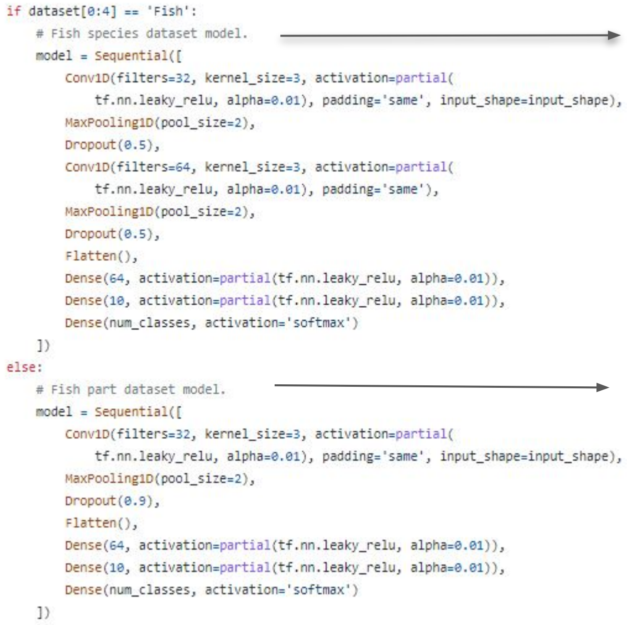
True label \ Predicted label	BCO	GUR	SNA	
BCO	6	0	0	0
GUR	0	3	0	0
SNA	0	0	3	0
TAR	0	0	0	3



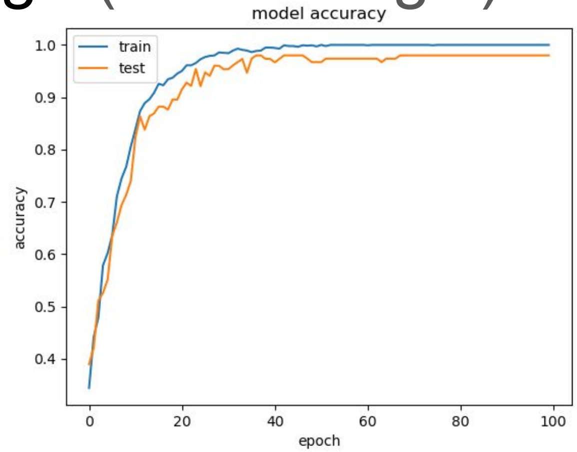
confusion_matrix_test

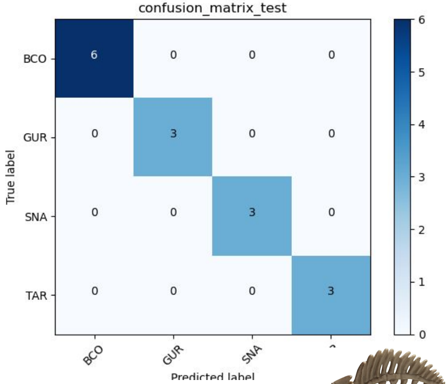
True label \ Predicted label	BCO	GUR	SNA	TAR		
BCO	1	0	0	0	1	0
GUR	0	3	0	1	0	0
SNA	0	0	2	0	0	0
TAR	0	0	0	3	0	0
	2	0	0	0	0	0
	0	1	0	0	0	1

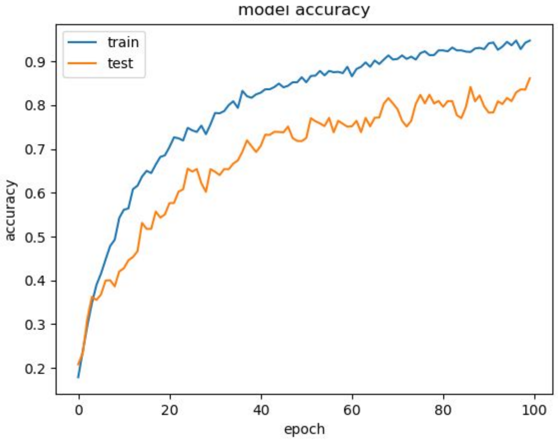
Why not deep learning? (Black magic)

