**Characterization of E-cigarette Smokes through Machine Learning Classification Algorithms**

Sunwoo Kang1, Ri Chen2, Sinharoy Pritam3, Eric Gross4

1-4 Department of Anesthesia, Perioperative and Pain Medicine, Stanford School of Medicine, Stanford, California, United States of America

**Method**

*Machine Learning (ML) Algorithm*

This research utilized six different supervised learning classification algorithm: k-Nearest Neighbors (KN), Naïve Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and Artificial Neural Network (ANN). Various ML algorithms were used since the study sought for the best way to distinguish the different e-cigarette smokes. These algorithms were implemented in Python script using scikit-learn (sklearn) package[[1]](#footnote-1). The details of implementation could be found here [pseudo-code].

The crucial part of learning involves pre-processing the sample data. For data pre-processing, this paper used Standard Scaler (SS) and Principal Component Analysis (PCA) available from Python sklearn package. In some instances, the algorithm showed higher performance with raw sample data. The scheme of analysis is shown in [figure: ML algorithm overview]. Based on the typical ML training scheme[[2]](#footnote-2), the size of training set was set as n=12, validating set was n=4, and testing set was n=4 for each e-cigarette brand.

*Sample size validation*

Since the study made no assumption regarding the identity or variance of potential compounds of interest[[3]](#footnote-3), we validated the sample size through drawing learning curve for each supervised learning algorithms[[4]](#footnote-4). Although the learning rate differed on each algorithm, five out of six algorithm conformed to the inverse power law model. With the threshold of 0.8 accuracy, sample size of 12 was selected for the training set [figure: Learning curve].

*Choice of E-cigarettes*

In order to identify the most popular, convenient, and highly rated electronic cigarettes, we searched and discerned several online polls, trends, and internet discussion forums. This led to the five brands of electronic cigarettes used in our study: Juul, Blu, Halo, V2, and VaporFi. Characteristics of each electronic cigarette are listed in table shown below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Product Name** | **Brand** | **Cartridge Type** | **Ingredients** | **Flavor** | **Labelled Nicotine Content (mg/mL)** |
| Juul | Juul | Juul Pod | Glycerol, Propylene Glycol, Flavor, Nicotine, Benzoic acid | Classic Tobacco | 50 |
| Blu Plus+ Rechargeable | Blu | Cartomizer | Vegetable Glycerin, Propylene Glycol, Nicotine, Natural and Artificial Flavors, Water | Classic Tobacco | 24 |
| Halo G6 Rechargeable | Halo | Cartomizer | Propylene Glycol, Glycerin, Flavorings, Nicotine | Prime 15 | 24 |
| V2 Red Disposable | V2 | Disposable | Propylene Glycol, Vegetable Glycerin, Nicotine, Natural Flavors, Artificial Flavors | Tobacco | 18 |
| VaporFi Express | VaporFi | Refillable Cartomizer | USP Propylene Glycol (66%), USP Glycerin (21%), Natural and Artificial Flavorings (30%), USP Nicotine (2.4%) | Classic Tobacco | 24 |

*Smoke Sample Collection & Measurement*

Experiments were set up to reflect real-life smoking patters of electronic cigarette users. Batteries were recharged before every trial. A trial consisted of drawing approximately 400 – 450 mL of vapor into 0.5 L Tedlar Bags. The vapor from electronic cigarettes was generated through Masterflex tubing (96410-16) by a Masterflex L/S pump head (model 7518-00), connected to a Masterflex L/S pump drive (model 77300-40) and operated by a Masterflex L/S digital modular drive, at a rate of 600 mL/min, or 10 mL/sec. The vapor in the Tedlar Bags was then transported to the Voice200ultra (Syft co.). From there, we ran Selected Ion Flow Tube-Mass Spectroscopy (SIFT-MS) tests on each sample. The machine had following setting during the sample measurement: carrier flow of 5.6tls, sample flow of 0.3tls, tube pressure of 0.69, tube temperature 119, and reaction time of 10 min for eight reagent ions (H3O+, NO+, O2+, OH-, O2-, O-, NO2-, NO3-). Each reagent ion yielded ion count measurement from mass to charge ratio range of 15 to 400. Thus, each sample measurement consisted of 3088 data points.

1. <http://scikit-learn.org/stable/> [↑](#footnote-ref-1)
2. <https://www.coursera.org/learn/machine-learning> [↑](#footnote-ref-2)
3. Exhaled volatile organic compounds for phenotyping chronic obstructive pulmonary disease: a cross-sectional study [↑](#footnote-ref-3)
4. Predicting sample size required for classification performance [↑](#footnote-ref-4)