# Automated Fish Classification

## Using Unprocessed Fatty Acid Chromatographic Data

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

[1]School of Engineering and Computer Science — Te Kura Mātai Pūkaha, Pūrorohiko
Victoria University of Wellington — Te Herenga Waka
[2]New Zealand Institute for Plant and Food Research Limited, Nelson, New Zealand

# Island Bay, Wellington, New Zealand

# PSO [1] inspired by social behaviour of animals

# Topics

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

# Have you been catfished? [2]

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.

Aussies! No surprises there...

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

**Nutrition Facts**

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

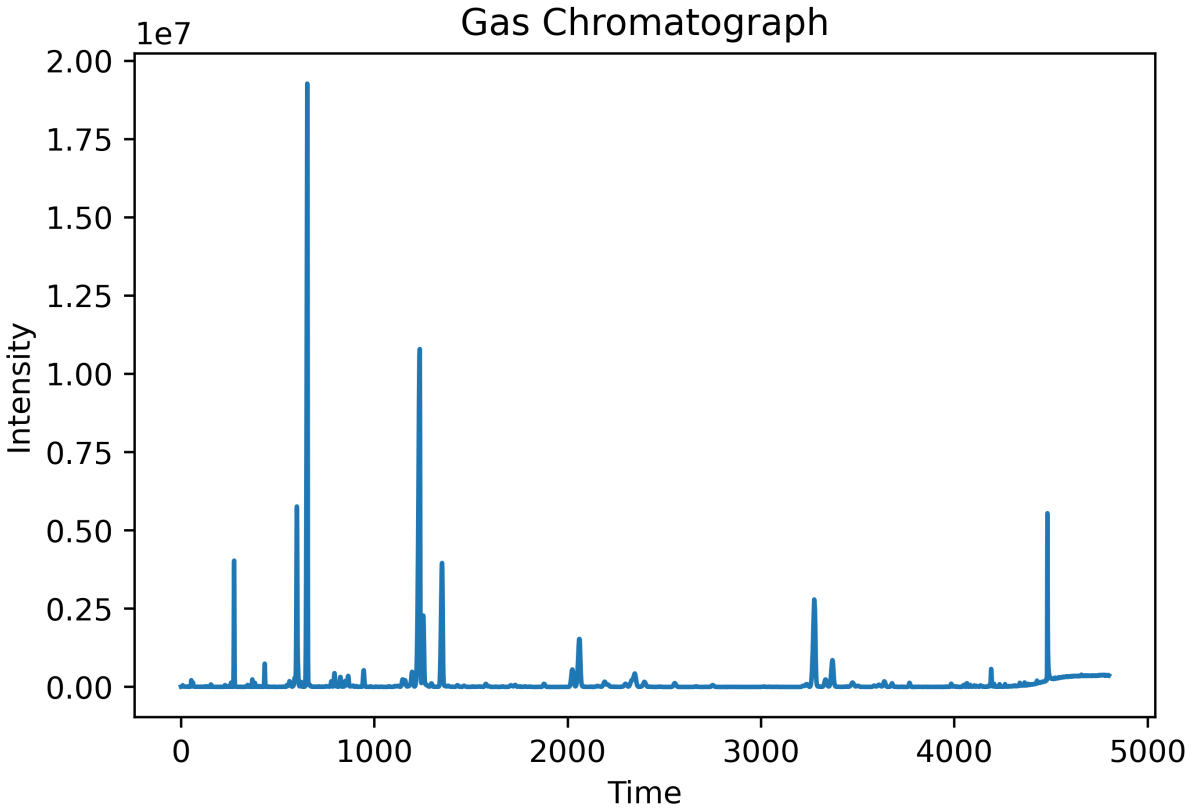
**Amount per serving**
**Calories** **200**

