Hyperspectral imaging generated adulterated honey dataset

Tessa Phillips, Waleed Abdulla

February 8, 2022

This is a generated dataset of hyperspectral images of adulterated honey. The dataset was generated using a variational autoencoder and a latent space transform based on the ground truth honey adulteration data published online [1]. The generated data utilises the larger dataset of botanical origins honey available at [2], and covers the same brands and classes of honey at each adulteration concentration $(0\%, 5\%, 10\%, 25\%, 50\%)$. This is a generated dataset, so is not strictly real ground-truth data, however the trends followed in the generated data are similar to the ground-truth adulteration, so this is a reasonable approximation of a larger dataset of adulterated honey.

The overall shape of the generated data is provided in figure 1. We can see that generally the adulterated data has a higher spectral response in the mid-wavelengths, however the spectral response is unique for each different type of honey.

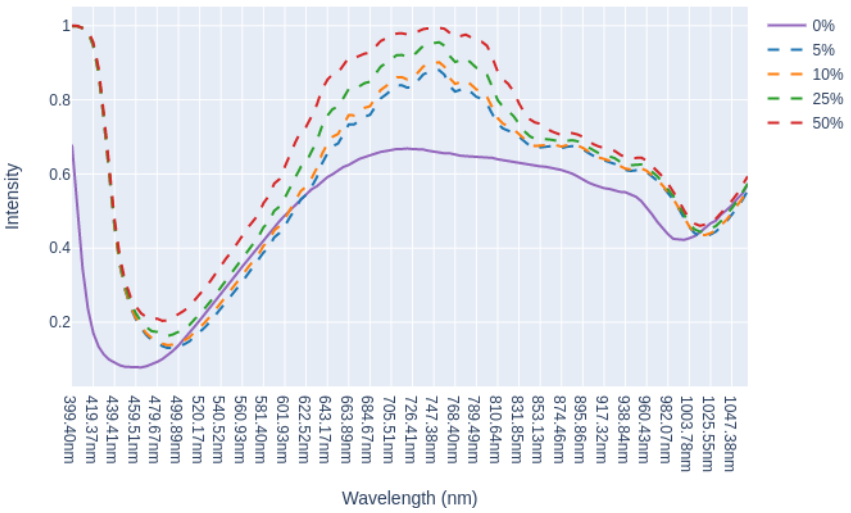

Figure 1: Average spectral response of generated adulterated honeys.

0.1 Attribute Description

The features from the hyperspectral images are the working wavelengths of the hyperspectral camera. Several additional attributes have been considered and can be useful for splitting the dataset into training and testing sets, as well as testing the generalisation ability of the algorithms.

The 'Brand' of honey represents the manufacturer that has supplied the honey. This attribute is included because it can be useful to test if a system can classify all honey types within a brand against each other. It can also be useful when developing general systems to check if we can exclude a brand from the training set and still have a good performance with the testing set. The brands have been anonymised for confidentiality reasons, the brand labels have been renamed as $C1, C2, ..., C11$.

The 'Acquisition' attribute represents the different sampling of images for the same type and brand of honey. For each unique jar of honey, there have been six samples taken and captured by the hyperspectral imaging system. Each image captured is numbered with an acquisition number between one and six. This attribute allows us to split the training and testing sets such that we obtain a balanced distribution of all the honey types. This also ensures that we do not have an instance in the testing set that comes from a segment included in the training set. For testing, we use acquisition number six, and for training, we use acquisitions one to five.

The class attribute indicates the class of honey, which is the botanical origin, and the UMF value if it is UMF rated Manuka honey. Botanical origins have a huge impact on the value of honey, where some types are precious such as pure Manuka honey, and others are much more common and not considered as valuable, such as multi-floral honey.

Finally there are two sugar concentration attributes. The first 'concentration' represents the actual concentration of sugar in the sample this could be for example 5.001. This attribute is used for regression type algorithms as it has the exact concentration of sugar that we have adulterated the pure honey sample with. The second attribute is 'concentration_class' this represents the class grouping of the adulterated sample as either '0', '5', '10', '25', or '50'. This attribute is used for classification problems to detect what group the adulterated honey should fit into.

References

[1] T. Phillips, B. Coleman, S. Takano, and W. Abdulla, "Hyperspectral imaging adulterated honey dataset," Aug 2021. [Online]. Available: https://auckland.figshare.com/articles/dataset/Hyperspectral_Imaging_adulterated_honey_dataset/16441686/1

[2] T. Phillips, A. Noviyanto, and W. Abdulla, "Hyperspectral imaging honey database," 4 2020. [Online]. Available: https://figshare.com/s/25afe30ff531b8f1e65f