The Classification of Cannabis Varieties Through Headspace Chemical Analysis 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 new intelligent analytical platform for discriminating between hemp and marijuana. The platform integrates headspace chemical analysis and transfer learning of a deep convolutional neural network (CNN). The transfer learning technique enables superior classification performance of cannabis varieties without manual intervention.

INTRODUCTION

Cannabis sativa L. (cannabis) has been known to produce cannabinoids that may have diverse bioactivities. The Agriculture Improvement Act of 2018 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 [1]. The new law limits the definition of marijuana only to include cannabis containing more than 0.3% THC. To discriminate between hemp and marijuana, proficient analytical workflows for cannabis samples are essential in forensic laboratories.

GC/MS is a commonly accepted analytical instrumentation to identify cannabinoids. However, interpreting GC/MS data typically involves an analyst's operation. Due to the heterogeneity of cannabis samples, extensive sample preparation is also required before instrumental analysis. A solid phase microextraction (SPME) headspace sampling approach is suited for increasing the throughput of extracting cannabinoids before using GC/MS for chemical identification. To decrease the turnaround time, an artificial intelligence (AI) classifier is exceptionally feasible for automating the process of data analysis. Deep learning models are the state-of-the-art machine learning algorithms that use multiple layers to extract lower to higher-level features from the raw input in a hierarchical manner. The technique has been proven to demonstrate a huge potential for image processing [2]. To improve the learning efficiency, transfer learning enables transferring knowledge from an existing machine learning model and reuses the knowledge to learn a new related task.

In this study, we explore the viability of the proposed workflow to transform headspace GC/MS data into 2-dimensional (2D) images and adopt transfer learning of a pre-trained CNN model to develop an AI-based classifier for discriminating between hemp and marijuana samples. The workflow requires minimal manual effort and gives high prediction performance, which are attractive features for modern forensic cannabis analysis.

RESULTS AND DISCUSSION

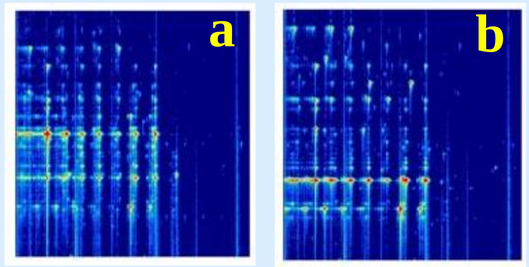
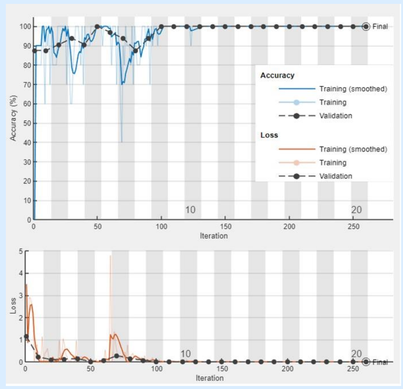
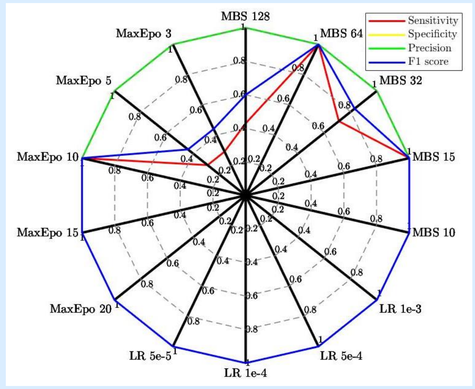


Figure 1: Examples of pseudo-color heat maps of a (a) hemp, THC (w/w) 0.08%, CBD (w/w) 3.4% and (b) marijuana samples, THC (w/w) 13.4%, CBD (w/w) 0.03%.





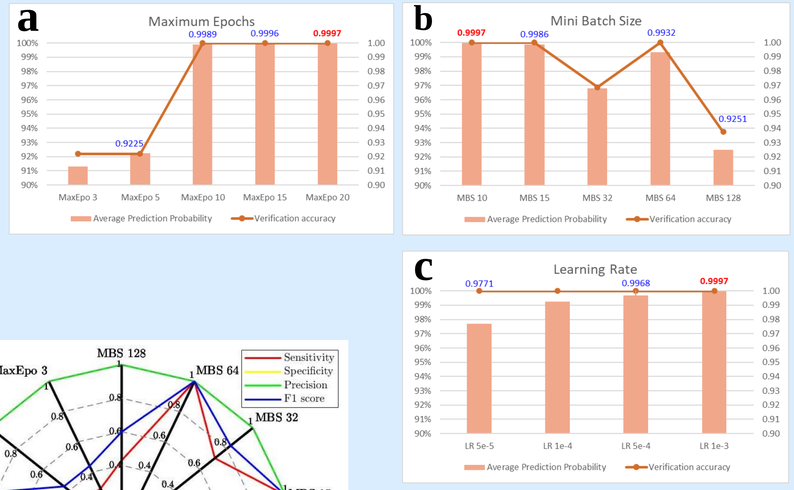


Figure 2 (Upper) : Accuracy and average prediction probabilities of the verification data in the hyper-parameter optimization tests: (a) maximum epochs (with LR 1e-3 and MBS 10); (b) learning rate (with MaxEpo 20 and MBS 10); (c) mini batch size (with MaxEpo 20 and LR 1e-3). In order to compare the variation of the average prediction probabilities, the values were rounded to 4 decimal places.

