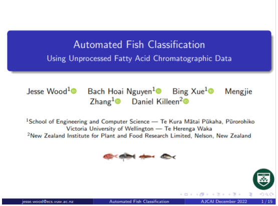
Introduction

- Hi, I'm Jesse from VUW
- Industry partnership with PFR
- Talk about fish oil, gas chromatography, and machine learning



Island Bay, Wellington, New Zealand

- Grandparents from Stromboli - a fishing village, volcanic island, in Italy.
- Migrated and settled in Island Bay, a quaint little fishing village where I was born and raised.
- Father is a cray fisherman.
- When I said I was taking up the fishing business, they were surprised that I meant postgraduate research.

PSO [1] inspired by social behaviour of animals

- A synchronicity to this research that is very apt.
- PSO (Kennedy 1995), an algorithm used in this paper, was invented trying to mimic social behaviour of animals.
- For example, the synchronous motion of school of fish in search of food or evading prey.
- We take an algorithm inspired by nature, and here we apply it to nature, *figuratively* releasing it back into the wild.

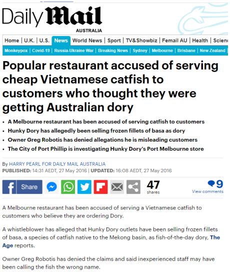
Topics

- These are the topics I will discuss today.

1. Catfishing
2. Fish Oil
3. Gas Chromatography
4. Classification
5. Intepretable
6. Feature Selection

Have you been catfished?

- (Pearl 2016) A story... A restaurant in Melbourne Australia with Australian Dory on the menu, but served catfish.
- Gut reaction, Aussies... No surprises there!



Daily Mail AUSTRALIA

Home | U.K. | U.S. | News | World News | Sport | TV&Showbiz | Femail AU | Health | Scien

Monkeypox | Covid-19 | Russia-Ukraine War | Breaking News | Sydney | Melbourne | Brisbane | New Zealand |

Popular restaurant accused of serving cheap Vietnamese catfish to customers who thought they were getting Australian dory

- **A Melbourne restaurant has been accused of serving catfish to customers**
- **Hunky Dory has allegedly been selling frozen fillets of basa as dory**
- **Owner Greg Robotis has denied allegations he is misleading customers**
- **The City of Port Phillip is investigating Hunky Dory's Port Melbourne store**

By HARRY PEARL FOR DAILY MAIL AUSTRALIA
PUBLISHED: 14:31 AEDT, 27 May 2016 | **UPDATED:** 16:08 AEDT, 27 May 2016

Share 47 shares 9 View comments

A Melbourne restaurant has been accused of serving a Vietnamese catfish to customers who believe they are ordering Dory.

A whistleblower has alleged that Hunky Dory outlets have been selling frozen fillets of basa, a species of catfish native to the Mekong basin, as fish-of-the-day dory, **The Age** reports.

Owner Greg Robotis has denied the claims and said inexperienced staff may have been calling the fish the wrong name.

Catfishing [2], Mislabelling [3], and Quality Assurance [4]

- Unfortunately, it's an international problem, a 2016 meta-analysis (Pardo 2016) of seafood industry found average mislabelling rate of 30%.
- Many steps supply chain from ocean-to-plate is prone to human error and criminal activity (like catfishing).
- Confident we know what we are eating, labels on seafood products are accurate.
- Need tools for quality assurance.

Nutrition Facts

6 servings per container
Serving size 4-5 ounces(187g)

Amount per serving
Calories 200

	% Daily Value*
Total Fat 5g	**6%**
Saturated Fat 0.5g	**3%**
Trans Fat 0g	
Cholesterol 80mg	**27%**
Sodium 610mg	**27%**
Total Carbohydrate 10g	**4%**
Dietary Fiber 0g	**0%**
Total Sugars 3g	
Includes 0g Added Sugars	**0%**
Protein 27g	
Vitamin D 2mcg	10%
Calcium 79mg	6%
Iron 3mg	15%
Potassium 519mg	10%

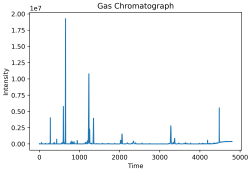
*The % Daily Value tells you how much a nutrient in a serving of food contributes to a daily diet. 2,000 calories a day is used for general nutrition advice.

