ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 6, Issue 7, July 2019

Exploration of Electronic-Nose Potential as Diabetes Urine Detection using Machine Learning Algorithms

Syahrul Fadholi GUMELAR,Hanif Mustafa BUDIYANTO,Muhammad Fawwaz MAYDA, Bakhtiar Alldino Ardi SUMBODO

Departement of Computer Science and Electronics, Faculty of Mathematics and Natural Science, Universitas Gadjah Mada – Yogyakarta, Indonesia

ABSTRACT: Diabetes is one of the most common chronic diseases among people. Based on data from the International Diabetic Federation (IDF) in 2017, the number of people with diabetes in Indonesia reached 10.3 million people. Therefore an early diabetes detection is needed based on the measurement of urine odor using an electronics-nose (E-nose). Eight types of gas sensors are used in E-nose, including MQ3, MQ135, TGS2600, TGS2611, MQ2, MQ137, MQ7, and TGS822. The urine sample used is a normal urine sample and a urine sample. For urine diabetes dextrose is added with variations of 100 mg/dL, 500 mg/dL, and 1000 mg/dL. Data processing stages include signal processing, feature extraction, dimension reduction and classification with machine learning algorithms including KNN (K-Nearest Neighbors), SVM (Support Vector Machine), DT (Decision Tree) and RF (Random Forest). The results to classify urine samples, the KNN algorithm has 100% accuracy, SVM algorithm 100%, DT algorithm 96%, and RF algorithm 96%. Therefore it can be concluded that E-nose can classify between normal urine and diabetes urine

KEYWORDS: Electronics-nose, urine, diabetes, classification, machine learning algorithms

1. INTRODUCTION

Electronic-nose (E-nose) is one of the electronic instruments that mimic the performance of the human nose to detect and analyze scents [1]. The e-nose consists of an array of chemical gas sensors that function to capture and convert scents into electrical signals or sensor responses. Then the electrical signal produced will form special patterns to detect the type of aroma that is captured. Then for classifying the pattern, a pattern recognition machine is used [2].

Diabetes is one of the common diseases caused by the pancreas that cannot produce enough insulin (a hormone that regulates glucose levels in the blood) or when the body cannot use insulin effectively [3]. Diabetes itself is also one of the most common chronic diseases among people. Based on data from the International Diabetic Federation (IDF) 8th edition of the 2017 Atlas, the number of diabetics in Indonesia has reached 10.3 million people. The number is predicted to increase to 16.7 million by 2045. Also, the number of diabetics in Indonesia who have never made a diagnosis of diabetes reached 7.6 million people or 73.7% of the total diabetics in Indonesia. The data shows that there is still a lack of public awareness in Indonesia to diagnose diabetes. Diabetes can be diagnosed by measuring glucose levels in the blood. This can be done through three methods, namely random blood sugar test, fasting blood sugar test, and oral glucose tolerance test. But it has weaknesses in the procedure which is quite complicated and time-consuming. Another method for the diagnosis of diabetes quickly is the measurement of glucose in the urine. In general, normal urine contains very little glucose; people with high blood sugar often have glucose in their urine because of diabetes. The effect of high sugar levels in urine can cause sweet smells. Therefore, it is possible to use an electronic nose for the diagnosis of diabetes through direct analysis of urine odor [4]. Therefore, it is made an electronic-nose (E-nose) that can replace the work of the human nose in smell. E-

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 6, Issue 7, July 2019

nose is expected to be used accurately to smell the scents created by urine so that it can distinguish the urine of normal people and urine of diabetics patient.

In general, the way the E-nose works is to read the sensor response from the scents that are being observed; then the sensor response data will be processed to be then carried out pattern recognition analysis using certain algorithms. Many pattern recognition methods have been implemented to identify or classify scents. For example, in predicting the quality of electronic nose-based tea with the fuzzy method [5], and the E-nose application for the introduction of beer using PCA, RBF, and backpropagation [6]. The difference in the method of recognition of the aroma pattern is due to the absence of a standard method for each sample scents measurement. Some studies on diabetes detection through urine using multilayer perceptron learning methods obtained results of the success rate of the program identification of 100% non-diabetes, and 30% diabetes [3] and other studies using the PCA and CA methods [7].

