Effects of Handling Missing Values of VOCs Gases Emitted From Human for Human Detection

Ahmed Qusay Sabri, Rayner Alfred

Abstract—The main objective of this paper is to investigate the effects of replacing missing values of Volatile Organic Compounds (VOC) that represent gases emitted from human on the accuracy of classifying individual person. These missing values will be replaced with three possibilities, which include 0, 1, and Random number between 0 and 1. The effects of using these predefined values on the classification accuracy are investigated by conducting experiments that involve applying a list of classification methods to classify 15 humans using human odour. Each person is characterized by their own pre-selected 15 gases emitted from their sweats. In this paper, we also study and determine the minimum number of gases that is required to produce acceptable results to correctly classify an individual person based on the gases emitted from their bodies. Based on the results obtained from the conducted experiments, the maximum and minimum allowable numbers of missing gases in human odour samples in reference to human emitted gases are 4 and 3. The best accuracy result when missing values are introduced in the odour dataset is the ensemble Bagged Trees.

Keywords: Tree Classification, SVM, KNN, Ensemble Classification, VOC, Odour

1. INTRODUCTION

The human body generates Volatile Organic Compounds (VOCs) that provide some information about diseases, behavior, emotional state and health status of a person. Also, body odor is one of the physical characteristics of human which can be used to identify people. There are researches conducted that apply human body odor obtained from armpits to classify individual persons. For instance, the electronic nose combines a sensing array with a data analysis system [23]. Human odour detection considered as non-physical contact biometric procedure to detect human. Odour of human emitted from varieties of body parts like urine, armpits and feet. The best human part used to collect VOCs emitted from person is armpits. Robust smell emitted from armpits part of human body based on huge number of glands and bacteria, hence this determines a unique pattern lead to person detection [32]. The body odour of humans consists of numerous forms of components. Research conducted earlier has focused to identify the relevant Volatile Organic Compounds (VOCs) which can be utilized for detecting gender [5]. Twenty samples of male and female humanodour were then collected with 15 out of 198 (VOCs) selected from each person using both feature selection entropy method and feature selection Chi-square method for detecting and classifying gender via artificial neural networks (ANN). Several topologies of ANN are tested through variation number of neurons and hidden layers as well [7]. Gender detection system reach accuracy percentage of 94.7%, 94.6%, 93.4% with four layers/thirty neurons, three layers/thirty neurons and two layers/twenty neurons of network structures tested respectively [6]. The factor of gender is one of the established demographic human being attributes; In addition to the detection of gender, the computer vision is also applied to recognize other human attributes and features like age and ethnicity to be used and server for security, monitoring, surveillance and biometrics [5, 9, 20].

Metabolic variations in human body are at times associated with run of the mill scents that may e estimated on breath of air, liquid sweat or other excreta from human. Personal odour comprises diverse VOCs which can be utilized as biomarkers for sicknesses. Investigations with highly prepared dogs demonstrated that they can notice infections as e.g. tumor disease, hypoglycemia illness or hyperglycemia. Nevertheless, till current time no one know precisely what compound these canines truly smell (which material is identified with a devoted illness). On the off chance that we would know the ailments smell segments we could create electronic noses as electrical detector of scent. An electronic nose (e-nose) framework is a specialized tool for assurance of VOCs, the amount of e-nose requested apps in the medicinal sector rapidly and developed quickly. A portion of these apps are the recognition of mycobacterium tuberculosis in sputum tests and cultures, diagnosis of urinary tract diseases, microbes clustering, define of schizophrenic patients and diabetes, and investigation of the breathed-out breath in various illnesses, such as, lung growth and asthma [19]. This is among many reliable biometric traits, like face, which is a very popular and has an accessibility reputation. However, the malicious environments, unfortunately, simplify several raiders to create a fake copy easily and break or travesty these systems of face recognition. Spoofing stands as an effort to acquire authentication into a biometric framework via
submitting a kind of fraud proof for a legal user. This face recognition sensitivity has induced serious care from the biometric society with several publications on countermeasure studies being published [6].

The issue of soft biometric traits classification related to age, ethnicity, gender, and odour has not yet been a resolved due to problematic exposure conditions [15, 21]. The biometric benefit considered more stable and stronger enough not to change or lose. Several human features and body portions considered to be biometric signatures such as eyes, teeth, ears, hand fingers, voices, hands, gaits, veins, feet, typing styles and signatures. Those biometric have weakness points and strength points as well, consequently biometric is employed in applications of Identification (authentication) [4]. The pillar of our investigation is to have GC/MS test (Gas Chromatograph Mass Spectrometry) used on a collection of Men and Women individuals sampled in various days of the week to guarantee the steadiness of transmitted VOCs (Volatile Organic Compounds). Our work achieved the best steady with an exact set of VOCs produced from human with gas name, obtaining the specific set of 15 VOCs to be utilized for the human recognition procedure.

