Feature Selection Method to Improve the Accuracy of Diabetes Mellitus Detection Instrument

Sari Ayu Wulandari\textsuperscript{1}, Sutikno Madnasri\textsuperscript{2}, Ratih Pramitasari\textsuperscript{3}, Susilo\textsuperscript{4}

\textsuperscript{1}Department of Biomedical Engineering, Dian Nuswantoro University, Semarang, Indonesia
\textsuperscript{2, 4}Department of Physic, Semarang State University, Semarang, Indonesia
\textsuperscript{3}Department of Public Health, Dian Nuswantoro University, Semarang, Indonesia

sari.wulandari@dsn.dinus.ac.id, sutikno@mail.unnes.ac.id, ratih.pramitasari@dsn.dinus.ac.id, susilo@mail.unnes.ac.id

\textbf{Article History}
Received Oct 30\textsuperscript{th}, 2020
Revised Dec 14\textsuperscript{th}, 2020
Accepted Dec 15\textsuperscript{th}, 2020
Published Dec, 2020

\textbf{Abstract}—The need for aroma recognition devices or often known as enose (electronic nose), is increasing. In the health field, enose can detect early diabetes mellitus (DM) type 2 from the aroma of urine. Enose is an aroma recognition tool that uses a pattern recognition algorithm to recognize the urine aroma of diabetics based on input signals from an array of gas sensors. The need for portable enose devices is increasing due to the increasing need for real-time needs. Enose devices have an enormous impact on the choice of the gas sensor array in the enose. This article discusses the effect of the number of sensor arrays used on the recognition results. Enose uses a maximum of 4 sensors, with a maximum feature matrix. After that, the feature matrix enters the PCA (Principal Component Analysis) feature extraction and clustering using the FCM (Fuzzy C Means) method. The number of sensors indicates the number of features. Enose using method for feature selection, it’s a variation from 4 sensors, where experiment 1 uses 4 sensors, experiment 2 uses a variation of 3 sensors and experiment 3 uses a variation of 2 sensors. Especially for sensors 3 and 4 using feature extraction method, PCA (Principal Component Analysis), to reduce features to only 2 best features. As for the variation of 2 sensors use primer feature matrix. After feature selection, the number of features is 2 out of 11 variations. Next, do the grouping using the FCM (Fuzzy C Means) method. The results show that using two sensors has a high accuracy rate of 92.5%.

\textbf{Keywords}—Number_of_array_sensor; Sensor_MQ; Enose
1 INTRODUCTION

Diabetes has become a disease that causes complications due to high sugar levels in the blood. Diabetes is a disease which the body cannot produce insulin (a blood sugar-regulating hormone), or produced insufficient insulin or insulin does not work correctly. Diabetes detecting using urine has characteristics which is distinctive odor that is the smell of acetone / C3H6O. When urine decomposes, an unpleasant ammonia odor arises. Chemical levels contained in the urine influences the smell of urine. Diabetes detecting using urine is often used as a reference for some test kits, for example, is urinalysis glucose diabetes strips urine test strip pack. This instrument detects like a pregnancy test strip. The use of a test strip is for one test, and the price is high, insensitive, and inaccurate. In a realtime system, we need an instrument that can detect blood sugar levels that are more sensitive, easy to use, and accurate.

Enose is new hope for diabetes mellitus (DM) detector. One of the substantial advantages of E-nose detection is the ability to adjust the chemical sensitivity of individual E-nose sensors in the sensor array to adjust the instrument to a chemical detection range that is very specific to VOCs in certain chemical classes, or even to a single component when this is sufficient to detect metabolic events or certain physiologically highly correlated with the release of the compound. E-specific nose applications with specific reference databases can focus on a very narrow range of analytes to simplify the detection of certain diseases or determine the health status of organ metabolism in the body. Liu [1], evaluate the detection of a single compound (acetone in human breath) as a promising diagnostic and non-invasive method for monitoring diabetes. A portable E-nose sensor, consisting of a nano structured film, 10% mol Si-mol, developed with a miniature sample chamber volume, senses optimized temperatures for the limit of low detection acetone (~ 20 ppb), and short response (10–15 seconds) and recovery time (35–70 seconds). Durable sensor (response) signal to be able to detect and monitor acetone levels continuously at varying air flow rates and a realistic relative humidity range (80-90%) in human breath.