Based on the problem of the inaccuracy of normal and diabetes urine classification, this study focuses on the potential of E-nose to detect diabetes urine using machine learning algorithms, including k-Nearest Neighbour (KNN), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF) and then look for which algorithms have the highest accuracy to classify normal urine and diabetes urine..

II. RESEARCH METHODS

A. SYSTEM ANALYSIS

The E-nose system used in this study, as shown in*Figure 1.* consists of a sample room, sensor room, processing and control room, and data analysis. The sample room consists of a sample in the form of urine and heater, which is used to regulate the temperature in the urine. The sensor room consists of a series of gas sensor arrays consisting of MQ3, MQ135, TGS2600, TGS2611, MQ2, MQ137, MQ7, and TGS822. These rows of sensors will later be used to detect the characteristics of the scents patterns produced by urine. In the processing section, there is a microcontroller that accepts analog signal input and converts it into a digital signal. The microcontroller will send the data from the sensor output to the computer through the universal serial bus (USB). This data will then go through the processing stages such as feature extraction, and dimension reduction. Until later, the urine scents data will be classified using the machine learning algorithm.

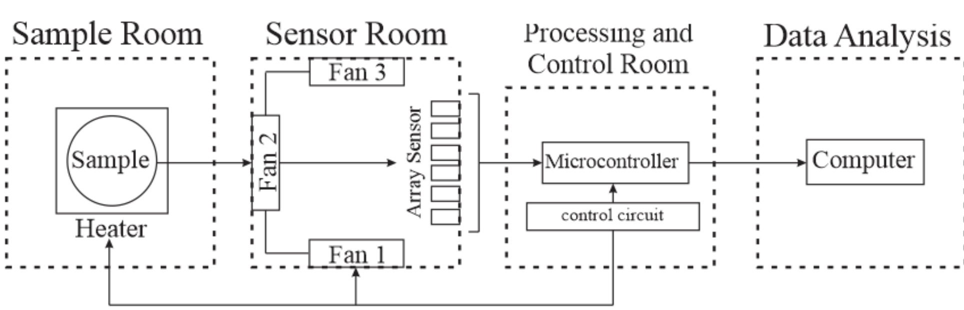


Figure 1. Diagram block of E-nose system

B. PROCEDURES AND DATA COLLECTION

B.1 SAMPLE PREPARATION

This study used 4 urine samples (normal urine, diabetes urine with variations of glucose levelof 100mg/dL, 500mg/dL and 1,000mg/dL). Each urine sample had the same volume of 40mL. Samples are placed in beaker glass of 50 mL.

B.2 DATA COLLECTION AND PROCESSING

In the process of data collection, obtained data in the form of sensor response signals, as shown in Figure *3*. The sensing process is shown in the odor ON section, while the flushing process is shown in the odor OFF section [8]. One cycle of data collection consists of one-time sensing and flushing. The sensing process itself is a sensor response when E-nose is exposed to sample odors, while the flushing process is a sensor response when E-nose is exposed to clean air. At the beginning of data collection, the E-nose system first performs a flushing process for 20 seconds. This aims to find the baseline value of each sensor so that it can be used as a reference for the baseline normalization process.

After the process of sampling odors data, the data is stored in a file with extension .csv to be processed first before the training process is carried out for the classification of urine samples. Data processing is meant by manipulating the baseline. This is because each sensor has different characteristics and sensitivity to certain substances so that each sensor has a different baseline value. The baseline value is the value when the sensor is exposed to air oxygen or fresh air. The baseline value for each sensor will be normalized so that it has the same baseline value. Then after normalization, the feature extraction process can be carried out using the mean method, namely by taking the average value of each cycle of sensing and flushing. This feature extraction data will be used as training data on machine learning algorithms. The following is the flow of data retrieval and processing procedures and expected outcomes, shown in Figure *2*.