One of the preprocess procedures conducted to improve classification is feature selection. As features in datasets may interact with each other in determining the class labels of instances in the dataset, it is very important to look at how these features affect the classification results when they are removed from the datasets [1, 2]. Our main target in this research is to investigate the performance of different classification methods in case of incomplete human odour samples with a missing of some VOCs emitted by human. In this experiment, using incomplete odour samples, the classification performance of a list of classifiers will be investigated. These classifiers include Tree Classification (Fine, medium and coarse), SVM Classification (Linear, Quadratic, Cubic, Gaussian, Medium Gaussian, Coarse Gaussian), KNN Classification (Fine, Medium, Coarse, Cosine, Cubic, Weighted) and Ensemble Classification (Ensemble Boosted Tress, Ensemble Bagged Tress, Ensemble Subspace Discriminant, Ensemble Subspace KNN and Ensemble RUSBoosted Tress). In other words, this paper investigates the effects of replacing missing values of Volatile Organic Compounds (VOC) that represent gases emitted from human with predefined values. Those missing values will be replaced with three possibilities, which include 0, 1, and Random number between 0 and 1. The minimum number of gases that is required to produce acceptable results to correctly classify an individual person based on the gases emitted from their bodies will also be studied and investigated.

The remaining of the paper will be as follows; firstly, some related works on human identification and human VOC’s roles will be highlighted. Secondly, the processes involved in human identification and classification are described. Next, the results obtained will be analyzed and discussed. Finally, the research conclusion is presented by offering some future work suggestions which can be constructed upon the obtained results in this research.

II. RELATED WORKS

Humans are observed to have distinctive scents like a fingerprint or signature. The axillary region plays a significant role in producing individual odour [22, 23]. A system for odour sensing through various sensors with a technique for pattern recognition and with partially overlapping particularities has been examined [30]. Odour classification in dynamically varying concentration can be beneficial in various applications such as environment monitoring, fire detection, etc. However, odour classification reliant on transient sensor responses has been conducted so far under the constant odour concentration condition [30]. Classification of odors in dynamically changing concentration is of great use in many applications such as fire detection, environment monitoring, etc. Odor classification even based upon transient sensor responses, however, has been so far carried out under the condition of the constant odor concentration [32]. Research has been successfully conducted in identifying two different people through detecting human armpit odour by means of desktop electronic nose (E-nose). It has been further found that applying deodorant does not have any effective element on the relative identification of those two different people. However, throughout the experiment, the volunteer persons were asked to control their daily ordinary activities and lifestyle to behave averagely in a similar way on agreed guidelines. In practice, these guidelines and rules are considered difficult to be followed in most real-life world situations and applications. Furthermore, desktop E-nose is rather big and involves consuming more power that makes it problematic to be used in sampling process. An original E-nose that is handheld is more desirable as it can be utilized to recognize and detect human odour in the armpit body part. The E-nose produced can operate through a Wi-Fi or local area network with low power consumption option [32]. Another study was conducted to examine the facility of an e-nose to detect variations in odour of patient peoples with diverse levels of renal dysfunction as compared with non-patient people. All healthy people (n=11) managed to be distinguished from renal failure people (n=62) using quadratic discriminant analysis and principal component analysis (PCA). The discrimination among study subjects with chronic renal failure and with patients have critical level of renal failure (n=42) was successful at a rate of 92.5 %. Primary outcomes of a clinical pilot study discovered a classification rate at 100% among primed alcoholic addicted patients, liver cirrhosis patients, and healthy subjects. The study objective was to apply e-nose system to diagnose heart failure through identifying disease-related VOCs from sweat produced by the skin surface and triggered by an impaired metabolic [19]. Efficient and accurate results obtained for fingerprint detection and face recognition using artificial neural net [4, 10, 13, 14, 17]. Palm print and biometric hands detection employing back-propagation NN with Levenberg-Marquardt algorithm applied for classification of Palm print Biometrics. Results acquired shown that classification percentage attained at 99.9% using 2 hidden layers [11, 12].
Consequently, in this study, the training algorithm of Levenberg adopted with two others neural net methods. Employing neural net in the field of biological informatics is extremely effective. In fact, neural net comprises structure of layers input, output and three hidden back-propagation procedure with gradient descent equation employed as person detection for classification reaches rate 97% [24]. Besides, a structure with multi-layer neural net using Levenberg algorithm was utilized as training algorithm for checking weights update of the hepatitis diseases diagnosis neural net. The results revealed a classification accuracy at a rate of 91.87% [25]. Classification of heart and brain biometrics by artificial neural network has accuracies at a rate of 92.4% to 95.1% [26].