Enose is an instrument that can feel like a human’s sense of smell. Enose uses gas sensors that are sensitive to volatile organic compounds (VOC) in the air. The gas sensors arranged in an array, have distinctive patterns that distinguish between scents. The scents are entered into the expert system for data training and will produce a model from the PARC (Pattern Recognition) device. In some DM detection studies using enose, there are several variations in the number of sensors, both of which use commercial gas sensors [1][2][3][4][5] or handmade sensor gas [6][7][8][9]. The number of sensors also varies, starting from 6 sensors [3], 7 sensor [4], 8 sensor [1][2] dan 12 sensor [5]. The sensors mostly use Figaro metal oxide sensors [1][3][4][5] and MQ [2]. Some enose portable applications will optimize the function of the sensor array, by reducing the number of sensors but with good performance.

One of the methods used by previous researchers is the feature selection method. Feature selection reduces the size of feature vectors, simplifies computation and increases recognition speed. Yon Yin et al. Introduced 6 types of features, with 100% accuracy [11]. Using a matrix reduction by taking 51 highly correlated with the release of the features. The best performance is using 3 sensors with KNN (K-Nearest Neighbor) method [12]. This shows that a large number of features does not guarantee an increase in recognition performance. The three methods from recent study, RBFNN, PNN and KNN use complex computational approaches. Of course it is not appropriate for enose portable. Research using enose portable for DM detection at first uses 4 sensors, while in this paper it will put to the features selection process, to becomes 2, 3 or 4 sensors that’s still have good recognition performance. Feature selection is an attempt to realize enose portables that are small, light and easy to carry, but still accurate.

Pedro research, using the KNN feature selection method. This method is good to use when the number of sensors is odd, where there is an 8-16% to 96% increase in accuracy. However, this method requires a large memory to store all feature points used as neighbors. In this paper, the feature selection process uses a variety of sensors, combined with the FCM cluster method. The results of the recognition are the number of sensors, the type of sensor used, and the cluster midpoint that can be easily embedded into Enose. This method uses a simple computational approach, saves memory and is more suitable for enose portable design.

2 METHOD

Enose uses a gas sensor array, where the selection of gas sensor variations is based on the number of molecules supporting the aroma. Several things are used as a standard in the selection of sensors, that is based on the primary material constituent, based on size, based on the type of constituent substances and based on the number of sensors.

1. Based on the basic constituent materials :
   a. Polimer [13][14] : the commercial one is Figaro (TGS) made in Japan [15][16] and MQ made in China [17].
   b. Metallic [13][18]
   c. Carbon Conductor [19][20]

2. Based on size:
   a. Nano [21][22] : widely used in terms of sensor durability and stability. Some scents which have a low density can also be identified properly using a nanosensor.
   b. Mikro [23]

3. Based on the type of constituent substances
   a. organic (Biosensor) [25][26][27]
   b. non-organic (Chemical Sensor) [28][22][29].
Table 1. Research Accomplished

<table>
<thead>
<tr>
<th>Sensors</th>
<th>Characterization</th>
<th>Accuracy</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>Six array sensor, types of gas sensors these are TGS 822, TGS 825, TGS 816, TGS 2620, TGS 2610, TGS 2611</td>
<td>Gas exhalation (standard deviation value)</td>
<td>[3]</td>
<td></td>
</tr>
<tr>
<td>Six array sensor, TGS825, TGS826, TGS822, TGS813, TGS8260, and TGS2611</td>
<td>Gas exhalation (standard deviation value)</td>
<td>[24]</td>
<td></td>
</tr>
<tr>
<td>Seven array sensor, TGS822, TGS2620, QS-01, TGS821, TGS2620, TGS8262, TGS826, TGS2610</td>
<td>Gas exhalation (Mean value)</td>
<td>[4]</td>
<td></td>
</tr>
<tr>
<td>Eight array sensor MOX</td>
<td>Gas exhalation (Mean value)</td>
<td>The accuracy of sample identification on fasting is 85%, and the accuracy rate is up to 98% one hour after the meal, and it is 92% two hours after the meal.</td>
<td>[1]</td>
</tr>
<tr>
<td>Eight array sensors, MQ3, MQ35, TGS2600, TGS2611, MQ2, MQ317, MQ7, and TGS822</td>
<td>Urine (Mean value)</td>
<td>[2]</td>
<td></td>
</tr>
<tr>
<td>12 array sensor, TGS2600, TGS2602, TGS2611, TGS2610-C00, TGS2610-D00, TGS2620, TGS825, TGS4161, TGS826, TGS2201, TGS822, TGS821</td>
<td>Gas exhalation (Median value)</td>
<td>[5]</td>
<td></td>
</tr>
<tr>
<td>Field-Asymmetric Ion Mobility Spectrometry (FAIMS) and FOX 4000 (Alpha M.O.S. Toulouse, Francis)</td>
<td>Urine (Maximum)</td>
<td>[6]</td>
<td></td>
</tr>
<tr>
<td>PIMA INDIAN DIABETES dataset and has been collected from UCI machine learning repository</td>
<td>Blood (minimum and maximum value)</td>
<td>[7]</td>
<td></td>
</tr>
<tr>
<td>Field-Asymmetric Ion Mobility Spectrometry (FAIMS)</td>
<td>Urine (mean and median value)</td>
<td>It was discovered that gas emissions (concentration and diversity) reduced over time. However, there was less variation in the initial nine months of storage with more excellent uniformity and stability of concentrations together with tighter clustering of the total number of chemicals released. It suggests that nine months could be considered a general guide to a sample shelf-life.</td>
<td>[8]</td>
</tr>
</tbody>
</table>