Fish Oil is Brain Food! [5, 6]

- Fish oil contains nutritious fatty acid (Simopoulous 2011), Omega 3.
- Bodies don't produce it naturally, often missing from western diets (Panse 2016), high demand for Omega-3 supplements.
- To reduce waste in fish processing, need tools to identify fish species and body parts suitable for use in Omega-3 supplements.

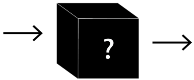
Fish oil analyzed with Gas Chromatography!

- Gas chromatography is an analytical chemistry technique
- Use to identify fish species (and body parts) suitable for use in Omega-3 supplements.
- Prepare/analyze Gas Chromatography is time-consuming and expensive as takes several hours and requires domain expertise, i.e. chemists (Black 2019).



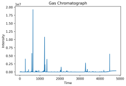
Fish oil analysis can't be blackbox! [7, 8]

- Previous works (Bi 2020, Matyushin 2020) employed blackbox CNNs to analysis of Gas Chromatography.
- We need models we can understand, and build trust in their predictions,
- For QA, need ability to troubleshoot/verify model,
- in order to deploy in real-world factory settings.



Gas Chromatography [4] ≈ Chemical Fingerprint

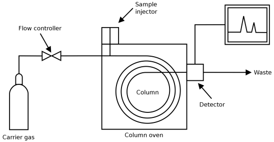
- Gas chromatography is a chemical fingerprint that tells us what something is made of.
- It produces high-dimensional low-sample size data
- (due to the cost of acquiring data).



Gas Chromatography: Steps

- Gist: separate molecules based on their distinct boiling temperatures.
- Each peak represents a known molecule.
- Can match these timestamps and peaks to reference samples to find out what something is made from.

1. Heat
2. Evaporate
3. Tube
4. Detector



Classification: Datasets

Dataset
Species 
Parts 

Classification: Methods

Dataset	**Method**
	KNN [9]
	RF [10]
Species 	DT [11]
Parts 	NB [12]
	SVM [13]

Classification: Balanced Accuracy, Cross-validation

Dataset	Method	**Train**	**Test**
	KNN [9]	83.57	74.88
	RF [10]	100.0	85.65
Species 	DT [11]	100.0	76.98
	NB [12]	79.54	75.27
	SVM [13]	100.0	98.33
	KNN	68.95	43.61
	RF	100.00	72.60
Parts 	DT	100.00	60.14
	NB	65.54	48.61
	SVM	100.00	79.86

Classification: Results

Dataset	Method	Train	Test
	KNN [9]	83.57	74.88
	RF [10]	100.0	85.65
Species 	DT [11]	100.0	76.98
	NB [12]	79.54	75.27
	SVM [13]	100.0	98.33
	KNN	68.95	43.61
	RF	100.00	72.60
Parts 	DT	100.00	60.14
	NB	65.54	48.61
	SVM	100.00	79.86

Classification: SVM near-perfect on fish species

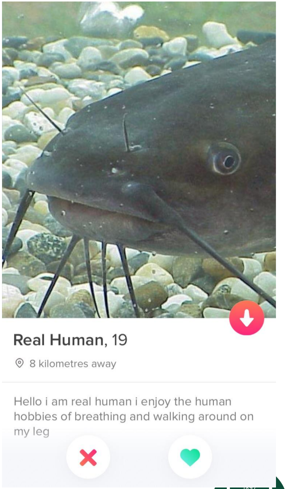
Dataset	Method	Train	Test
	KNN [9]	83.57	74.88
	RF [10]	100.0	85.65
Species 	DT [11]	100.0	76.98
	NB [12]	79.54	75.27
	SVM [13]	**100.0**	**98.33**
	KNN	68.95	43.61
	RF	100.00	72.60
Parts 	DT	100.00	60.14
	NB	65.54	48.61
	SVM	100.00	79.86