III. HUMAN IDENTIFICATION PROCESS FROM ODOUR

A. Sampling procedure: GC/MS (Gas Chromatograph Mass Spectrometry) Test

Research participants attend laboratory for sampling in consecutive days during the week, variety of genders and age group considered. They will go through a seven-day washout period before the test program during which they are permitted to only wash their armpit with a kind of neutral soap. They need to be present at the lab every morning to wash as a standardized procedure. They will be asked to wear a cotton t-shirt with pads under the armpits for a period of 4 hours during which they must adhere to their normal day activities. To have the best odour representation, study subjects are requested to run up and down the stairs for approximately 3 times with 3 minutes per time round. The individuals are requested to carry cotton pads to the lab after 4 hours.

Cotton pads with target sweat are presented to a Nalophan bag of 1.5 L nitrogen. The pad is then heated for 30 minutes up to 90 °C. Subsequently, the headspace will be gathered from that Nalophan bag of every sample into the thermal desorption tube, namely Tenax/Carbograph, so that it is up to 1000ml total volume. Moreover, one more thermal desorption tube that contain no sample using similar circumstances of sampling placed into nalophan bag that was applied by sampling process. These study tubes were blanks; this is done to detect all produced compounds from sweat. During the study, 15 persons of different genders and ages were tested. The X axis signifies time measurement (minutes) needed for tube to stated VOC recognition machine, whereas Y axis refer to detected VOC condensation. Table 1 summaries set of 15 VOCs designated out of 198 gas that was emitted from study subject for using as input data to the classifier for the stated human identification aim [18]. In Table 1 below only 15 gases are picked for human detection and identification procedure based on multiple feature selection algorithms applied [18].

Table 1. 15 Selected gases to be used for person detection

<table>
<thead>
<tr>
<th>No</th>
<th>VOCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Acetaldehyde (*)</td>
</tr>
<tr>
<td>2</td>
<td>2-Propenal</td>
</tr>
<tr>
<td>3</td>
<td>Methacrolein</td>
</tr>
<tr>
<td>4</td>
<td>Furan</td>
</tr>
<tr>
<td>5</td>
<td>Furan, 2-pentyl</td>
</tr>
<tr>
<td>6</td>
<td>Acetone</td>
</tr>
<tr>
<td>7</td>
<td>2-Butanone</td>
</tr>
<tr>
<td>8</td>
<td>3-Penten-2-one</td>
</tr>
<tr>
<td>9</td>
<td>2-Pentanone</td>
</tr>
<tr>
<td>10</td>
<td>Furan, 2-pentyl</td>
</tr>
<tr>
<td>11</td>
<td>5-Hepten-2-one, 6-methyl-</td>
</tr>
<tr>
<td>12</td>
<td>Ethanol</td>
</tr>
<tr>
<td>13</td>
<td>Propanal, 2-methyl-</td>
</tr>
<tr>
<td>14</td>
<td>Butanal, 3-methyl-</td>
</tr>
<tr>
<td>15</td>
<td>2,3 Butanediene</td>
</tr>
</tbody>
</table>

B. Common Techniques in Data Classification

Probabilistic Methods

Probabilistic methods are a group of classification techniques marked by the fact that their output is a probabilistic measure of similarity between the object under test and some hypothesized class or classes. The output of this kind of measure compared with a hard decision or unconstrained score has numerous advantages. Initially the attraction for such a measure is that it is easily understood by those using such a system; however, the advantages extend beyond the ease of human utility [33]. The most popular among Probabilistic methods refer to as Bayes classifier that well known as a generative model.

\[
P(Y(T) = i|x_1 ... x_d) = \frac{P(x_1 ... x_d|Y(T) = i).P(Y(T) = i)}{P(x_1 ... x_d)} \tag{1}
\]

Alternative probabilistic method is commonly known as discriminative model, Logistic regression is a prevalent classifier of this type, here the posterior probability is directly modeled through discriminative function learner that map the input data directly to the output class, with the target of immediately approximate the posterior probability \(P(Y(T) = i|X)\) from the training data. Equation 2 below illustrate the logistic regression model: \([1, 2]\)

\[
P(x_1 ... x_d|Y(T) = i) = \prod_{i=1}^{d} P(x_j|Y(T) = i) \tag{2}
\]

where \(\theta\) required to estimate vector

Decision Trees

Decision Tree classify the instance by sorting them down the tree from the root to some leaf node, which provide the classification of the instance. Each node in the tree specify a test of some attributes of the instance, and each branch descending from that node correspond to one of the possible values for this attribute. An instance is classified by starting from the root node of the tree, testing the attribute specified by this node then moving down the tree branch corresponding...
to the value of the attribute [34].