4. Based on the number of sensors
   a. 3 sensors [13][15][18]
   b. 4 sensors [30][31]
   c. and multi-sensor [16], that is 6 sensor [3], 7 sensor [4], 8 sensor [1][2] dan 12 sensor [5].

The electronic nose is an instrument that uses several sensor elements that can be operated to offer information about that category, as well as odor intensity. In this sense, a sensor includes two sensor elements in the electronic nose. This instrument is a desktop, not a portable type in many cases. Due to the increasing number of sensor elements, some previous researchers conducted several experiments using more than two sensor arrays. It is shown in Table 1.

The number of sensors most commonly used is the number of even sensors, 6 and 8 sensors. The use of sensor is not proportional to the amount of accuracy. The level of accuracy depends on the selection of the right sensor, if the sensor is right, the use of a few sensors will be more accurate than many sensors.

In the experiments prepared are urine samples from 60 urine data of normal blood sugar levels (starting now referred to as standard data), between 70 to 200 mg / dL and 60 urine data of high blood sugar levels (starting now referred to as DM data which is higher than 200 mg / dL). Urine here functions as an odorant or source of the aroma. Urine is brought close to the sensor for ADC (Analog Digital Converter) values from VOCs (Volatile organic compounds), which are trapped in the sensor, and cause changes in the voltage value of the gas sensor array. ADC value is taken every 1 second 100 data, taken for 1 minute 4 seconds, to produce 1000 sample data in each aroma sample. Data were collected for 60 standard and DM data, respectively. In one ADC value retrieval, it consists of 4 columns because the number of sensors is 4, so the total data is 4 x 120 x 1000. Every 1000 ADC data is taken the maximum value to minimize the dimensions, so the dimension of the data matrix generated from the pre-processing process is 120 x 4.

2.1 Variation Feature Selection Method

Feature selection is a process of finding the resulting features are correlated with each other without using the entire result of the extraction features [10]. The number of sensors indicates the number of features. Enose using method for feature selection, it’s a variation from 4 sensors, where experiment 1 uses 4 sensors, experiment 2 uses a variation of 3 sensors and experiment 3 uses a variation of 2 sensors that’s it mean of variation feature selection method [12]. Especially for sensors 3 and 4 using feature extraction method, PCA, to reduce features to only 2 best features. As for the variation of 2 sensors use primer feature matrix. After feature selection, the number of features is 2 out of 11 variations.
The sensors used in the experiment as well as the gases that can be detected are shown in Table 2. The sensor array used consists of MQ2, MQ5, MQ6, and MQ138. The urine aroma of DM contains acetone compounds. At high levels, the aroma of acetone is very pungent and can even be detected using only organoleptic testing. Acetone is one of the ketone functional groups often called propanone, dimethyl ketone, so the sensor selection is based on propane and ketone gas. The selection of a hydrogen sensor, based on that hydrogen is also used as an acetone propagation carrier in the air.

Table 2. Array Sensor Specification

<table>
<thead>
<tr>
<th>Name</th>
<th>Sensor</th>
<th>Gas Detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensor 1</td>
<td>MQ2</td>
<td>H2, LPG, CH4, CO, Alcohol, Smoke or Propane</td>
</tr>
<tr>
<td>Sensor 2</td>
<td>MQ5</td>
<td>LPG, i-butane, methane, alcohol, Hydrogen, smoke</td>
</tr>
<tr>
<td>Sensor 3</td>
<td>MQ6</td>
<td>LPG, iso-butane, propane.</td>
</tr>
<tr>
<td>Sensor 4</td>
<td>MQ138</td>
<td>Aldehydes, ketones, alcohols</td>
</tr>
</tbody>
</table>

After getting the feature vector from the pre-processing process which is 120 x 4, the next process is selecting the sensor pair, where in the pair of 2 sensors, the data is directly scattered and a feature selection process is carried out, whereas in the 3 and 4 pair sensor the pre-process data results first extracted using the PCA feature extraction. The feature extraction process is carried out so that the data can be represented using 2D cartesian coordinates.