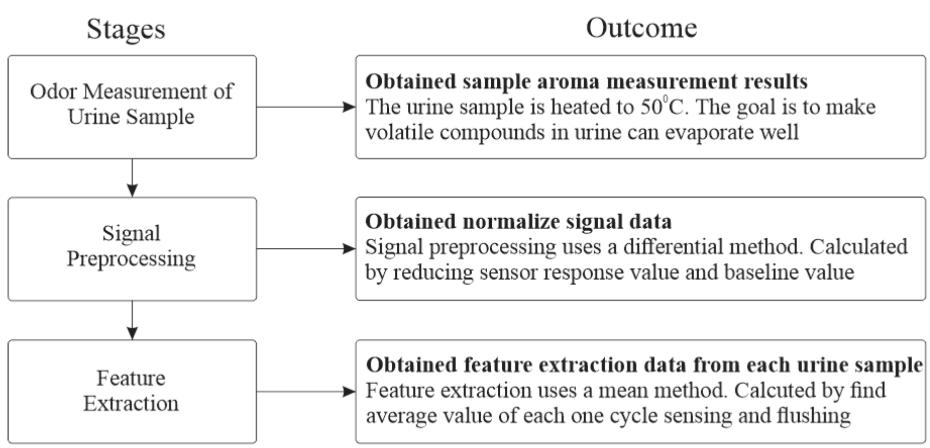


Figure 2. Stages of data collection and processing

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 6, Issue 7, July 2019

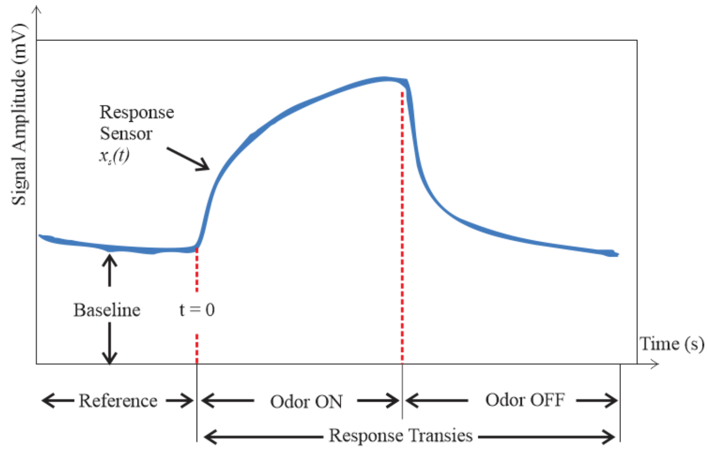


Figure 3. E-nose response signal

B.3 DIMENSIONAL REDUCTION AND DATA CLASSIFICATION

Before entering the classification stage, data from feature extraction was first reduced using the principal component analysis (PCA) method, eight variables representing the number of E-nose sensors to be only two variables. How the PCA method works are to reduce the dimensions of the correlated variables to be reduced variables that are not linearly correlated, which are called principle components to explain as much as possible the variance that occurs to a minimum [8].