Rule-Based Methods

Rule-based systems have been extensively used as an effective way to store knowledge and to do logic inference. Furthermore, based on the given inputs and the rule database, we can manipulate the stored knowledge for interpreting the generated outputs as well as for decision making. Particularly, rules and rule-based classification systems have been widely applied in various expert systems, such as fault diagnosis for aerospace and manufacturing, medical diagnosis [35].

Instance-Based Learning

In its most basic form it assigns the label of its nearest neighbor to an observation data. Hence, all we need to implement it is a distance measure between pairs of observations. Note that this distance need not even be symmetric. This means that nearest neighbor classifiers can be extremely flexible [35].

Support Vector Machines SVM Classifiers

SVMs are set of related supervised learning methods used for classification and regression. They belong to a family of generalized linear classification. SVM map input vector to a higher dimensional space where a maximal separating hyperplane is constructed. Two parallel hyperplanes are constructed on each side of the hyperplane that separate the data. The separating hyperplane is the hyperplane that maximize the distance between the two parallel hyperplanes. An assumption is made that the larger the margin or distance between these parallel hyperplanes the better the generalization error of the classifier will be [35].

Ensemble classifier

Here, a combination of models to increase accuracy and Combine a series of learned models, with the aim of creating an improved model [35].

Nearest neighbor Classifier

Refer to as k-nearest neighbor’s algorithm (k-NN) here the classification achieved based on closest(nearest) training samples in the feature space. K-NN is a type of instance-based learning, or lazy learning. KNN may apply for regression [35].

C. Classifiers Input Data Set

The input dataset for all different classifier methods contains 15 Gases obtained from 15 persons [7]. Table 1 shows all stable VOCs emitted by human body. The actual input dataset for all classifiers is shown in Table 2 in which 15 persons’ data are applied as input to the classifiers and each person has 15 gases.

In this paper, several settings will be used to investigate the effects of replacing missing values of Volatile Organic Compounds (VOC) that represent gases emitted from human on the accuracy of classifying individual person is investigated. Here are the settings:

a) Classifiers are tested using full datasets of 15 VOCs gases for 15 individuals. The best set of classifiers will be selected for the following tasks.

b) Using only three main classifiers (having 100% detection accuracy) chosen from Step 1, the effects of replacing missing values of Volatile Organic Compounds (VOC) using 0, 1, and Random number between 0 and 1, on the detection accuracy will be investigated.

c) The investigation will be conducted by replacing gases number 15 to 11 first, then replacing gases number 1 to 5 next and finally replacing gases number 6 to 10.

