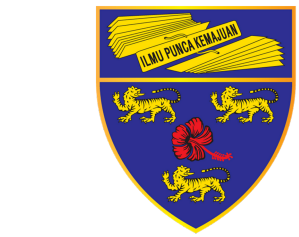

UNIVERSITY OF MALAYA

The Leader in Research & Innovation

Explainable Machine Learning Classification of Forensically Important Flies using Wing Venation Data

Ling Min Hao

Supervisor:
Assoc Prof. Dr Khang Tsung Fei

September 2021

Contents

1 Introduction **2**

2 Literature Review **3**

3 Significance of Research **4**

4 Methodology **5**

5 Work Schedule **6**

References **7**

1 Introduction

Classification of species is an important task for taxonomists so that scientists can better communicate biological information to solve real-life problems. For instance, two important families of flies (Calliphoridae and Sarcophagidae) are forensically important and they are usually used as an indicator to predict the minimum post-mortem interval in forensic studies [1]. The advancement of various machine learning technologies greatly aids taxonomists to classify the species in these two families. Using random forest model [2, 3] on geometric morphometric data generated using geometric morphometric analysis [4], Khang et al. [5] gave a promising proof of concept towards an accurate classification of species. Next, Goh & Khang (2021) [6] showed that whole images of wing venation patterns in Calliphoridae and Sarcophagidae provide better prediction quality as compared to using geometric morphometric image data. In fact, prediction via this approach is extremely accurate. However, for the results of the machine learning prediction to be contextually valuable (i.e. explainable) to the taxonomists, the region where the machine learning model learns to discriminate between different species needs to be identified.

The goal of this research is to explore what exactly are these regions of the wings venation pattern important for the classification of fly species. Indeed, these regions can be extracted by taxonomists through empirical works and experiences. Nevertheless, it takes a substantial amount of time and money to train a well-qualified taxonomist which becomes less practical due to the shortage of research funding [7]. Also, the taxonomists might misidentify regions that they thought to be useful for classification.

Successful implementation of this project will be extremely helpful to mitigate these issues, with the machine as the tour guide to facilitate the classification of species. Besides, it can be used to cross-validate characteristics that were previously used by taxonomists for classification. This approach has also a huge potential to discover new features that are previously unidentified by the taxonomists, which could be helpful to advance future taxonomy research.

2 Literature Review

Generally, there are different methods to classify fly species. The traditional method is based on morphological analysis but it does not work well for the Sarcophagidae family because species within this family appear morphologically similar to the untrained eye [8]. The continuous development of high throughput sequencing has made fly species identification using DNA sequencing analysis possible. While this method is powerful, it has limitations to correctly identify specifies from certain families under certain circumstances [9, 10, 11, 12, 13]. Also, the maintenance cost is high and thus DNA-based identification for routine species identification is unsustainable [14]. Thus, alternative methods are required to alleviate these issues.

Recently, Khang et al.[5] provides a proof of concept that geometric morphometric wings image data can accurately classify species in the Calliphoridae and Sarcophagidae families. These geometric morphometric wings datas' are generated by capturing homologous landmarks on the left-wing of males (Figure 1). The estimated percentage of concordance between species identities predicted using the random forests model and those inferred using DNA-based identification was about 80.6% with approximately 95% Bayesian credible interval = $[68.9\%, 92.2\%]$. Later, Khang & Goh (2021) [6] further improved the prediction performance of the random forest model by using the whole wings image venation pattern as a dataset instead of only the homologous landmarks. The images of wing venation patterns in Calliphoridae and Sarcophagidae, when extracted using a class of discrete orthogonal moments known as the Krawtchouk moments (Figure 2), contain species-specific features that enable highly accurate machine learning prediction of species identity. (100 % accurate classification results based on 74 samples from 15 species, 2 families). These methods are cost-effective as opposed to DNA-based identification as only basic optical instruments and mature analysis software are required.