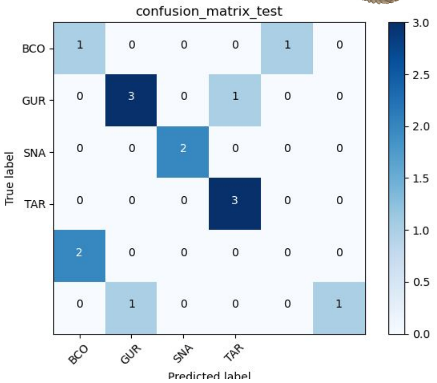


```
if dataset[0:4] == 'Fish':
    # Fish species dataset model.
    model = Sequential([
        Conv1D(filters=32, kernel_size=3, activation=partial(
            tf.nn.leaky_relu, alpha=0.01), padding='same', input_shape=input_shape),
        MaxPooling1D(pool_size=2),
        Dropout(0.5),
        Conv1D(filters=64, kernel_size=3, activation=partial(
            tf.nn.leaky_relu, alpha=0.01), padding='same'),
        MaxPooling1D(pool_size=2),
        Dropout(0.5),
        Flatten(),
        Dense(64, activation=partial(tf.nn.leaky_relu, alpha=0.01)),
        Dense(10, activation=partial(tf.nn.leaky_relu, alpha=0.01)),
        Dense(num_classes, activation='softmax')
    ])
else:
    # Fish part dataset model.
    model = Sequential([
        Conv1D(filters=32, kernel_size=3, activation=partial(
            tf.nn.leaky_relu, alpha=0.01), padding='same', input_shape=input_shape),
        MaxPooling1D(pool_size=2),
        Dropout(0.9),
        Flatten(),
        Dense(64, activation=partial(tf.nn.leaky_relu, alpha=0.01)),
        Dense(10, activation=partial(tf.nn.leaky_relu, alpha=0.01)),
        Dense(num_classes, activation='softmax')
    ])
```









Genetic Programming: Motivations

- Interpretable, can easily identify important features, interactions, and relate to domain expertise in chemistry.
  - Expression requires knowledge basic numeracy,
  - GP tree requires computer science
  - Both are covered by target audience of biochemists and ML practitioners.
- Embedded feature selection is embedded in the learning process.
  - Global search ability, to handle combinatorial explosion of possible feature subsets.
- Domain expertise in chemistry not required to find a good model.
- Parellisation - efficient computation on distributed computing systems.

Genetic Programming

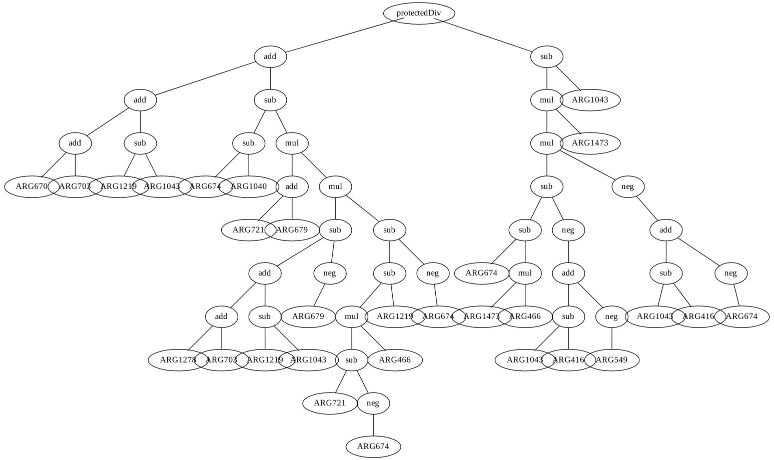
- Experimental Settings:
    - Operators: [+,-,*,/ (protected)], Terminals: 4800 features, random constant.
    - Population: 100, Mutation: 1%, Crossover: 95%, Generations: 30 (Grefenstette 1986)
    - Maximum depth: 8, due to memory limitations (Koza 1992).
    - Elitism: 10% of best individuals kept. Guarantees monotonous improvement.
- Two approaches:
    - Single Tree GP
    - Multi-tree GP