| | % 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% |

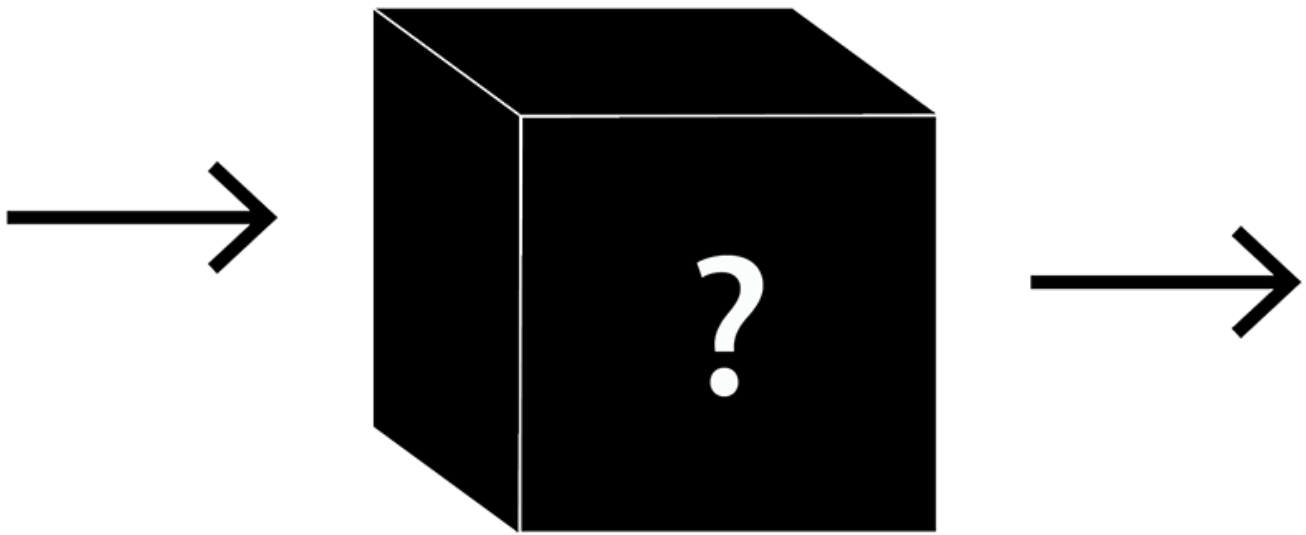
*The % Daily Value tells you how much a nutrient in a serving of food contributes to a daily diet. 2,000 calories a day is used for general nutrition advice.

# Fish oil is brain food! [5, 6]

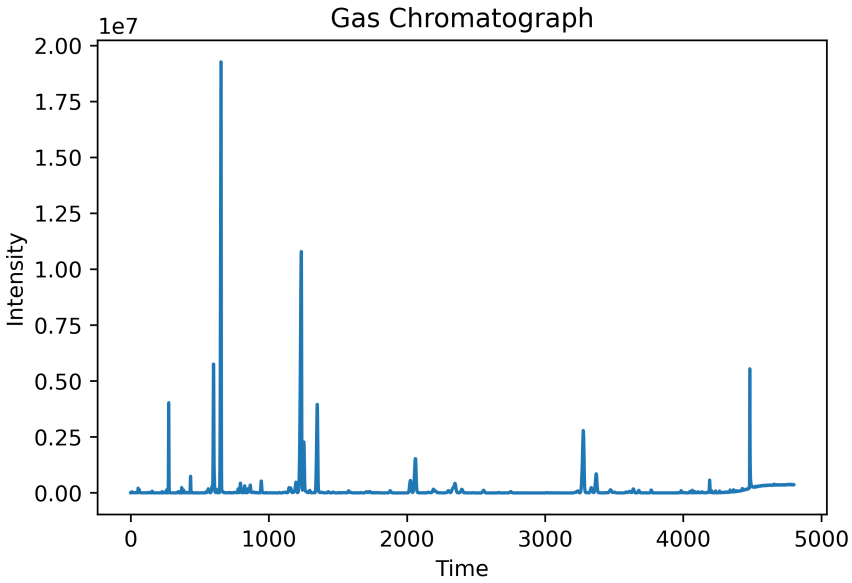
# Fish oil analyzed with Gas Chromatography! [7]