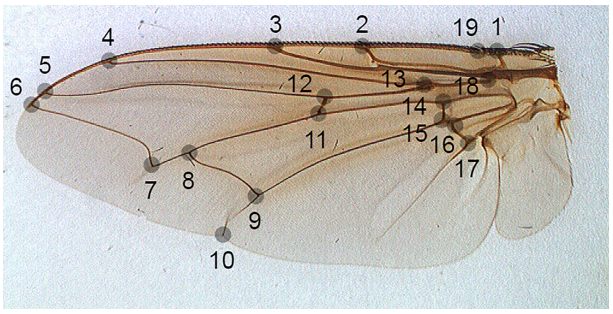

Figure 1: Raw image of left wing of a male fly. There are 19 homologous landmarks, numbered from 1-19.

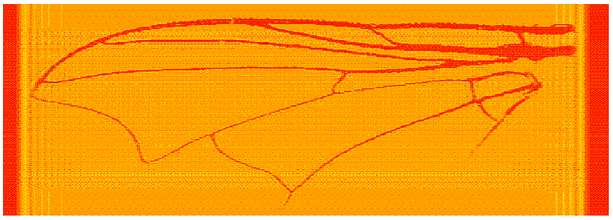

Figure 2: Reconstructed image from raw image using Krawtchouk moments

3 Significance of Research

Despite the success in the prediction performance of the random forest model, we do not really understand how the model extracts meaningful features from the data and concludes the decisions. It is paramount because the exact identification of useful features would be helpful for a data scientist to convince the taxonomists that the predictions are trustworthy. Only with this, a taxonomy features database for classification could possibly be built and used to assist future taxonomy work. Therefore, it is important to learn how a machine model learns. This framework is also commonly known as explainable artificial intelligence [15], an active research area in data science. In this project, we will study the explainable machine learning classification of Calliphoridae and Sarcophagidae using wing venation data.

4 Methodology

The present work proposes a workflow that can potentially achieve the desired level of model explainability when predicting the species identity of forensically important flies using their wing venation patterns. An existing database of 74 wing image data reported in Khang et al. [5] is prepared. These raw images are first converted to binary images. After manual denoising various effects due to artifacts etc, the images are then reconstructed into a 200×200 matrix using Krawtchouk moment invariants of order 200 [16]. This matrix is known as the matrix of Krawtchouk moments, denoted as Q and the entries of the matrix are known as moment features. Actually, Q can be related to the raw image intensity function A by the equation

$$Q = K_1 A K_2^T$$

where K_1, K_2 are the $200 \times N$ and $200 \times M$ orthogonal matrix of normalized Krawtchouk polynomials.

Following the use of cluster analysis for feature selection [17], we then capture the relatedness between features using the relatively robust Spearman rank correlation. Hierarchical clustering of the resultant 40,000-moment features is then done using the Ward algorithm. Representative feature in each of the clusters induced by a suitable cut-off parameter is then used to reduce the dimensionality of the data set. Following this step, feature selection using filter methods such as the ANOVA F-statistic can be done to identify informative features for clustering work. These informative features will then be used to classify the fly samples. Suitable machine learning models such as kernel density estimation, support vector machine, etc. can then be used to evaluate prediction performance, with an out-of-bag estimate of generalization error. If the representative features used produce good prediction performance, features that cluster with these representative features are identified and their entries in matrix Q are traced. Now, let Q^* be the matrix of Krawtchouk moments where the entry values of unimportant features are shrunk to 0. By the orthogonality of K_1 and K_2, we can then recover the matrix of image intensity

function A^* from Q^* (with unimportant features reassigned to 0) using

$$A^* = K_1^T Q^* K_2$$

It is important to notice that computational complexity will surge tremendously without the orthogonality property of these matrices, and this is why this method is useful. As a result, the wing venation pattern (A^*) thus reconstructed can then be compared with that in the raw image (A) to potentially identify regions that remain unperturbed, where such regions potentially explain how the machine learning model learns to discriminate between species.