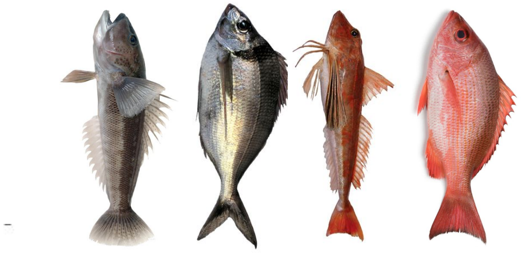
THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

https://xkcd.com/1838/

Genetic Programming: Single-Tree

- Single-Tree GP with classification map (Smart 2005)
- Minimize a single objective, the classification error.
- Classification map, class regions sequentially on floating point number line.
- 69% training accuracy on fish species (underfitting!).





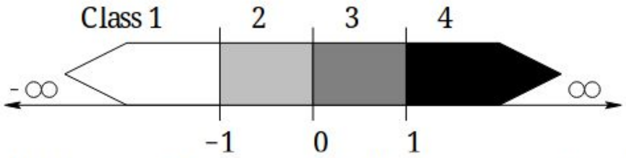
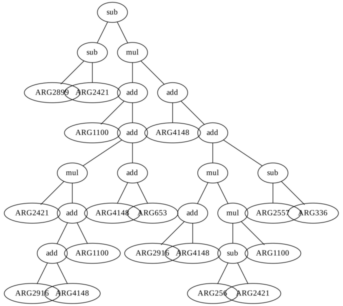


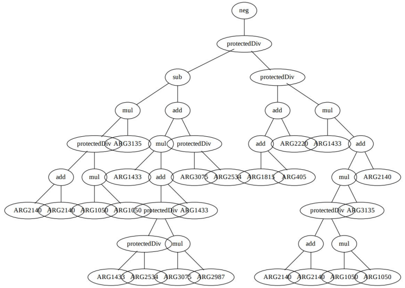
Figure 4.2: An example program classification map for four classes.

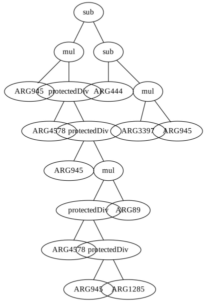
(Smart 2005)

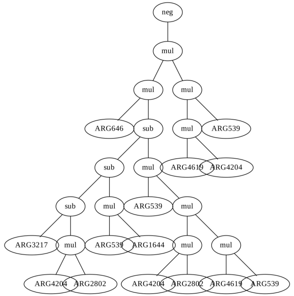
Genetic Programming: Multi-tree GP

- Multi-tree GP (Tran 2019), one-vs-all approach, similar SVM (Kennedy 1995).
- Each class has a classifier tree with one-vs-all prediction, balanced accuracy.
- Operators: crossover limited to subtree, mutation applied at random to ree.
- Maximize balanced accuracy, the aggregate of all classifiers.
- 72% training accuracy on fish species (better than Single-Tree GP!)



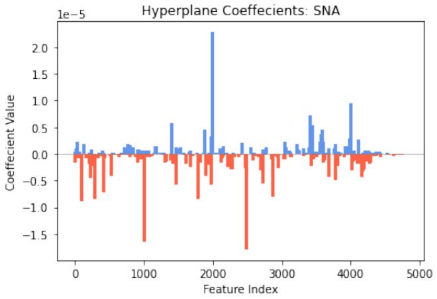




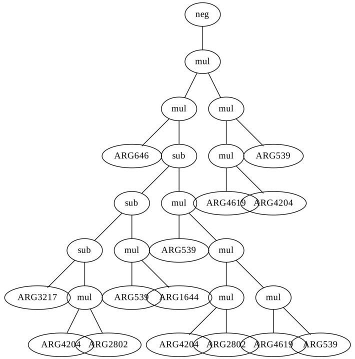


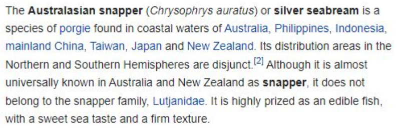
Genetic Programming: SVM Comparison

- Accuracy:
  - SVM better,
  - GP Tree worse.
- Features:
  - SVM - 150 features,
  - GP - 14 features,
    (max: 2^n=16 with max depth = n = 8)
- Tradeoff: Accuracy vs. Interpretability
- A simpler explanation, may lose some precision, but is easier to understand.
- Future work: Add distance measure (Tran 2019) as regularization term to maximize distance terms between different classes, and minimize distance between same class .