Classification: Body parts harder than fish species

Dataset	Method	Train	Test
Species 	KNN [9]	83.57	74.88
	RF [10]	100.0	85.65
	DT [11]	100.0	76.98
	NB [12]	79.54	75.27
	SVM [13]	**100.0**	**98.33**
Parts 	KNN	68.95	43.61
	RF	100.00	72.60
	DT	100.00	60.14
	NB	65.54	48.61
	SVM	**100.00**	**79.86**

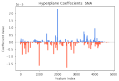
Classification: Avoid Catfishing [2] & Mislabelling [3]

- Avoid getting catfished (Pearl 2016)
- (if you carry a gas chromatograph in your pockets/purse to restaurants)
- Accurate and interpretable model to determine fish species and avoid mislabelling (Pardo 2016) for QA.
- Identify high-value fish oil for use in Omega-3. supplements.



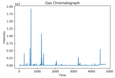
Intepretable Model - A Hyperplane

- Linear SVM uses a one-vs-rest approach, where it constructs a hyperplane coefficient for each class.
- Here we give hyperplane coefficient for snapper.
- Positive weights (in blue), molecules at these timestamps push classifier towards Snapper class.
- Negative weights (in red), molecules at these timestamps push classifier away from Snapper class.



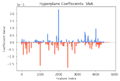
Interpretable Instance - A Chromatograph

- Hyperplane coefficient $\approx$ weight vector,
- Instance is a gas chromatograph.
- Dot product as a loose measure of the similarity of two vectors (Grokking 2016).
- Dot product of weight vector and instance, hyperplane and gas chromatograph, predicts which class the instance belongs to.

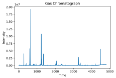


Intepretable Comparison - Hyperplane vs. Chromatograph

- With domain expertise we can see if the model learned semantically meaningful features.
- semantically meaningful $\approx$ known molecules and timestamps expected in that species or body part.
- Post-hoc analysis of the model to build trust in its predictions.
- Unlike blackbox models, it can be trusted, verified for use in factory setting.



post hoc analysis to build trust in the prediction 



Feature Selection: Dataset

Dataset
Species 
Parts 

Feature Selection: Methods

Dataset	**Method**
	ReliefF [14]
Species 	mRMR [15]
	χ^2 [16]
Parts 	PSO [1]
	Full

Feature Selection: # Features given for Best Run

Dataset	Method	**# Features**
Species 	ReliefF [14]	359
	mRMR [15]	1500
	χ^2 [16]	3250
	PSO [1]	1192
	Full	4800
Parts 	ReliefF	1650
	mRMR	1500
	χ^2	1550
	PSO	1223
	Full	4800

Feature Selection: Balanced Accuracy, Cross-validation

Dataset	Method	# Features	**Train**	**Test**
Species 	ReliefF [14]	359	100.0	98.33
	mRMR [15]	1500	100.0	99.17
	χ^2 [16]	3250	100.0	98.33
	PSO [1]	1192	100.0	99.17
	Full	4800	100.0	98.33
Parts 	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

Feature Selection: Results

Dataset	Method	# Features	Train	Test
	ReliefF [14]	359	100.0	98.33
	mRMR [15]	1500	100.0	99.17
Species 	χ^2 [16]	3250	100.0	98.33
	PSO [1]	1192	100.0	99.17
	Full	4800	100.0	98.33
	ReliefF	1650	100.0	84.44
	mRMR	1500	100.0	86.94
Parts 	χ^2	1550	100.0	82.50
	PSO	1223	100.0	84.31
	Full	4800	100.0	79.86

Feature Selection: PSO & MRMR improve accuracy!

