A new framework to support headspace chemical forensics using wavelet scalogram visualization of mass spectrometry data and transfer learning

Ting-Yu Huang, M.S.*, and Jorn Chi-Chung Yu, PhD

Department of Forensic Science, Sam Houston State University, Huntsville, TX 77340

MEMBER THE TEXAS STATE UNIVERSITY SYSTEM

ABSTRACT

This poster presents a novel intelligent framework for classifying hemp and marijuana from the images transformed from gas chromatography and mass spectrometry (GC/MS) data. The data visualization approach using continuous wavelet transform (CWT) and transfer learning of a deep convolutional neural network (CNN) is presented and discussed.

INTRODUCTION

The high efficiency of GC/MS in separating components allows the analysts to perform qualitative and quantitative chemical analyses of forensic evidence. Because the interpretation of GC/MS data highly relies on the analyst's skill and experience, the turnaround time for evidence analysis in forensic laboratories may be affected. Additionally, the classification of forensic evidence is often required to provide investigative leads. In such case, chemometric techniques can be adopted to analyze GC/MS data.

A deep CNN is an emerging machine learning technique that uses multiple layers to identify features in an image in a hierarchical manner and making inferences on categorical classification. However, this technique may fail to perform well when dealing with a small amount of data. Transfer learning is a more efficient learning approach than training a CNN from scratch with randomly initialized weights [1]. It has also been reported that, by using transfer learning, higher classification performance can be achieved.

Recently, digital image analysis has been investigated to process chromatograms and mass spectra to produce various 2-dimensional (2D) images. It is believed that distinctive patterns can be generated from raw GC/MS data to allow the analysts conducting a comparative analysis of the samples. CWT is a technique that has been applied to denoise or detect signals in bioinformatics, Raman spectroscopy, and mass spectrometry. One type of the mother wavelet in CWT is Morse wavelet. The wavelet is widely utilized in processing modulated signals with time-varying amplitude and frequency [2].

In this study, we proposed that the 2D images transformed from the summed ion mass spectrum by CWT contained characteristic patterns for evidence classification. The patterns could be further recognized by a pre-trained CNN model through transfer learning. We chose hemp and marijuana samples to evaluate the proposed workflow. The Agriculture Improvement Act of 2018 has provided a new statutory definition of hemp, which defines the cannabis plant, and any part of it, with a delta-9-tetrahydrocannabinol (THC) concentration of not more than 0.3% on a dry weight basis [3]. An intelligent classifier for discriminating between hemp and marijuana can benefit forensic cannabis analysis greatly. The proposed workflow requires minimal manual effort and gives high prediction performance, which are attractive features for forensic laboratories.

RESULTS AND DISCUSSION

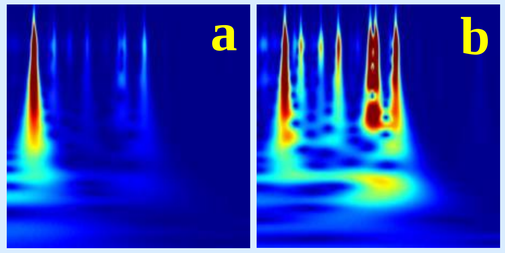

Figure 1: Examples of scalograms of a (a) hemp, THC (w/w) 0.08%, CBD (w/w) 3.4% and (b) marijuana samples, THC (w/w) 13.4%, CBD (w/w) 0.03%.

Table 3: Comparison of (a) means and uncertainties at 95% confidence intervals for confusion matrices (b) accuracies and uncertainties at 95% confidence intervals for the proposed classifier and five machine learning algorithms using 100 bootstrap sampling of verification data

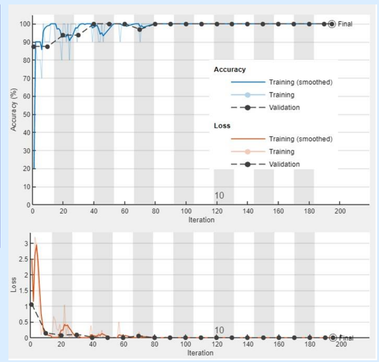

Figure 2 : Training progress of the proposed classifier using the combination of optimal values of the hyper-parameters.