# Fish oil analysis can't be blackbox! [8, 9]

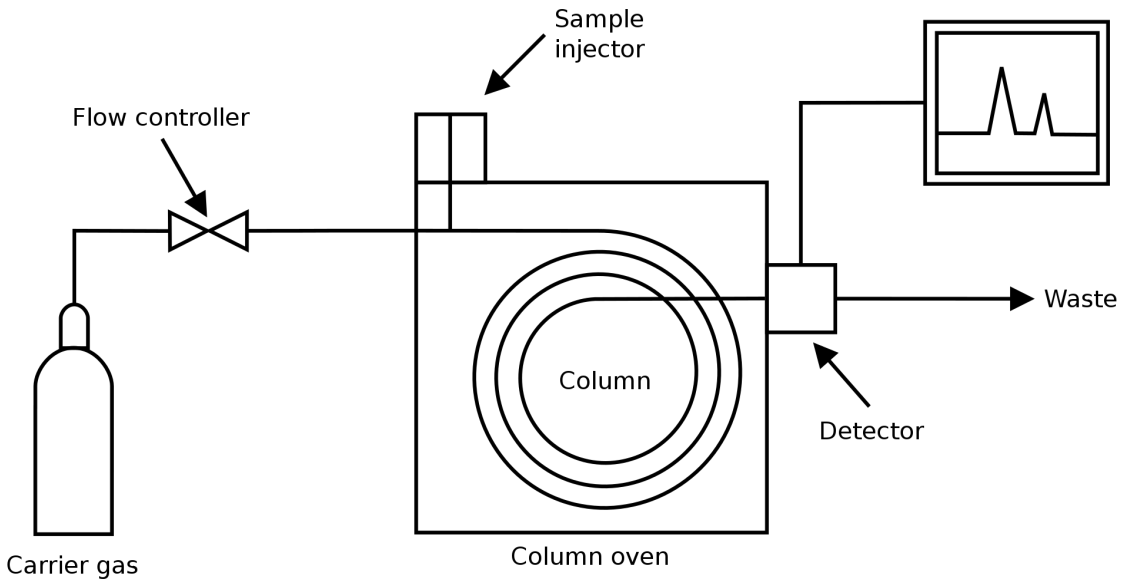


# Gas Chromatrography [4] ≈ Chemical Fingerprint



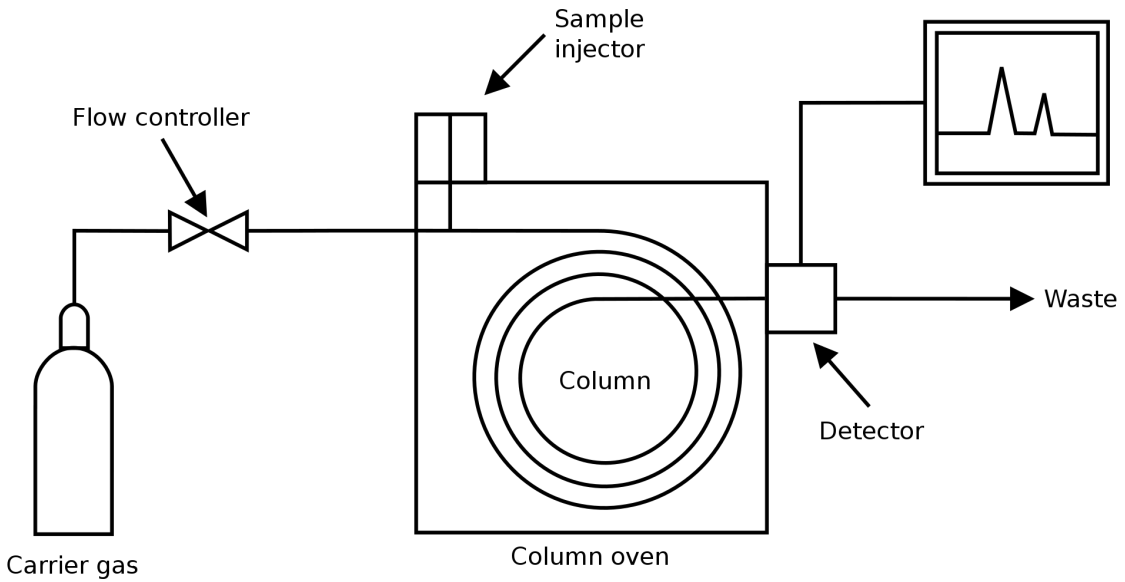

# Gas Chromatrography: Steps

1. Apply heat to liquid.
2. Evaporate into gas.
3. Travel through long tube.
4. Detector measures intensity.



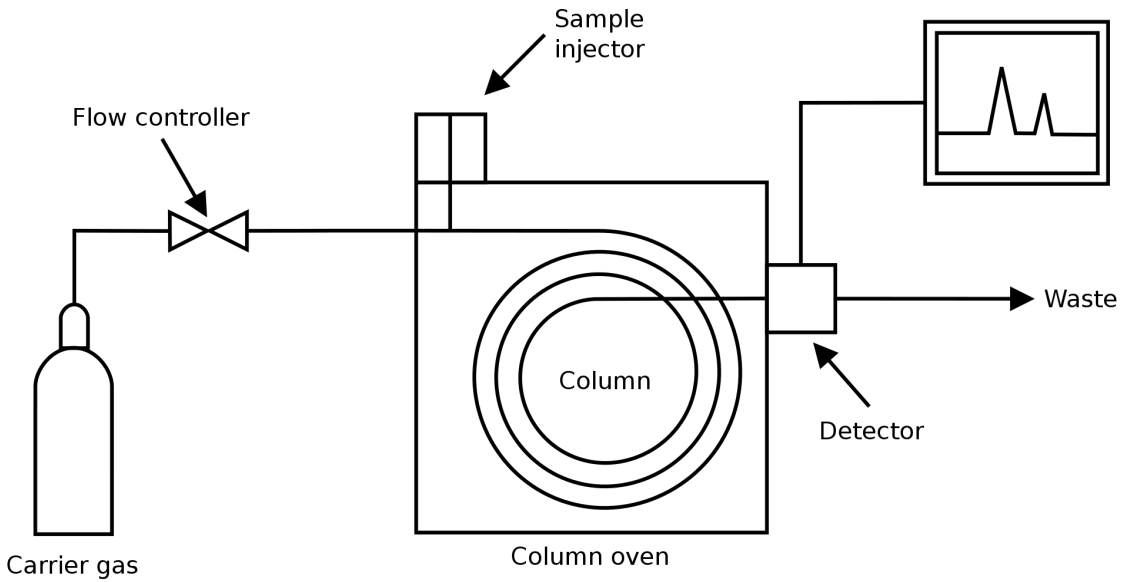

# Gas Chromatrography: Steps

1. Apply heat to liquid.
2. Evaporate into gas.
3. Travel through long tube.
4. Detector measures intensity.



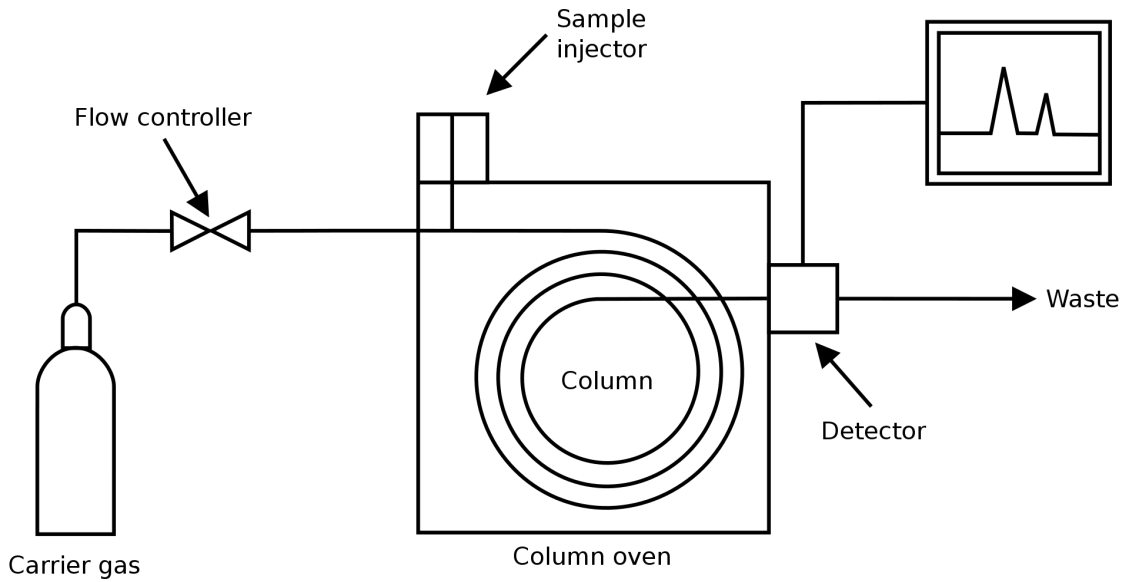
# Gas Chromatrography: Steps

1. Apply heat to liquid.