Table 2. List of 15 VOCs emitted from individuals

<table>
<thead>
<tr>
<th></th>
<th>Person1</th>
<th>Person2</th>
<th>Person3</th>
<th>Person4</th>
<th>Person5</th>
<th>Person6</th>
<th>Person7</th>
<th>Person8</th>
<th>Person9</th>
<th>Person10</th>
<th>Person11</th>
<th>Person12</th>
<th>Person13</th>
<th>Person14</th>
<th>Person15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gas</td>
<td>1.1</td>
<td>1.2</td>
<td>1.2</td>
<td>2.1</td>
<td>2.9</td>
<td>4.1</td>
<td>16.3</td>
<td>7.8</td>
<td>10.0</td>
<td>8.0</td>
<td>13.8</td>
<td>198.3</td>
<td>25.7</td>
<td>17.9</td>
<td>16.3</td>
</tr>
<tr>
<td>Gas</td>
<td>0.4</td>
<td>0.3</td>
<td>0.3</td>
<td>0.0</td>
<td>0.6</td>
<td>0.6</td>
<td>0.1</td>
<td>1.0</td>
<td>1.1</td>
<td>0.9</td>
<td>15.5</td>
<td>3.3</td>
<td>2.3</td>
<td>0.8</td>
<td>0.0</td>
</tr>
<tr>
<td>Gas</td>
<td>0.3</td>
<td>0.8</td>
<td>0.6</td>
<td>0.3</td>
<td>0.4</td>
<td>0.2</td>
<td>1.3</td>
<td>1.7</td>
<td>0.9</td>
<td>1.6</td>
<td>0.9</td>
<td>23.7</td>
<td>10.4</td>
<td>8.5</td>
<td>2.9</td>
</tr>
<tr>
<td>Gas</td>
<td>16.5</td>
<td>14.7</td>
<td>17.0</td>
<td>25.2</td>
<td>35.2</td>
<td>25.4</td>
<td>25.9</td>
<td>32.5</td>
<td>52.4</td>
<td>43.7</td>
<td>24.5</td>
<td>25.4</td>
<td>41.0</td>
<td>9.6</td>
<td>10.1</td>
</tr>
<tr>
<td>Gas</td>
<td>0.4</td>
<td>2.2</td>
<td>0.4</td>
<td>2.3</td>
<td>0.3</td>
<td>1.3</td>
<td>0.8</td>
<td>3.4</td>
<td>1.8</td>
<td>0.1</td>
<td>12.0</td>
<td>23.1</td>
<td>9.1</td>
<td>3.8</td>
<td>1.9</td>
</tr>
<tr>
<td>Gas</td>
<td>1.3</td>
<td>0.5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>38.8</td>
<td>6.0</td>
<td>2.0</td>
<td>1.8</td>
<td>1.8</td>
</tr>
<tr>
<td>Gas</td>
<td>0.6</td>
<td>1.5</td>
<td>1.3</td>
<td>1.2</td>
<td>3.9</td>
<td>0.6</td>
<td>1.2</td>
<td>1.9</td>
<td>10.7</td>
<td>5.4</td>
<td>4.2</td>
<td>19.5</td>
<td>22.4</td>
<td>7.3</td>
<td>4.0</td>
</tr>
<tr>
<td>Gas</td>
<td>1.7</td>
<td>2.5</td>
<td>1.2</td>
<td>1.6</td>
<td>14.4</td>
<td>6.5</td>
<td>2.8</td>
<td>2.2</td>
<td>28.7</td>
<td>12.8</td>
<td>2.4</td>
<td>0.0</td>
<td>30.4</td>
<td>9.4</td>
<td>3.1</td>
</tr>
<tr>
<td>Gas</td>
<td>3.2</td>
<td>0.9</td>
<td>1.9</td>
<td>1.8</td>
<td>0.7</td>
<td>2.6</td>
<td>1.3</td>
<td>2.1</td>
<td>6.2</td>
<td>0.8</td>
<td>2.6</td>
<td>18.5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Gas</td>
<td>10.2</td>
<td>1.2</td>
<td>1.1</td>
<td>3.4</td>
<td>0.6</td>
<td>2.9</td>
<td>4.6</td>
<td>0.8</td>
<td>1.2</td>
<td>2.3</td>
<td>3.2</td>
<td>5.6</td>
<td>13.5</td>
<td>98.8</td>
<td>1.1</td>
</tr>
<tr>
<td>Gas</td>
<td>11.0</td>
<td>0.2</td>
<td>0.0</td>
<td>0.2</td>
<td>0.2</td>
<td>0.4</td>
<td>0.0</td>
<td>0.6</td>
<td>0.6</td>
<td>0.3</td>
<td>0.6</td>
<td>0.7</td>
<td>0.3</td>
<td>0.0</td>
<td>2.2</td>
</tr>
<tr>
<td>Gas</td>
<td>12.2</td>
<td>2.1</td>
<td>0.0</td>
<td>2.0</td>
<td>12.3</td>
<td>2.3</td>
<td>15.9</td>
<td>3.4</td>
<td>10.5</td>
<td>21.6</td>
<td>2.6</td>
<td>19.2</td>
<td>25.1</td>
<td>32.7</td>
<td>10.4</td>
</tr>
<tr>
<td>Gas</td>
<td>13.0</td>
<td>0.4</td>
<td>1.0</td>
<td>9.6</td>
<td>3.8</td>
<td>0.0</td>
<td>22.8</td>
<td>12.7</td>
<td>46.6</td>
<td>56.1</td>
<td>12.5</td>
<td>0.0</td>
<td>43.1</td>
<td>39.5</td>
<td>46.0</td>
</tr>
<tr>
<td>Gas</td>
<td>14.0</td>
<td>29.9</td>
<td>8.2</td>
<td>21.6</td>
<td>36.2</td>
<td>4.6</td>
<td>19.3</td>
<td>0.0</td>
<td>38.5</td>
<td>18.0</td>
<td>23.4</td>
<td>0.0</td>
<td>0.0</td>
<td>10.4</td>
<td>10.1</td>
</tr>
<tr>
<td>Gas</td>
<td>15.0</td>
<td>0.2</td>
<td>2.0</td>
<td>2.6</td>
<td>23.5</td>
<td>70.6</td>
<td>0.2</td>
<td>6.3</td>
<td>0.3</td>
<td>29.0</td>
<td>5.0</td>
<td>0.6</td>
<td>0.0</td>
<td>0.0</td>
<td>8.3</td>
</tr>
</tbody>
</table>