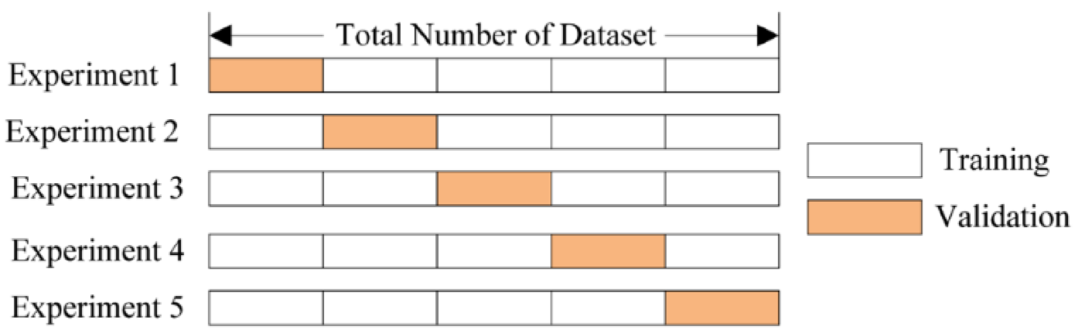


Figure 4. k-Fold Cross Validation process

After the data is successfully reduced, the next step is data classification with the machine learning algorithms. In this study, the algorithms used are KNN (K-Nearest Neighbors), SVM (Support Vector Machine), DT (Decision Tree) and RF (Random Forest). Each of these algorithms will be evaluated for accuracy in sample classification using the k-Fold Cross Validation method, which is one of the most commonly used methods to evaluate the success rate of a modeling system. The way it works is by dividing the data into two parts, the first part is used for the training process, and the other part is used as the test data [9]. This k-Fold Cross Validation will repeat times or as many pieces as part of the dataset used, for example in the first iteration process used as validation data is the first part of the data, then the other data will be used as training data. Thus for the second iteration, use the second part data as test data and so on for the next iteration. The process of validation with k-Fold Cross Validation is shown in Figure 4 below.

III. RESULTS&DISCUSSION

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 6, Issue 7, July 2019

A. ANALYSIS OF URINE DATA COLLECTION

Sampling data is taken by taking sensor response data from each of the existing gas sensors. From the sensor response data, profile information is obtained for urine samples being tested. The signal generated from the E-nose can be seen in Figure 5.

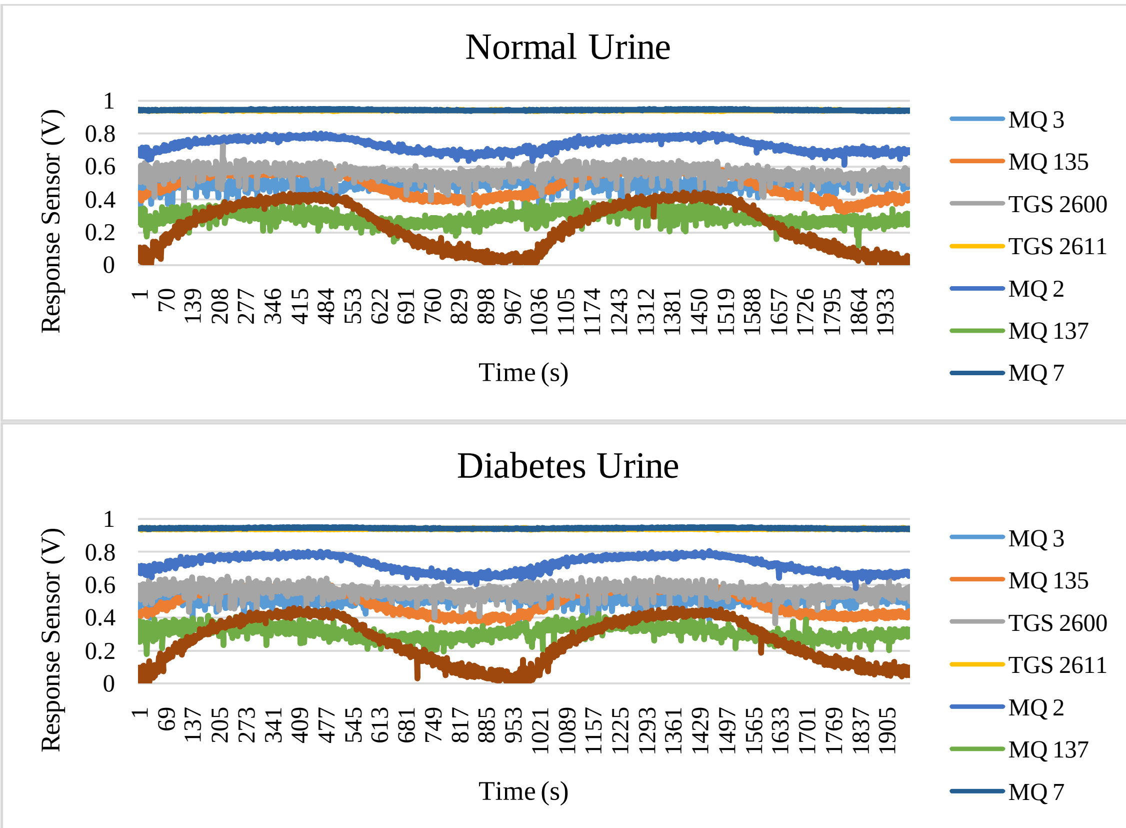


Figure 5. E-nose sensor reading result

Based on the picture that there is not much difference between normal urine and diabetic urine. This is likely because the response of each sensor is not compared to the reference baseline value. Therefore, the next process of signal processing needs to be done