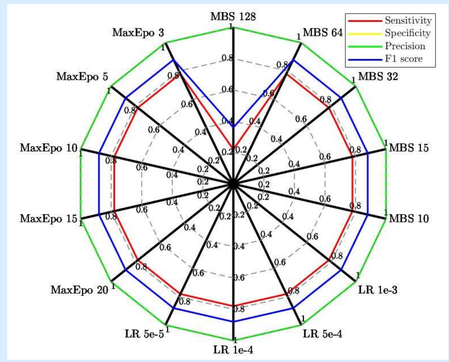

Figure 3 : Radar chart showing the evaluation measures of the verification data in the hyper-parameter optimization tests.

(a)	CNN (GoogLeNet)		KNN		Discriminant Analysis		Naive Bayes		SVM		Ensemble	
	Hemp	Marijuana	Hemp	Marijuana	Hemp	Marijuana	Hemp	Marijuana	Hemp	Marijuana	Hemp	Marijuana
Hemp	7.25± 0.2	1.75± 0.2	5.01± 0.25	3.99± 0.25	6.82± 0.21	2.18± 0.21	5.97± 0.26	3.03± 0.26	4.81± 0.26	4.19± 0.26	2.9± 0.24	6.1± 0.24
Marijuana	0 ± 0	57 ± 0	0 ± 0	57 ± 0	0.93± 0.18	56.07± 0.18	6.27± 0.4	50.73± 0.4	0 ± 0	57 ± 0	3.49± 0.30	53.51± 0.30

(b)	CNN (GoogLeNet)	KNN	Discriminant Analysis	Naive Bayes	SVM	Ensemble
Accuracy	0.97± 0.12	0.94± 0.08	0.97± 0.1	0.86± 0.1	0.94± 0.08	0.85± 0.09

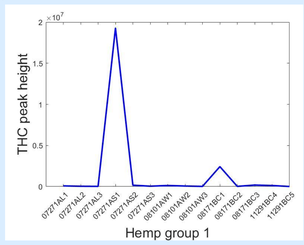

Figure 4 : Comparison of THC peak heights in hemp samples (group 1)

Hyper-parameter optimization tests

The average probabilities of the classifier increased as the number of epochs rose before epochs 15. For the learning rates, the prediction probabilities gradually improved when the value of the hyper-parameter increased. Higher epochs and a learning rate aided in decreasing the prediction error for our proposed classifiers. For the mini batch size, the average probabilities and accuracy decreased as it rose and fell by the lowest values for mini batch size 128. A small mini batch size ensured a higher classification performance for the proposed classifier. Overall, the combination of optimal values of the hyper-parameters were MaxEpo 15, LR 1e-3, and MBS 10.

Training progress and overall performance

The classifier reached 100% validation accuracy and 0% validation loss after training (Figure 2), suggesting that the GoogLeNet model with the proposed hyper-parameter values has successfully developed an intelligent classification system for classifying scalograms transformed from GC/MS data by CWT into correct hemp or marijuana classes. The scores of the evaluation measures indicate that the classifier did not rely on optimizing critical hyper-parameters to secure good classification performance (Figure 3). It is suggested that the features of a scalogram were easy to be extracted and recognized by the GoogLeNet model.

Comparison test

As seen in Table 3 (a), the classifier performs well (57 ± 0) in classifying marijuana samples using 100 bootstrap sampling of verification data. The slightly decreased performance on classifying hemp samples should be attributed to samples 07271AS1 and 08171BC1, whose THC peak intensities were deviated from the median of the THC peak intensities of all group 1 hemp samples (Figure 4). When comparing the accuracies of the models, as shown in Table 3 (b), the classifier and discriminant analysis have better performance (0.97 ± 0.12 and 0.97 ± 0.1, respectively) than the other ML models. Overall, the results suggest that the wavelet transform approaches offered satisfactory capabilities to transform GC/MS data for transfer learning.

CONCLUSIONS

- The summed ion mass spectra could be used to generate 2D images by CWT.
- The transformed scalograms provided distinctive patterns for the GoogLeNet to discriminate hemp or marijuana.
- Pre-processing of GC/MS data, such as peak alignment could be eliminated.
- The AI-powered classifier achieved high performance in accuracy, sensitivity, specificity, precision, and F1 score.
- Homogenization of cannabis plant samples and increasing the sample mass for HHS-SPME are highly recommended.