IV. RESULTS AND DISCUSSION

A. Classification Results using Complete Dataset

Fifteen set of VOCs odours obtained from 15 persons are applied to be classified using different classification techniques. The performance accuracies of these classifiers are shown in Table 3 below. Based on Table 3 and Fig 2, it was found that the SVM techniques (various models) produced a more consistent high result for the human odour detection, while KNN classification is the poorest algorithm in detection human from odour. The actual output results obtained from different Classifiers can be referred Table 4. Based on Table 3, three classifiers having 100% detection accuracies are selected to be used in investigating the effects of replacing missing values of Volatile Organic Compounds (VOC) on the accuracy of detecting human using odours. In table 4 below, YES means the person is detected by the classification method, NO means the person did not detected by the classification method. Only Person1 detected by all
classifiers
Based on the results obtained shown in Fig. 1, three
classifiers having 100% detection accuracies will be
selected to be used in investigating the effects of replacing
missing values of Volatile Organic Compounds (VOC).
These algorithms include Quadratic SVM, Fine KNN and
Ensemble Bagged Trees.

Table 3. List of 15 VOCs emitted from individuals

<table>
<thead>
<tr>
<th>Classification Techniques</th>
<th>Models</th>
<th>Acc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree Classification</td>
<td>Fine Tree</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Medium Tree</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Coarse Tree</td>
<td>40</td>
</tr>
<tr>
<td>SVM Classification</td>
<td>Linear</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>Quadratic</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Cubic SVM</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Gaussian SVM</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Medium Gaussian</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Coarse Gaussian</td>
<td>100</td>
</tr>
<tr>
<td>KNN Classification</td>
<td>Fine KNN</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Medium KNN</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Coarse KNN</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Cosine KNN</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Cubic KNN</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Weighted KNN</td>
<td>100</td>
</tr>
<tr>
<td>Ensemble Classification</td>
<td>Ensemble Boosted Tress</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Ensemble Bagged Tress</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Ensemble Subspace Discriminant</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Ensemble Subspace KNN</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Ensemble RUSBoosted Tress</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 4. Output results obtained from different Classifiers

Fig. 1. Output Results Chart obtained from different Classifiers

B. Classification Results using Incomplete Dataset
When emitted gases cannot be extracted, this will result in
missing values for some of the VOCs gases, in which in most
cases it may happen. Hence, the 15 gases dataset will consist
of missing values, even when we have decided to choose the
short listed 15 gases [7]. Replacing the missing gases with
other values such as 1 or 0 or random number between 0 and
1 will be another option. With these options, several
classification algorithms that manage to produce 100% correct
detection of classification will be used. These
algorithms include Quadratic SVM, Fine KNN and Ensemble
Bagged Trees as shown in Fig. 2.
The first investigation is to determine appropriate value
used to replace the missing gas. In this work, three gas values
Gas13, Gas14 and Gas15 (for gases {Propanal, 2-methyl-},
{Butanal, 3-methyl-}, {2,3 Butanediol}) will be replaced
with new values, value of 0, value of 1 and value of random
between (0,1) respectively. Table 5, Table 6 and Table 7
show our results of replacing the missing gases. We also
replaced the gases from gas 11 to gas 15 and used value 0 on
missing gases the results in Table 5, we also used value 1 on
missing gases the results in Table 6, and finally we used value
Random between (0,1) on missing gases the results in Table
7. Based on results shown in Tables 5, 6 and 7, we have
decided to replace the missing gases with random number
between (0, 1), and the recommended algorithm is
Ensemble Bagged for the incomplete samples. Gas
replacement procedure is shown in Table 8, variety of
combinations of missing gases were tested.
Table 8 shows the Gas Hiding procedure, here we
investigate the missing of consecutive gases from the
samples (Row 1 to row 14), and we also investigate missing
of non-consecutive gases from the samples (Row 15 to row
21). The results of classification column are computed based...
on Ensemble Bagged Tress which succeeded to detect better than Quadratic SVM, Fine KNN.

![Figure 2. Output Results Chart obtained from actual correct odour samples](image)

**Table 5. Accuracy performances when replacing missing gases with 0 (Missing gases No: 11 – 15)**

<table>
<thead>
<tr>
<th>Number Missing Gases</th>
<th>Performance Accuracy (%)</th>
<th>Quadric SVM</th>
<th>Fine KNN</th>
<th>Ensemble Bagged Tress</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 missing gases</td>
<td>87%</td>
<td>60%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>3 missing gases</td>
<td>60%</td>
<td>27%</td>
<td>73%</td>
<td></td>
</tr>
<tr>
<td>4 missing gases</td>
<td>60%</td>
<td>27%</td>
<td>73%</td>
<td></td>
</tr>
<tr>
<td>5 missing gases</td>
<td>53%</td>
<td>20%</td>
<td>47%</td>
<td></td>
</tr>
</tbody>
</table>