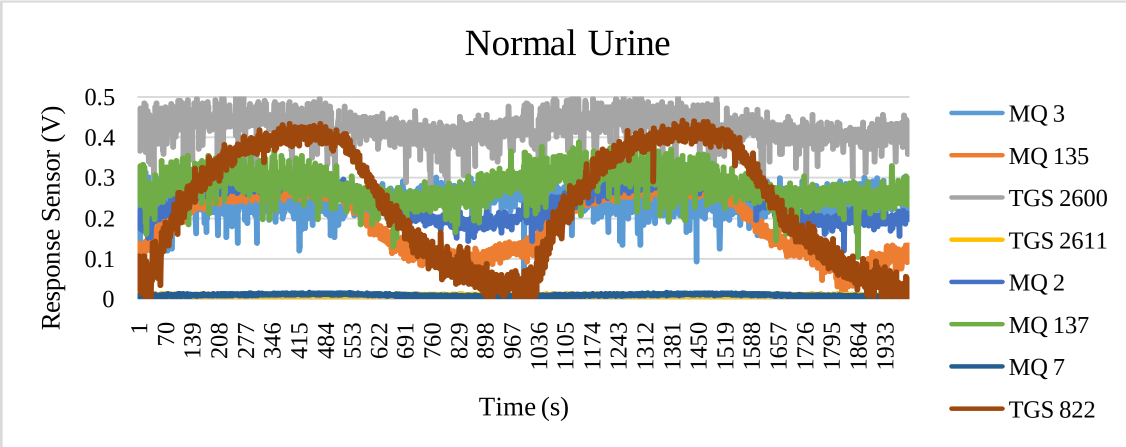
B. SIGNAL PREPROCESSING

The signal processing process used in this study is by adjusting the response signal to the reference signal (baseline). In doing baseline manipulation, this is done differentially, namely by measuring the value of the signal response with its baseline [10]. The baseline taken is the sensor response value when exposed to fresh air (flushing).

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 6, Issue 7, July 2019



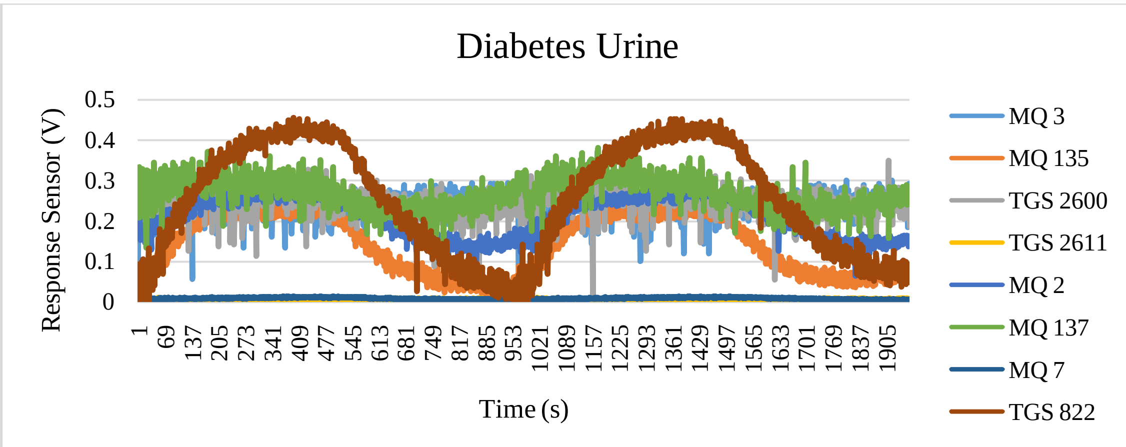


Figure 6. Signal preprocessing result

From Figure 6, it is known that several sensors show different sensor response values when exposed to diabetes urine aroma. So it can be concluded that E-nose produces different sensor response characteristics in normal urine samples and diabetic urine