REFERENCES

[1] S. Khan, N. Islam, Z. Jan, I. Ud Din, J.J.P.C. Rodrigues, A novel deep learning based framework for the detection and classification of breast cancer using transfer learning, Pattern Recognition Letters. 125 (2019) 1–6. https://doi.org/10.1016/j.patrec.2019.03.022.
[2] S.Y. Sarraf, R. Trappen, S. Kumari, G. Bhandari, N. Mottaghi, C.Y. Huang, M.B. Holcomb, Application of wavelet analysis on transient reflectivity in ultra-thin films, Opt. Express. 27 (2019) 14684. https://doi.org/10.1364/oe.27.014684.
[3] H.R.2 - Agriculture Improvement Act of 2018, https://www.congress.gov/bill/115th-congress/house-bill/2/text (accessed 9 September 2021).
[4] A. McDaniel, L. Perry, Q. Liu, W.C. Shih, J.C.C. Yu, Toward the identification of marijuana varieties by headspace chemical forensics, Forensic Chem. 11 (2018) 23–31. https://doi.org/10.1016/j.forc.2018.08.004.

MATERIALS AND METHODS

Cannabis Data Set and Headspace Chemical Analysis The cannabis data set was built from data previously collected by headspace solid phase microextraction (HHS-SPME) GC/MS analysis in our research group [4]. Fourteen different cannabis varieties with known levels of THC and cannabidiol (CBD) were provided by the National Institute on Drug Abuse (NIDA) and used as the standard reference samples. Ten mg of each cannabis sample was measured by an analytical balance and placed into a 20-mL headspace vial. A 23 gauge and 100 µm polydimethylsiloxane (PDMS) SPME fiber obtained from Sigma-Aldrich (St. Louis, MO) was installed onto a PAL autosampler supplied by Agilent (Santa Clara, CA) for headspace sampling. Agilent 7890B coupled with a 5977A mass selective detector was used for sample analysis. Detailed experimental parameters can be found in [4].

A total number of 228 GC/MS data were collected, including 30 data in the hemp class and 198 data in the marijuana class. The GC/MS data were separated into training (162 data) and verification (66 data) subsets for transfer learning and the confirmation of the training outcomes respectively.

GC/MS Data Processing The GC/MS data were first converted into NetCDF format using the Agilent ChemStation. Then, the data were processed using the Bioinformatics Toolbox in MATLAB (Natick, MA). To resample raw GC/MS data into equally spaced signals, we constructed a data structure involving retention times (denoted as vector "Time"), mass to charge ratio and intensity values (denoted as matrix "MZ") for each GC/MS data. Table 1 shows a range of data characteristic to THC, CBD, and cannabinol (CBN) in vector Time and matrix MZ. The fingerprint regions were extracted and used to produce the summed ion mass spectra by summing up the intensities of each m/z value over retention times in the extracted region. The summed ion mass spectra were then computed by the CWT filter bank (Table 2) to create scalograms (Figure 1). Each scalogram was resized in an RGB image format that was an array of size 224-by-224-by-3.

Transfer Learning of GoogLeNet A pre-trained CNN, GoogLeNet, was fine-tuned to recognize the features in the scalograms to classify hemp and marijuana samples. The final three layers (pool5-drop_7x7_s1, loss3-classifier, output) were replaced with new layers to adapt to the data in the work. Several hyper-parameters, including maximum epochs, learning rate, and mini batch size were selected to optimize the classification performance (Table 2). During transfer learning, the training data were randomly divided into 80% (130 images) and 20% (32 images) for training and validation phases.

Table 1: Table 1: Data extraction of retention time (TIC) and major ion fragments (mass spectrum)

Data structure	Original data range	Extracted region
Time	0 to 16 minutes	6 to 13 minutes
MZ	m/z 40 to 450	m/z 210 to 393

Table 2: Parameters of CWT and transfer learning

Types of Analysis	Parameter	Input
CWT	Signal Length	1000
	Sampling Frequenc	Fs
	Voices Per Octave	12
Transfer Learning	Mini Batch Size	Varied as 10, 15, 32, 64, 128
	Max Epochs	Varied as 3, 5, 10, 15, 20
	Initial Learn Rate	Varied as 0.00005, 0.0001, 0.0005, 0.001
	Validation Frequenc	10

ACKNOWLEDGEMENTS

This work is partly funded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice [Award #2014-R2-CX-K005]. The opinions, findings, and conclusions or recommendations expressed in this presentation are those of the author(s) and do not necessarily reflect those of the Department of Justice. The authors would like to acknowledge Austin McDaniel for his assistance in performing laboratory experiments.