Table 6. Accuracy performances when replacing missing gases with 1 (Missing gases No: 11 – 15)

<table>
<thead>
<tr>
<th>Number Missing gases</th>
<th>Performance Accuracy (%)</th>
<th>Quadric SVM</th>
<th>Fine KNN</th>
<th>Ensemble Bagged Tress</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 missing gases</td>
<td>80%</td>
<td>40%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>3 missing gases</td>
<td>80%</td>
<td>27%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>4 missing gases</td>
<td>47%</td>
<td>100%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 missing gases</td>
<td>53%</td>
<td>27%</td>
<td>73%</td>
<td></td>
</tr>
</tbody>
</table>

**Table 7. Accuracy performances when replacing missing gases with random number from 0 to 1 using Ensemble Bagged Tress (Missing gases No: 11 – 15 in Table 1)**

<table>
<thead>
<tr>
<th>Number Missing Gases</th>
<th>Performance Accuracy (%)</th>
<th>Quadric SVM</th>
<th>Fine KNN</th>
<th>Ensemble Bagged Tress</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 missing gases</td>
<td>80%</td>
<td>40%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>3 missing gases</td>
<td>73%</td>
<td>27%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>4 missing gases</td>
<td>60%</td>
<td>27%</td>
<td>87%</td>
<td></td>
</tr>
</tbody>
</table>

Based on results shown in Tables 9, it can be concluded the following list of gases shown in Table 10 below can be best replaced with random number from 0 to 1 and the recommended algorithm is Ensemble Bagged for the incomplete samples. Based on results obtained from Table 11, the following list of gases shown in Table 12 below can be replaced with random number from 0 to 1 and the recommended algorithm is Ensemble Bagged for the incomplete samples.

**Table 8. Gases replacement procedure**

<table>
<thead>
<tr>
<th>Inputs</th>
<th>Result of classification (Ensemble Bagged Trees)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Hide gas 1</td>
<td>100 %</td>
</tr>
<tr>
<td>2 Hide gas 1 +2</td>
<td>100 %</td>
</tr>
<tr>
<td>3 Hide gas 1 +2 +3</td>
<td>100 %</td>
</tr>
<tr>
<td>4 Hide gas 1 +2 +3 +4</td>
<td>100 %</td>
</tr>
<tr>
<td>5 Hide gas 1 +2 +3 +4 + 5</td>
<td>100 %</td>
</tr>
<tr>
<td>6 Hide gas 1 +2 +3 +4 + 5 + 6</td>
<td>System Fail</td>
</tr>
<tr>
<td>7 Hide gas 6</td>
<td>100 %</td>
</tr>
<tr>
<td>8 Hide gas 6 +7</td>
<td>100 %</td>
</tr>
<tr>
<td>9 Hide gas 6 +7 +8</td>
<td>100 %</td>
</tr>
<tr>
<td>10 Hide gas 6 +7 +8 +9</td>
<td>100 %</td>
</tr>
<tr>
<td>11 Hide gas 6 +7 +8 +9 +10</td>
<td>System Fail</td>
</tr>
<tr>
<td>12 Hide gas 15</td>
<td>100 %</td>
</tr>
<tr>
<td>13 Hide gas 14 +15</td>
<td>100 %</td>
</tr>
<tr>
<td>14 Hide gas 13 +14 +15</td>
<td>System Fail</td>
</tr>
<tr>
<td>15 Hide gas 1+6</td>
<td>100 %</td>
</tr>
<tr>
<td>16 Hide gas 1+6+15</td>
<td>100 %</td>
</tr>
<tr>
<td>17 Hide gas 2+7+14</td>
<td>100 %</td>
</tr>
<tr>
<td>18 Hide gas 1+2+6+15</td>
<td>100 %</td>
</tr>
<tr>
<td>19 Hide gas 1+2+7+14</td>
<td>100 %</td>
</tr>
<tr>
<td>20 Hide gas 1+2+7+14+15</td>
<td>100 %</td>
</tr>
<tr>
<td>21 Hide gas 1+2+6+7+14+15</td>
<td>System Fail</td>
</tr>
</tbody>
</table>

**Table 9. Accuracy performances when replacing missing gases with random number from 0 to 1 (Missing gases number 1 to 5 in Table 1)**

<table>
<thead>
<tr>
<th>Number Missing Gases</th>
<th>Performance Accuracy (%)</th>
<th>Quadric SVM</th>
<th>Fine KNN</th>
<th>Ensemble Bagged Tress</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 missing gases</td>
<td>93%</td>
<td>100%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>3 missing gases</td>
<td>93%</td>
<td>93%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>4 missing gases</td>
<td>67%</td>
<td>93%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>5 missing gases</td>
<td>67%</td>
<td>93%</td>
<td>100%</td>
<td></td>
</tr>
</tbody>
</table>

Based on results shown in Tables 9, it can be concluded the following list of gases shown in Table 10 below can be best replaced with random number from 0 to 1 and the recommended algorithm is Ensemble Bagged for the incomplete samples. Based on results obtained from Table 11, the following list of gases shown in Table 12 below can be replaced with random number from 0 to 1 and the recommended algorithm is Ensemble Bagged for the incomplete samples.