C. FEATURE EXTRACTION

The feature extraction used in this study is the mean method. This method looks for average values in one cycle of sensing and flushing. From this average value that will be used for the process of training machine learning algorithms The results of feature extraction that has been obtained have a matrix size of 25 x 8 for each urine sample, where 25 lines represent the amount ofdata recorded and 8 columns represent the characteristics of the response signal that can be taken by each gas sensor. Because there are a total of 4 urine samples, the matrix is obtained.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 6, Issue 7, July 2019

D. DIMENSIONAL REDUCTION USING PRINCIPAL COMPONENT ANALYSIS (PCA)

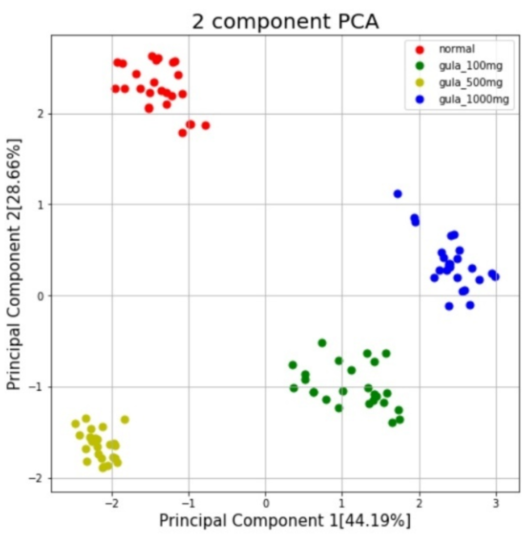


Figure 7. Dimensional reduction using PCA

Before conducting training using a machine learning algorithm,the feature extraction result data first reduced to from dimensions of 8 variables representing the number of sensors used in this study into two dimensions consisting of the first principal component (PC1) and the second principal component (PC2).

Based on Figure 7, the results of dimensional reduction using PCA obtained the percentage of data variation of 44,19 % for PC1, and 28,66% for PC2. Based on the figure, it can also be seen that a cluster of data points for normal urine samples can be separated from the collection of data points for diabetic urine samples with variations of glucose level 100 mg/dL, 500 mg/dL and 1000 mg/dL. Therefore it can be concluded qualitatively that E-nose can distinguish normal urine and diabetes urine based on the aroma characteristics of the sample.

E. CLASSIFICATION USING MACHINE LEARNING ALGORITHM

From the dimensional reduction data, qualitatively it can be seen that the sample data is well classified, but quantitative analysis needs to be done to make it more measurable. So that the data of urine samples odors obtained from the E-nose conducted further analysis by pattern recognition engine [11]. In this study, we used machine learning algorithms to recognize sample odors patterns and used the k-Fold Cross Validation method to measure the accuracy of the machine learning algorithm. The number of k used for validation in this study is 5.

E.1 KNN (K-NEAREST NEIGHBOUR)

One of the pattern recognition methods applied to the E-nose sample is the k-Nearest Neighbor (k-NN) algorithm. This k-NN algorithm has advantages such as being easy to adapt, both for training data that are many and few, and can be used for large test data as well. The workings of this algorithm are to place new objects in the existing class based on the results of the voting close to the object with the surrounding class [11]. In this study, the KNN algorithm uses a parameter of the number of neighbors (n neighbors) of 5. The following data from the accuracy of KNN algorithm are shown in Table 1.

Table 1. Accuracy result of KNN algorithm

	Split 1	Split 2	Split 3	Split 4	Split 5
Fit time	0.009475	0.003772	0.002867	0.00315	0.00339
Score time	0.008654	0.007241	0.006554	0.007853	0.006791
Test score	100%	100%	100%	100%	100%
Mean Accuracy	100%				

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 6, Issue 7, July 2019

E.2 SVM (SUPPORT VECTOR MACHINE)

Support Vector Machine or commonly abbreviated as SVM, is one of the machine learning methods that is often used, and this method is based on statistical theory combined with theories such as kernelling and optimization methods. SVM can classify data, both linearly and non-linearly [12]. The following data from the accuracy of the SVM algorithm are shown in Table 2.