Figure 3 (Left) : Training progress of the proposed classifier using the combination of optimal values of the hyper-parameters: MaxEpo 20, LR 1e-3, and MBS 10.

Figure 4 (Middle): Radar charts showing the evaluation measures of the verification data in the hyper-parameter optimization tests.

Hyper-parameter optimization tests

The accuracy and average probabilities of the classifier increased as the number of epochs rose (Figure 2 (a)). For all learning rates, the classifier obtained the highest score for accuracy. Further comparison of the average probabilities among all learning rates indicates that the prediction probabilities gradually improved when the value of the hyper-parameter increased, which means that a higher learning rate aided in decreasing the prediction error for our proposed classifiers (Figure 2 (b)). For mini batch size, a sudden enhancement in the performance was found for mini batch size 64, then it fell by the lowest values for mini batch size 128. Setting a small mini batch size ensured a higher classification performance for the proposed classifier (Figure 2 (c)). Overall, the combination of optimal values of the hyper-parameters were MaxEpo 20, LR 1e-3, and MBS 10.

Training progress and overall performance

The classifier reached 100% validation accuracy and 0% validation loss after training (Figure 3), suggesting that the GoogLeNet model with the proposed hyper-parameter values has successfully developed an intelligent classification system for classifying pseudo-color heat maps transformed from GC/MS data into correct hemp or marijuana classes.

The scores of the evaluation measures indicate that the classifier relied on optimizing critical hyper-parameters to secure good classification performance (Figure 4). Final cores demonstrates that the classifier provides outstanding and robust performance for cannabis varieties discrimination.

REFERENCES

[1] H.R.2 - Agriculture Improvement Act of 2018 https://www.congress.gov/bill/115th-congress/house-bill/2/text (accessed 6 January 2022).

[2] S. Deepak, P. Ameer, Brain tumor classification using deep CNN features via transfer learning, Computers in Biology and Medicine. 111 (2019) 103345. https://doi.org/10.1016/j.compbiomed.2019.103345.

[3] A. McDaniel, L. Perry, Q. Liu, W.C. Shih, J. Yu, Toward the identification of marijuana varieties by headspace chemical forensics, Forensic Chemistry. 11 (2018) 23–31. https://doi.org/10.1016/j.forc.2018.08.004.

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.

MATERIALS AND METHODS

Sample Description and Headspace Chemical Analysis The cannabis data set was built from data previously collected by headspace SPME GC/MS analysis in our research group [3]. The standard reference cannabis samples, which were provided by the National Institute on Drug Abuse (NIDA), contained 14 different varieties with various known levels of THC and cannabidiol (CBD). Those samples comprised mixed dry botanical structures including buds, leaves, and stems. A total number of 228 GC/MS data were in the data set, 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 the purposes of transfer learning and confirmation of the training outcomes consecutively.

GC/MS Data Processing The GC/MS data were first converted into NetCDF format using the Agilent ChemStation (Agilent Technologies, Inc., California, USA) for the subsequent data processing in the Matlab environment (Matlab 2021a and Bioinformatics Toolbox, MathWorks, Natick, Massachusetts, USA). To resample raw GC/MS data into equally spaced signals, data matrixes involving retention times (denoted as vector "Time"; recorded from 0 to 16 minutes), mass to charge ratio and intensity values (denoted as matrix "MZ"; scanned from 40 to 450 m/z) were constructed. To facilitate feature recognition and discrimination between the sample headspace chemical signatures of hemp and marijuana in the transfer learning process, a range of data characteristic to the cannabinoids including THC, CBD, and cannabinol (CBN) in vector Time and matrix MZ were extracted (Table 1). The 2D visualization of GC/MS data was the use of pseudo-color heat maps. The image displays the intensities for the spectra after a log transformation of the selected range of m/z values at the chosen range of retention times (Figure 1).

Transfer Learning of GoogLeNet GoogLeNet, a pre-trained CNN, was adapted and fine-tuned to learn the task of cannabis classification ("hemp" or "marijuana"). Initial layers in the GoogLeNet structure were retained, while the final three layers (pool5-drop_7x7_s1, loss3-classifier, output) were replaced with new layers. 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: Data extraction of retention time (TIC) and major ion fragments (mass spectrum)

Data structure	Extracted region in original data	Extracted region in data structure
Time	6 to 13 minutes	927th to 2397th elements
MZ	210 to 393 m/z	835th to 1730th elements

Table 2: Parameters of transfer learning

	Parameter	Input
Fine-tuning the weights of GoogLeNet	Max Epochs (MaxEpo)	Varied as 3, 5, 10, 15, 20
	Initial Learn Rate (LR)	Varied as 0.00005, 0.0001, 0.0005, 0.001
	Mini Batch Size (MBS)	Varied as 10, 15, 32, 64, 128

CONCLUSIONS

- An AI-powered classifier for cannabis variety discrimination is developed without collecting a large reference data set.
- Achieving 100% high performance in accuracy, sensitivity, specificity, precision, and F1 score.
- No user involvement in sample preparation, data pre-processing, and interpretation.
- The probability-based classification outcome aids in statistical evaluations for interpreting forensic evidence in legal proceedings.