2.2 FCM Cluster Method

Fuzzy C-Means (FCM) is a data clustering technique in which the existence of each data point in a cluster is determined by of membership degree [32]. The FCM result is the midpoint of the cluster that will be embedded into the portable Enose. The two resulting feature points show the location of the DM and non-DM feature points. If there is a new feature point as data testing, then the new test point will be compare for the distance using euclidean distance. The closest distance from two cluster points shows the results of detection. Flowchart of FCM method is shown in Figure 2 [32].

Figure 1. Experiments Scheme

![Experiments Scheme](Image)

The sensors used in the experiment as well as the gases that can be detected are shown in Table 2. The sensor array used consists of MQ2, MQ5, MQ6, and MQ138. The urine aroma of DM contains acetone compounds. At high levels, the aroma of acetone is very pungent and can even be detected using only organoleptic testing. Acetone is one of the ketone functional groups often called propanone, dimethyl ketone, so the sensor selection is based on propane and ketone gas. The selection of a hydrogen sensor, based on that hydrogen is also used as an acetone propagation carrier in the air.

Table 2. Array Sensor Specification

<table>
<thead>
<tr>
<th>Name</th>
<th>Sensor</th>
<th>Gas Detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensor 1</td>
<td>MQ2</td>
<td>H2, LPG, CH4, CO, Alcohol, Smoke or Propane</td>
</tr>
<tr>
<td>Sensor 2</td>
<td>MQ5</td>
<td>LPG, i-butane, methane, alcohol, Hydrogen, smoke</td>
</tr>
<tr>
<td>Sensor 3</td>
<td>MQ6</td>
<td>LPG, iso-butane, propane.</td>
</tr>
<tr>
<td>Sensor 4</td>
<td>MQ138</td>
<td>Aldehydes, ketones, alcohols</td>
</tr>
</tbody>
</table>

After getting the feature vector from the pre-processing process which is 120 x 4, the next process is selecting the sensor pair, where in the pair of 2 sensors, the data is directly scattered and a feature selection process is carried out, whereas in the 3 and 4 pair sensor the pre-process data results first extracted using the PCA feature extraction. The feature extraction process is carried out so that the data can be represented using 2D cartesian coordinates.

2.2 FCM Cluster Method

Fuzzy C-Means (FCM) is a data clustering technique in which the existence of each data point in a cluster is determined by of membership degree [32]. The FCM result is the midpoint of the cluster that will be embedded into the portable Enose. The two resulting feature points show the location of the DM and non-DM feature points. If there is a new feature point as data testing, then the new test point will be compare for the distance using euclidean distance. The closest distance from two cluster points shows the results of detection. Flowchart of FCM method is shown in Figure 2 [32].

Figure 2. Flowchart of FCM Method [32]
3 RESULT AND DISCUSSION

Sensor data retrieval is done using the ADC. The waveforms from the ADC data record results between standard data and DM data each have different forms. The difference in waveforms is shown in Figure 3.

![Figure 3](image1)

(a) Data ADC from Normal (b) Data ADC from DM

Figure 4. Spider Plot

The value of data distribution from the sensor array can be shown through the spider diagram, Figure 3. From the spider graph, it can be seen that there are two pieces of data, namely, standard data (Data 1-30) and DM data (Data 31-60). Sensors 1, 3, and 4 have a different distribution of the 2 data, wherein sensors 1 and 3, standard data are smaller on average than DM data, whereas, on sensor 4, standard data on average are more prominent than data DM. Sensor 2 has a distribution with the same average value.

![Figure 4](image2)

(a) Boxplot For Normal Data (b) Boxplot For DM Data, (c) Boxplot For All Data

The value of data distribution from the sensor array can be shown through the spider diagram, Figure 4. From the spider graph, it can be seen that there are two pieces of data, namely, standard data (Data 1-30) and DM data (Data 31-60). Sensors 1, 3, and 4 have a different distribution of the 2 data, wherein sensors 1 and 3, standard data are smaller on average than DM data, whereas, on sensor 4, standard data on average are more prominent than data DM. Sensor 2 has a distribution with the same average value.