Table 2. Accuracy result of SVM algorithm

	Split 1	Split 2	Split 3	Split 4	Split 5
Fit time	0.004104	0.005827	0.003158	0.002097	0.002712
Score time	0.001028	0.00151	0.001179	0.000962	0.000984
Test score	100%	100%	100%	100%	100%
Mean Accuracy	100%				

III.5.1 DT (DECISION TREE)

Decision Tree or also known as a decision tree, is a method that can be used in machine learning. This algorithm will create a kind of tree where each node is the attribute tested, and each leaf is the result of a prediction. The following data from the accuracy of the DT algorithm are shown in Table 3.

Table 3. Accuracy result of DT algorithm

	Split 1	Split 2	Split 3	Split 4	Split 5
Fit time	0.019637	0.008729	0.008174	0.007698	0.008011
Score time	0.002229	0.001512	0.001552	0.00153	0.001577
Test score	95%	85%	100%	100%	100%
Mean Accuracy	96%				

III.5.1 RF (RANDOM FOREST)

Random Forest is a machine learning method developed from the Decision Tree. This method is to collect the results of predictions from various Decision Tree in making final predictions. The following data from the accuracy of the RF algorithm are shown in Table 4.

Table 4. Accuracy result of RF algorithm

	Split 1	Split 2	Split 3	Split 4	Split 5
Fit time	0.03003	0.020709	0.028728	0.020427	0.022349
Score time	0.001759	0.001667	0.001701	0.001681	0.001678
Test score	95%	85%	100%	100%	100%
Mean Accuracy	96%				

F. VISUALIZE CLASSIFICATION RESULT

Every machine learning algorithm used in this study has its way of making decisions. This decision-making process can be seen through the data visualization process in Figure 8

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 6, Issue 7, July 2019

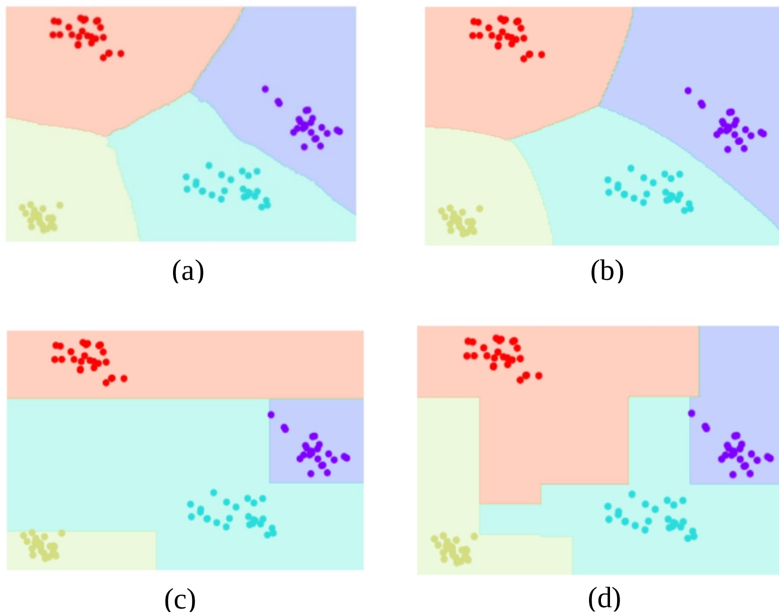


Figure 8. Visualization of ecision boundary for each algorithm (a) KNN, (b) SVM, (c) DT and (d) RF

Based on the visualization above, it can be seen that the KNN and SVM algorithms have succeeded in making a good decision boundary because the classified data is located in the middle, allowing the two algorithms to classify data with 100% accuracy. Then for DT and RF algorithms, there is data at the edge of the decision boundary, so that this can lead to a slight misclassification of the sample.

G. ACCURACY COMPARISON OF MACHINE LEARNING ALGORITHMS

Based on the results of the average accuracy of each machine learning algorithm used can be summarized the results of the accuracy comparison in Table 5 below.

Table 5. Accuracy comparison for each algorithm

Algorithm	Accuracy
KNN (K-Nearest Neighbour)	100%
SVM (Support Vector Machine)	100%
DT (Decision Tree)	96%
RF (Random Forest)	96%

It can be concluded that in general electronic nose (E-nose) can distinguish between normal urine and diabetes urine properly. Then from the four algorithms used, the KNN and SVM algorithms have the highest accuracy of 100 %.