**Table 10. List of gases replaced with random number between 0 and 1 in case of missing values**

<table>
<thead>
<tr>
<th>No</th>
<th>VOCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Acetaldehyde (*)</td>
</tr>
<tr>
<td>2</td>
<td>2-Propenal</td>
</tr>
<tr>
<td>3</td>
<td>Methacrolein</td>
</tr>
<tr>
<td>4</td>
<td>Furan</td>
</tr>
</tbody>
</table>
Table 11. Summary of accuracy performances obtained by replacing missing gases with random number from 0 to 1 using Ensemble Bagged Tress (Missing gases number 6 to 10 in Table 1)

<table>
<thead>
<tr>
<th>Number</th>
<th>Missing gases</th>
<th>Ensemble Bagged Trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>40%</td>
<td></td>
</tr>
</tbody>
</table>

Table 12. List of Gases replaced with random number between 0 and 1 in case of missing

<table>
<thead>
<tr>
<th>No</th>
<th>VOCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Acetone</td>
</tr>
<tr>
<td>7</td>
<td>2-Butanone</td>
</tr>
<tr>
<td>8</td>
<td>3-Penten-2-one</td>
</tr>
<tr>
<td>9</td>
<td>2-Pentanone</td>
</tr>
</tbody>
</table>

Table 13. Accuracy performances obtained by replacing missing gases with random number from 0 to 1 using Ensemble Bagged Tress (Missing gases number 15 to 11 in Table 1)

<table>
<thead>
<tr>
<th>Number</th>
<th>Missing Gases</th>
<th>Ensemble Bagged Trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>87%</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>73%</td>
<td></td>
</tr>
</tbody>
</table>

Table 14. List of Gases replaced with random number between 0 and 1 in case of missing

<table>
<thead>
<tr>
<th>No</th>
<th>VOCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>Propanal, 2-methyl-</td>
</tr>
<tr>
<td>14</td>
<td>Butanal, 3-methyl-</td>
</tr>
<tr>
<td>15</td>
<td>2,3 Butanediione</td>
</tr>
</tbody>
</table>

V. CONCLUSION

In short, in this paper, the effects of replacing missing values of Volatile Organic Compounds (VOC) that represent gases emitted from human on the accuracy of classifying individual person is investigated. It was found that the best human identification results are obtained when using the Quadratic SVM, Fine KNN and Ensemble Bagged Trees. All these classifiers have produced 100% accuracies identifying human based on odour using a complete set Volatile Organic Compounds (VOC) obtained from the odour samples. Further inspection has been carried out to measure the attainment of three classifiers Quadratic SVM, Fine KNN and Ensemble Bagged Tress on incomplete odour samples. The maximum and minimum allowable numbers of consecutive missing gases in human odour samples in reference to human emitted gases are 4 and 3. This is obtained when top 4 gases listed in Table 1 are missing, 4 consecutive missing gases in the middle of gas list in Table 1, and bottom 3 gases listed in Table 1 are also missing. For non-consecutive gas missing may reach 5 gas with full accurate detection. The best accuracy result when missing values are introduced in the odour dataset is the ensemble Bagged Tress. And the best replacement value of missing gases is random number between 0 and 1. For all types of decision tree models, the performance accuracy results obtained are 40%. These results indicate that handling missing values using decision trees are not suitable and recommended. Most ensemble techniques work based on hypothesis, this hypothesis, however, is not necessarily contained within the hypothesis space of the models from which it is built, thus some ensembles such as (Ensemble Boosted Tress, Ensemble Subspace Discriminant, Ensemble RUSBoosted Tress) have produced very bad results such as 7% only based on mismatch of the used hypothesis, while some ensembles have produced 100% accuracy such as (Ensemble Bagged Tress, Ensemble Subspace KNN). Ensembles tend to yield better results when there is a significant diversity among the models. In our case of data samples, we may not have noticeable diversity among the samples, some of the samples are very close and similar.

REFERENCES

Effects Of Handling Missing Values Of VOCS Gases Emitted From Human For Human Detection