2. Evaporate into gas.
3. Travel through long tube.
4. Detector measures intensity.



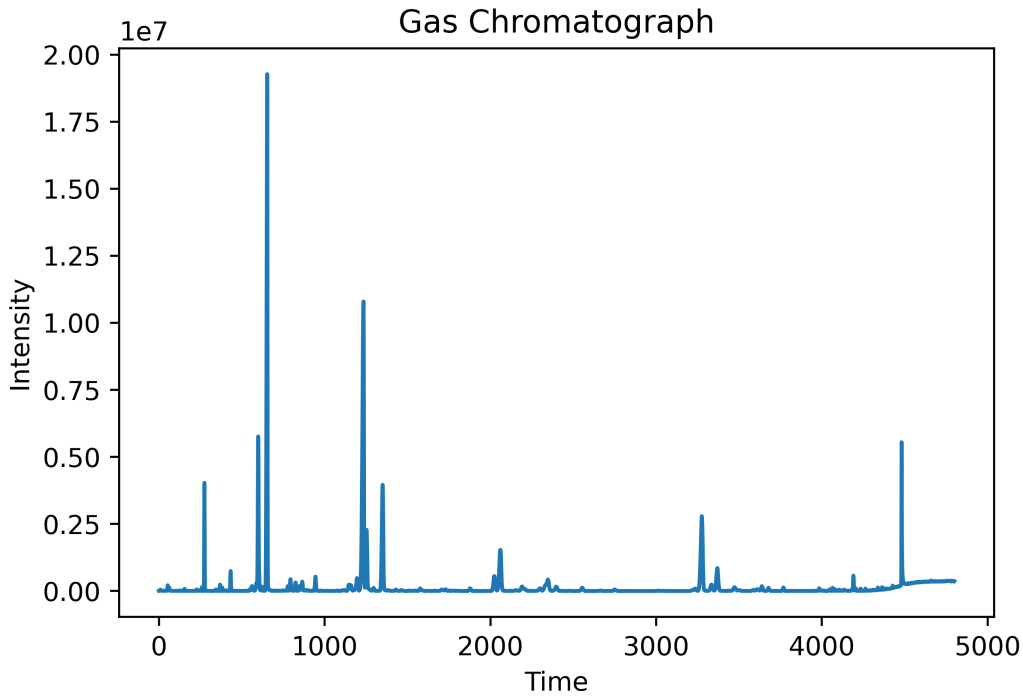
# Gas Chromatrography: Steps

1. Apply heat to liquid.
2. Evaporate into gas.
3. Travel through long tube.
4. Detector measures intensity.



# Gas Chromatrography: Steps

1. Apply heat to liquid.
2. Evaporate into gas.
3. Travel through long tube.
4. Detector measures intensity.



# Classification: Datasets

**Dataset**

Species 

Parts 

# Classification: Methods

| Dataset | **Method** |
|---|---|
| Species  | KNN [10] |
| Parts  | RF [11] |
| | DT [12] |
| | NB [13] |
| | SVM [14] |

# Classification: Balanced Accuracy, Cross-validation

| Dataset | Method | **Train** | **Test** |
|---|---|---|---|
| Species  | KNN [10] | 83.57 | 74.88 |
| | RF [11] | 100.0 | 85.65 |
| | DT [12] | 100.0 | 76.98 |
| | NB [13] | 79.54 | 75.27 |
| | SVM [14] | 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: Results

| Dataset | Method | Train | Test |
|---|---|---|---|
| Species  | KNN [10] | 83.57 | 74.88 |
| | RF [11] | 100.0 | 85.65 |
| | DT [12] | 100.0 | 76.98 |
| | NB [13] | 79.54 | 75.27 |
| | SVM [14] | 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: SVM near-perfect on fish species