5 Work Schedule

For the completion of this project, working days will be 5 days per week (Mon. – Fri.) at an average of 8 hours per day, totaling approximately 40 hours per week. Due to COVID-19 pandemic, I will be temporarily working remotely at Penang, with weekly (or bi-weekly) meeting with project supervisor via video conferencing application (eg. Zoom). A brief schedule is shown below:

Timeframe	Task Description
September-December 2021	Study relevant materials as a preparation for research work.
January-March 2022	Convert the raw image into matrix of Krawtchouk moments.
April - May 2022	Identify representative features using various statistical techniques.
June - July 2022	Train the machine learning model using representative features
August - September 2022	Evaluate the performance of the training model
October - December 2022	Write the results into a paper for journal publication.
January - March 2023	Review the paper for re-submission.
April - June 2023	Write master thesis report.
July 2023	Prepare slides for thesis defense.

References

1 Amendt, J., Richards, C. S., Campobasso, C. P., Zehner, R., & Hall, M. J. (2011). Forensic entomology: Applications and limitations. *Forensic science, medicine, and pathology*, *7*(4), 379–392.

2 Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd international conference on document analysis and recognition*, *1*, 278–282.

3 Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

4 Bookstein, F. L. (1997). *Morphometric tools for landmark data: Geometry and biology*. Cambridge University Press.

5 Khang, T. F., Mohd Puaad, N. A. D., Teh, S. H., & Mohamed, Z. (2021). Random forests for predicting species identity of forensically important blow flies (diptera: Calliphoridae) and flesh flies (diptera: Sarcophagidae) using geometric morphometric data: Proof of concept. *Journal of Forensic Sciences*, *66*(3), 960–970.

6 Goh, J. Y., & Khang, T. F. (2021). On the classification of simple and complex biological images using krawtchouk moments and generalized pseudo-zernike moments: A case study with fly wing images and breast cancer mammograms (in press).

7 BRITZ, R., HUNDSDÖRFER, A., & FRITZ, U. (2020). Funding, training, permits—the three big challenges of taxonomy. *Megataxa*, *1*(1), 49–52.

8 Tan, S. H. (2012). *Studies of forensically important flies of calliphoridae and sarcophagidae in malaysia: Morphological taxonomy, geographical and ecological distribution, species succession on carcasses, and dna-based identification/tan siew hwa* (Doctoral dissertation). University of Malaya.

9 Nelson, L. A., Lambkin, C. L., Batterham, P., Wallman, J. F., Dowton, M., Whiting, M. F., Yeates, D. K., & Cameron, S. L. (2012). Beyond barcoding: A mitochondrial genomics approach to molecular phylogenetics and diagnostics of blowflies (diptera: Calliphoridae). *Gene*, *511*(2), 131–142.

10 Hurst, G. D., & Jiggins, F. M. (2005). Problems with mitochondrial dna as a marker in population, phylogeographic and phylogenetic studies: The effects of inherited symbionts. *Proceedings of the Royal Society B: Biological Sciences*, *272*(1572), 1525–1534.

11 Whitworth, T., Dawson, R., Magalon, H., & Baudry, E. (2007). Dna barcoding cannot reliably identify species of the blowfly genus protocalliphora (diptera: Calliphoridae). *Proceedings of the Royal Society B: Biological Sciences*, *274*(1619), 1731–1739.

12 Williams, K., & Villet, M. H. (2013). Ancient and modern hybridization between lucilia sericata and l. cuprina (diptera: Calliphoridae). *European Journal of Entomology*, *110*(2).

13 Sonet, G., Jordaens, K., Braet, Y., & Desmyter, S. (2012). Why is the molecular identification of the forensically important blowfly species lucilia caesar and l. illustris (family calliphoridae) so problematic? *Forensic Science International*, *223*(1-3), 153–159.

14 Cameron, S., Rubinoff, D., & Will, K. (2006). Who will actually use dna barcoding and what will it cost? *Systematic biology*, *55*(5), 844–847.

15 Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE access*, *6*, 52138–52160.

16 Yap, P.-T., Paramesran, R., & Ong, S.-H. (2003). Image analysis by krawtchouk moments. *IEEE Transactions on image processing*, *12*(11), 1367–1377.

17 Park, C. H. (2013). A feature selection method using hierarchical clustering. *Mining intelligence and knowledge exploration* (pp. 1–6). Springer.