![Figure 5](image3)
In the boxplot graph, both standard data and DM data are 60 x of data retrieval, respectively. In standard data, there is only one minimum outlier in sensor 1, outlier value 773, with eight outliers. In DM data, there are several outliers in sensors 1 and 3, in sensor one, the maximum outlier value is 845 with a total of 7 outliers, while in sensor three the maximum outlier is 707 with several one outliers and a value of 700 with a total of 5 outliers. In the combined data, there is one minimum outlier on sensor 1, outlier value 773, with eight outliers. Figure 5. (c) shows the data distribution from sensor 4. Among the four sensors, sensor four can be used as a single sensor for DM urine detectors, which have a considerable variety of values, with a median value of 987, a maximum value of 1023, and a minimum value of 641 with no outlier.

3.1 Use 2 Features

Feature selection is a suitable method for determining sensor pairs with maximum accuracy from 4 gas sensors used. Six sensor pairs can be varied from 4 sensors. The sensor pairs are sensors 1 and 2, 1 and 3, 1 and 4, 2 and 3, 2 and 4, and 3 and 4. The scatter graph of all sensor pairs is shown in Figure 6. In tests that use more than two sensors, it is performed by using feature extraction, to simplify computing and reduce feature vector dimensions. In the use of 3 sensors, there are four pair variations, namely 1-2-3 sensor pairs, 1-2-4 sensors, 1-3-4 sensors, and 2-3-4 sensor pairs. Scatter graphs of all sensor pairs are shown in Figure 6. In the use of 4 sensors, there is one variation of pairs, namely 1-2-3-4 sensor pairs. A scatter graph of all sensor pairs is shown in Figure 7.

Figure 6. Testing 2 Sensors (a) 1-2, (b) 1-3, (c) 1-4, (d) 2-3, (e) 2-4, (f) 3-
The accuracy of all sensors is shown in Table 3. Almost all sensor array pairs have an accuracy of 92.5%, were from 2 classes; each has one outlier detected incorrectly. However, there is one sensor variation that has lower accuracy, namely sensor pairs 2 and 3, where the accuracy is 87.5%, with five outliers detected incorrectly.

Table 3. Array Sensor Specifications

<table>
<thead>
<tr>
<th>No</th>
<th>Sensors</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>S1-S2</td>
<td>92.5</td>
</tr>
<tr>
<td>2</td>
<td>S1-S3</td>
<td>92.5</td>
</tr>
<tr>
<td>3</td>
<td>S1-S4</td>
<td>92.5</td>
</tr>
<tr>
<td>4</td>
<td>S2-S3</td>
<td>87.5</td>
</tr>
<tr>
<td>5</td>
<td>S2-S4</td>
<td>92.5</td>
</tr>
<tr>
<td>6</td>
<td>S3-S4</td>
<td>92.5</td>
</tr>
<tr>
<td>7</td>
<td>S1-S2-S3</td>
<td>92.5</td>
</tr>
<tr>
<td>8</td>
<td>S1-S2-S4</td>
<td>92.5</td>
</tr>
<tr>
<td>9</td>
<td>S1-S3-S4</td>
<td>92.5</td>
</tr>
<tr>
<td>10</td>
<td>S2-S3-S4</td>
<td>92.5</td>
</tr>
<tr>
<td>11</td>
<td>S1-S2-S3-S4</td>
<td>92.5</td>
</tr>
</tbody>
</table>

From table 3, it can be seen that the addition of the number of sensors is not proportional to the increase in system accuracy. The use of capable sensors can only be done by using only one sensor, sensor four, where sensor 4 has a considerable variation in values, with a median value of 987, a maximum value of 1023, and a minimum value of 641 with no outliers. Whereas the use of 2 sensors can be done using sensor pairs 1-2, 2-3, 1-4 and 2-4.

4 CONCLUSION

The addition of the number of sensors is not directly proportional to the increase in system accuracy. The use of a capable sensor can be done only by using only one sensor, the MQ138 sensor, which has a considerable variation in values, with a median value of 987, a maximum value of 1023, and a minimum value of 641 with no outliers. Whereas the use of 2 sensors can be done using sensor pairs 1-2, 2-3, 1-4, and 2-4.
ACKNOWLEDGMENT

This research was partially funded by the Indonesian Ministry of Research, Technology and Higher Education of the Republic of Indonesia.

REFERENCES


[21] B. Szulczyński et al., “Different ways to apply a measurement instrument of E-nose type to evaluate ambient air quality with respect to odour nuisance in a vicinity of municipal processing plants,” Sensors (Switzerland), vol. 17, no. 11, 2017.