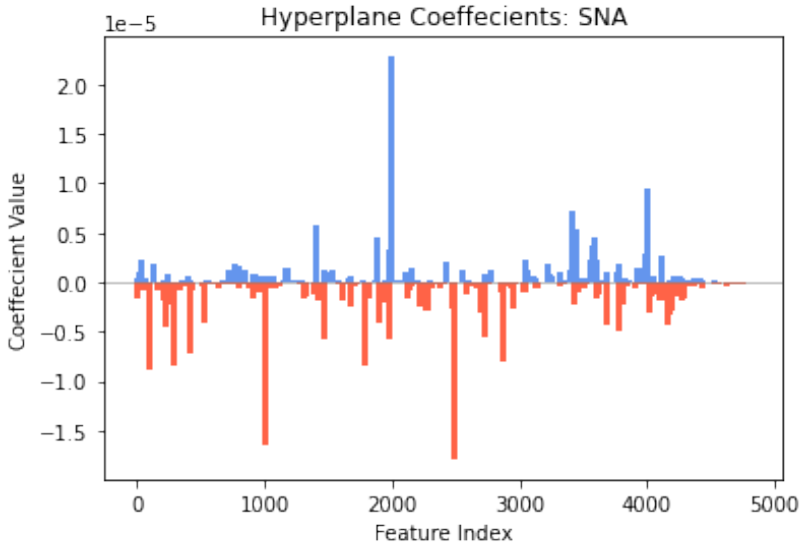
| Dataset | Method | Train | Test |
|---|---|---|---|
| Species  | KNN [10] | 83.57 | 74.88 |
| | RF [11] | 100.0 | 85.65 |
| | DT [12] | 100.0 | 76.98 |
| | NB [13] | 79.54 | 75.27 |
| | **SVM** [14] | **100.0** | **98.33** |
| Parts  | KNN | 68.95 | 43.61 |
| | RF | 100.00 | 72.60 |
| | DT | 100.00 | 60.14 |
| | NB | 65.54 | 48.61 |
| | SVM | 100.00 | 79.86 |

# Classification: Body parts harder than fish species

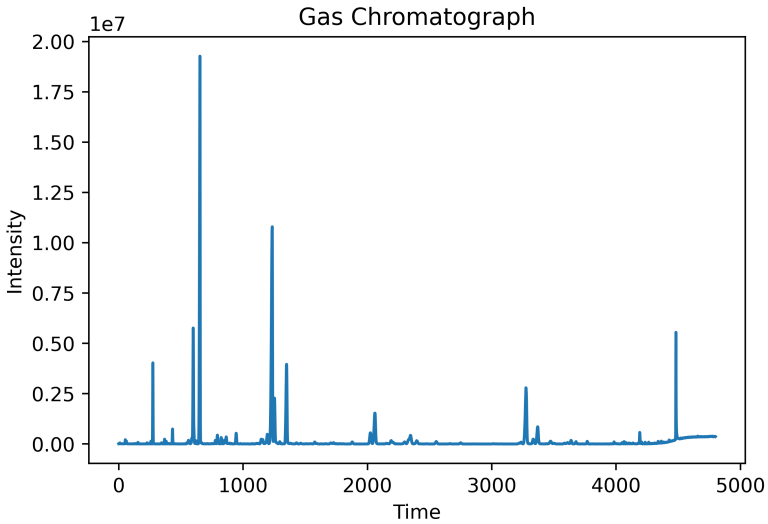
| Dataset | Method | Train | Test |
|---|---|---|---|
| Species  | KNN [10] | 83.57 | 74.88 |
| | RF [11] | 100.0 | 85.65 |
| | DT [12] | 100.0 | 76.98 |
| | NB [13] | 79.54 | 75.27 |
| | **SVM** [14] | **100.0** | **98.33** |
| Parts  | KNN | 68.95 | 43.61 |
| | RF | 100.00 | 72.60 |
| | DT | 100.00 | 60.14 |
| | NB | 65.54 | 48.61 |
| | **SVM** | **100.00** | **79.86** |