(a) Snapper





The **Australasian snapper** (*Chrysophrys auratus*) or **silver seabream** is a species of porgie found in coastal waters of Australia, Philippines, Indonesia, mainland China, Taiwan, Japan and New Zealand. Its distribution areas in the Northern and Southern Hemispheres are disjunct.[2] Although it is almost universally known in Australia and New Zealand as **snapper**, it does not belong to the snapper family, Lutjanidae. It is highly prized as an edible fish, with a sweet sea taste and a firm texture.

References

1. Eder, K. (1995). Gas chromatographic analysis of fatty acid methyl esters. Journal of Chromatography B: Biomedical Sciences and Applications, 671(1-2), 113-131.
2. Tomasi, G., Van Den Berg, F., & Andersson, C. (2004). 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.
3. Zhang, D., Huang, X., Regnier, F. E., & Zhang, M. (2008). Two-dimensional correlation optimized warping algorithm for aligning GC× GC− MS data. Analytical Chemistry, 80(8), 2664-2671.
4. Bi, K., Zhang, D., Qiu, T., & Huang, Y. (2019). GC-MS Fingerprints Profiling Using Machine Learning Models for Food Flavor Prediction. Processes, 8(1), 23.
5. Matyushin, D. D., & Buryak, A. K. (2020). Gas Chromatographic Retention Index Prediction Using Multimodal Machine Learning. Ieee Access, 8, 223140-223155.
6. Fix, E., & Hodges, J. L. (1989). Discriminatory analysis. Nonparametric discrimination: Consistency properties. International Statistical Review/Revue Internationale de Statistique, 57(3), 238-247.
7. Hand, D. J., & Yu, K. (2001). Idiot's Bayes—not so stupid after all?. International statistical review, 69(3), 385-398.
8. Ho, T. K. (1995, August). Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE.
9. Loh, W. Y. (2011). Classification and regression trees. Wiley interdisciplinary reviews: data mining and knowledge discovery,
10. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.
11. Liu, H., & Setiono, R. (1995, November). Chi2: Feature selection and discretization of numeric attributes. In Proceedings of 7th IEEE international conference on tools with artificial intelligence (pp. 388-391). IEEE.Liu, H., & Setiono, R. (1995, November). Chi2: Feature selection and discretization of numeric attributes. In Proceedings of 7th IEEE international conference on tools with artificial intelligence (pp. 388-391). IEEE.
12. Ding, C., & Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. Journal of bioinformatics and computational biology, 3(02), 185-205.
13. Robnik-Šikonja, M., & Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. Machine learning, 53(1), 23-69.
14. Eberhart, R., & Kennedy, J. (1995, November). Particle swarm optimization. In Proceedings of the IEEE international conference on neural networks (Vol. 4, pp. 1942-1948).
15. Nguyen, H. B., Xue, B., Andreae, P., & Zhang, M. (2017, June). Particle swarm optimisation with genetic operators for feature selection. In 2017 IEEE Congress on Evolutionary Computation (CEC) (pp. 286-293). IEEE.
16. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.
17. Wang, B., Sun, Y., Xue, B., & Zhang, M. (2018, July). Evolving deep convolutional neural networks by variable-length particle swarm optimization for image classification. In 2018 IEEE Congress on Evolutionary Computation (CEC) (pp. 1-8). IEEE.
18. Smart, W. R. (2005). Genetic programming for multiclass object classification. BSc (Honours) Research Project.
19. Tran, B., Xue, B., & Zhang, M. (2019). Genetic programming for multiple-feature construction on high-dimensional classification. Pattern Recognition, 93, 404-417.
20. Koza, J. R. G. P. (1992). On the programming of computers by means of natural selection. Genetic programming.