IV. CONCLUSION

Based on the research that has been conducted, it can be concluded that the electronic-nose (E-nose) with eight types of gas sensors, including MQ3, MQ135, TGS2600, TGS2611, MQ2, MQ137, MQ7, and TGS822 successfully distinguish normal urine and diabetes urine. The results to classify urine samples, the KNN algorithm have 100% accuracy, SVM algorithm 100%, DT algorithm 96%, and RF algorithm 96%. Based on the comparison of classification methods, the KNN and SVM algorithms have the highest accuracy of 100%.

V. FUTURE WORK

Therefore this study recommends following suggestion that the E-nose system can be developed to be better in terms of aesthetics and can be directly connected with the software so that the data retrieval process and classification can be done at the same time. Also needed to test with more samples and sample variations to improve the accuracy and validation of aroma measurements by E-nose.

VI. THANKS

At the end of this research, we would like to sincerely thank Republic of Indonesia Ministry of Research, Technology and Higher Education through student creativity program has funded the entire research process

REFERENCES

[1] T. Thepudom, N. Sricharoenchai, and T. Kerdcharoen, "Classification of instant coffee odors by electronic nose toward quality control of production," in *2013 10th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications, and Information Technology*, 2013, pp. 1–4.

[2] I. N. Rahman, D. Lelono, and K. Triyana, "Klasifikasi Kakao Berbasis e-nose dengan Metode Neuro Fuzzy," *IJEIS (Indonesian J. Electron. Instrum. Syst.*, vol. 8, no. 1, p. 49, Apr. 2018.

[3] A. Vivi, "Identifikasi Penyakit Diabetes dengan Sensor Gas Menggunakan Metoda Pembelajaran Multilayer Perceptron," *Diploma thesis, Univ. Andalas.*, Feb. 2013.

[4] L. Capelli *et al.*, "Application and uses of electronic noses for clinical diagnosis on urine samples: A review," *Sensors (Switzerland)*. 2016.

[5] A. Modak, R. B. Roy, B. Tudu, R. Bandyopadhyay, and N. Bhattacharyya, "A novel fuzzy based signal analysis technique in electronic nose and electronic tongue for black tea quality analysis," in *2016 IEEE 1st International Conference on Control, Measurement and Instrumentation, CMI 2016*, 2016.

[6] M. Siadat, E. Losson, M. Ghasemi-Varnamkhasti, and S. S. Mohtasebi, "Application of electronic nose to beer recognition using supervised artificial neural networks," in *Proceedings - 2014 International Conference on Control, Decision and Information Technologies, CoDIT 2014*, 2014.

[7] T. Seesaard, C. Sriphrapradang, T. Kitiyakara, and T. Kerdcharoen, "Self-screening for diabetes by sniffing urine samples based on a hand-held electronic nose," in *BMEiCON 2016 - 9th Biomedical Engineering International Conference*, 2017.

[8] F. Rosyad and D. Lelono, "Klasifikasi Kemurnian Daging Sapi Berbasis Electronic Nose dengan Metode Principal Component Analysis," *IJEIS (Indonesian J. Electron. Instrum. Syst.*, 2016.

[9] S. Yadav and S. Shukla, "Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification," in *Proceedings - 6th International Advanced Computing Conference, IACC 2016*, 2016.

[10] C. A. Lintang, T. W. Widodo, and D. Lelono, "Rancang Bangun Electronic Nose untuk Mendeteksi Tingkat Kebusukan Ikan Air Tawar," *IJEIS (Indonesian J. Electron. Instrum. Syst.*, 2017.

[11] E. F. Anggara, T. W. Widodo, and D. Lelono, "Deteksi Daging Sapi Menggunakan Electronic Nose Berbasis Bidirectional Associative Memory," *IJEIS (Indonesian J. Electron. Instrum. Syst.*, 2017.

[12] R.- Rendyansyah, A. P. P. Prasetyo, and K. Exaudi, "Implementasi Electrinic Nose Dan Support Vector Machine Pada Aplikasi Olfactory Mobile Robot Dalam Mengenali Gas," *J. Nas. Tek. ELEKTRO*, 2018.