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

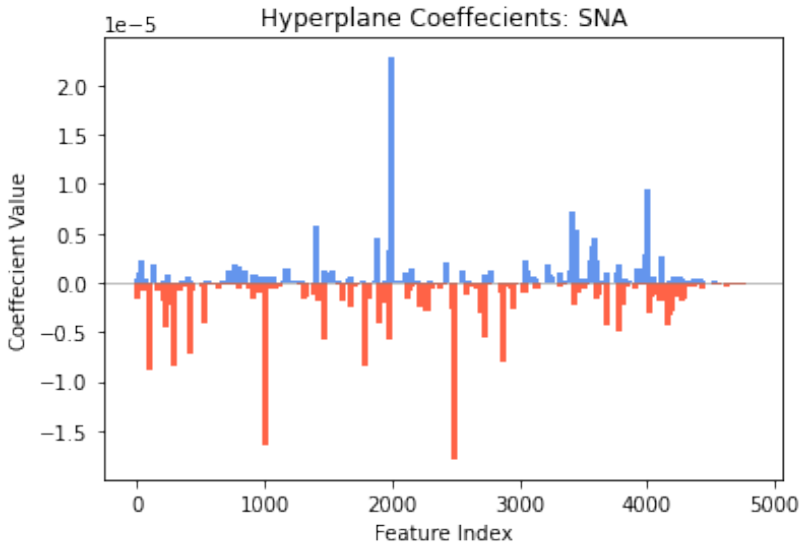
# Intepretable Model - A Hyperplane



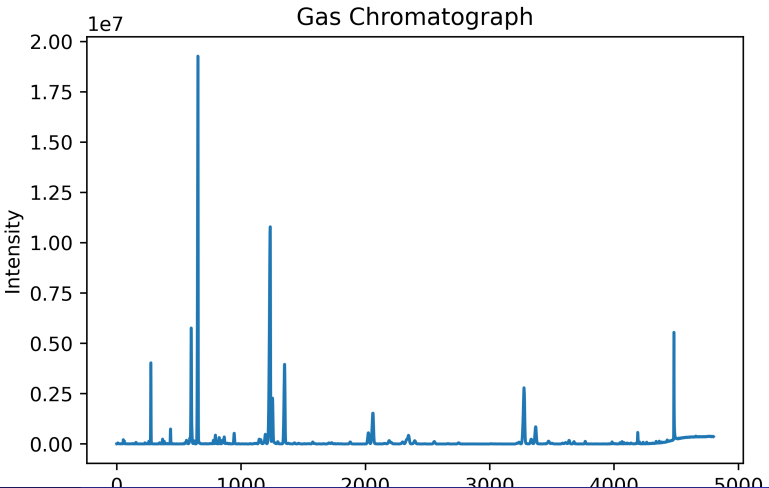

# Interpretable Instance - A Chromatograph




# Intepretable Comparison - Hyperplane vs. Chromatograph




post hoc analysis to build trust in the prediction 



# Feature Selection: Dataset

**Dataset**

Species 

Parts 

# Feature Selection: Methods

| Dataset | **Method** |
|---|---|
| Species  | ReliefF [15] |
| | mRMR [16] |
| Parts  | $\chi^2$ [17] |
| | PSO [1] |
| | Full |

# Feature Selection: # Features given for Best Run

| Dataset | Method | **# Features** |
|---|---|---|
| Species  | ReliefF [15] | 359 |
| | mRMR [16] | 1500 |
| | $\chi^2$ [17] | 3250 |
| | PSO [1] | 1192 |
| | Full | 4800 |
| Parts  | ReliefF | 1650 |
| | mRMR | 1500 |
| | $\chi^2$ | 1550 |
| | PSO | 1223 |
| | Full | 4800 |

# Feature Selection: Balanced Accuracy, Cross-validation

| Dataset | Method | # Features | **Train** | **Test** |
|---|---|---|---|---|
| Species  | ReliefF [15] | 359 | 100.0 | 98.33 |
| | mRMR [16] | 1500 | 100.0 | 99.17 |
| | $\chi^2$ [17] | 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 |
| | $\chi^2$ | 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 |
|---|---|---|---|---|
| Species  | ReliefF [15] | 359 | 100.0 | 98.33 |
| | mRMR [16] | 1500 | 100.0 | 99.17 |
| | $\chi^2$ [17] | 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 |
| | $\chi^2$ | 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 |
|---|---|---|---|---|
| Species  | ReliefF [15] | 359 | 100.0 | 98.33 |
| | **mRMR** [16] | **1500** | **100.0** | **99.17** |
| | $\chi^2$ [17] | 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** |
| | $\chi^2$ | 1550 | 100.0 | 82.50 |
| | PSO | 1223 | 100.0 | 84.31 |
| | Full | 4800 | 100.0 | 79.86 |