Dataset	Method	# Features	Train	Test
	ReliefF [14]	359	100.0	98.33
	mRMR [15]	**1500**	**100.0**	**99.17**
Species 	χ^2 [16]	3250	100.0	98.33
	PSO [1]	**1192**	**100.0**	**99.17**
	Full	4800	100.0	98.33
	ReliefF	1650	100.0	84.44
	mRMR	**1500**	**100.0**	**86.94**
Parts 	χ^2	1550	100.0	82.50
	PSO	1223	100.0	84.31
	Full	4800	100.0	79.86

Feature Selection: PSO uses 1/4 features, x4 faster!

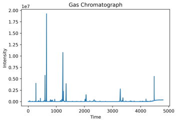
Dataset	Method	**# Features**	Train	Test
	ReliefF [14]	359	100.0	98.33
	mRMR [15]	**1500**	**100.0**	**99.17**
Species 	χ^2 [16]	3250	100.0	98.33
	PSO [1]	**1192**	**100.0**	**99.17**
	Full	4800	100.0	98.33
	ReliefF	1650	100.0	84.44
	mRMR	**1500**	**100.0**	**86.94**
Parts 	χ^2	1550	100.0	82.50
	PSO	1223	100.0	84.31
	Full	4800	100.0	79.86

Feature Selection: MRMR best for body parts!

Dataset	Method	# Features	Train	Test
Species 	ReliefF [14]	359	100.0	98.33
	mRMR [15]	**1500**	**100.0**	**99.17**
	χ^2 [16]	3250	100.0	98.33
	PSO [1]	**1192**	**100.0**	**99.17**
	Full	4800	100.0	98.33
Parts 	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

Feature Selection: Reduce GC time [4], simpler models [17]

- Reduce time taken for Gas Chromatography (Eder 1995).
- Stop early once important timestamps have been analyzed.
- Less features $\approx$ simpler model, less moving parts,
- Reduce dimensionality, easier to understand (Zhao 2019), build trust in predictions.



TLDR;

Linear SVM can accurately predict fish species, **PSO** makes that process 4 times faster, producing an **accurate**, **interpretable** and **efficient** model for **Gas Chromatography**.

Download the slides, paper, poster.

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

[2] H. P. F. D. M. Australia, "Melbourne restaurant hunky dory accused of serving catfish to customers instead of dory," May 2016. [Online]. Available: https://www.dailymail.co.uk/news/article-3611999/ Melbourne-restaurant-Hunky-Dory-accused-serving-catfish-customers-ins html

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

[4] K. Eder, "Gas chromatographic analysis of fatty acid methyl esters," *Journal of Chromatography B: Biomedical Sciences and Applications*, vol. 671, no. 1-2, pp. 113–131, 1995.

[5] A. P. Simopoulos, "Evolutionary aspects of diet: the omega-6/omega-3 ratio and the brain," *Molecular neurobiology*, vol. 44, no. 2, pp. 203–215, 2011.

[6] M. L. Panse and S. D. Phalke, "World market of omega-3 fatty acids," *Omega-3 Fatty Acids*, pp. 79–88, 2016.

[7] K. Bi, D. Zhang, T. Qiu, and Y. Huang, "Gc-ms fingerprints profiling using machine learning models for food flavor prediction," *Processes*, vol. 8, no. 1, p. 23, 2020.

[8] D. D. Matyushin and A. K. Buryak, "Gas chromatographic retention index prediction using multimodal machine learning," *Ieee Access*, vol. 8, pp. 223 140–223 155, 2020.

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

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

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

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

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

[14] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of relieff and rrelieff," *Machine learning*, vol. 53, no. 1, pp. 23–69, 2003.

[15] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of bioinformatics and computational biology*, vol. 3, no. 02, pp. 185–205, 2005.

[16] H. Liu and R. Setiono, "Chi2: Feature selection and discretization of numeric attributes," in *Proceedings of 7th IEEE International*

Conference on Tools with Artificial Intelligence. IEEE, 1995, pp. 388–391.

[17] Z. Zhao, R. Anand, and M. Wang, "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)*. IEEE, 2019, pp. 442–452.