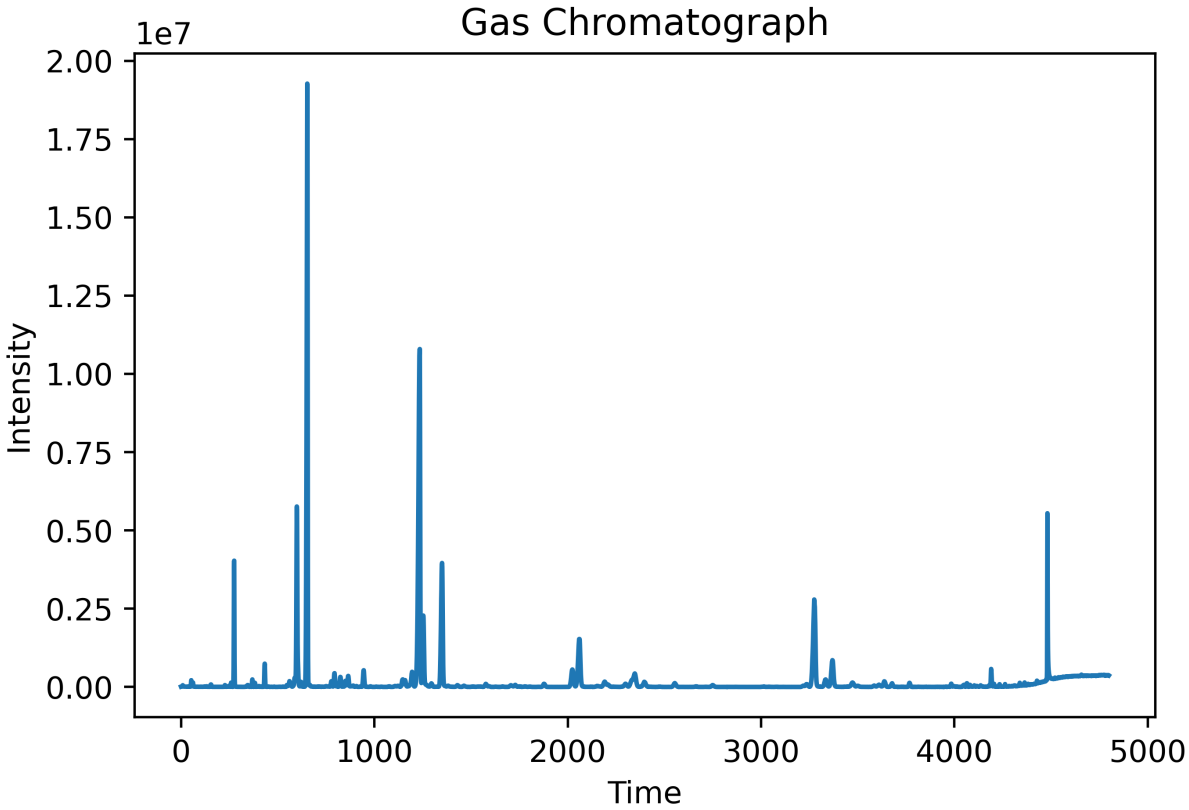
# Feature Selection: PSO uses 1/4 features, x4 faster!

| Dataset | Method | # **Features** | Train | Test |
|---|---|---|---|---|
| Species  | ReliefF [15] | 359 | 100.0 | 98.33 |
| | **mRMR** [16] | **1500** | **100.0** | **99.17** |
| | $\chi^2$ [17] | 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** |
| | $\chi^2$ | 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 [15] | 359 | 100.0 | 98.33 |
| | **mRMR** [16] | **1500** | **100.0** | **99.17** |
| | $\chi^2$ [17] | 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** |
| | $\chi^2$ | 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 [18]




# 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] C. Black, O. P. Chevallier, K. M. Cooper, S. A. Haughey, J. Balog, Z. Takats, C. T. Elliott, and C. Cavin, "Rapid detection and specific identification of offals within minced beef samples utilising ambient mass spectrometry," *Scientific reports*, vol. 9, no. 1, pp. 1–9, 2019.

[8] 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.

[9] 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.

[10] 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.

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

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

[13] 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.

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

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

[16] 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.

[17] 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.

[18] 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.