Detection of Diabetes from Gas Analysis of Human Breath using E-Nose

Hariyanto[1], Riyanarto Sarno[2]
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
[1]hariyanto13@mhs.if.its.ac.id., [2]riyanarto@if.its.ac.id

Dedy Rahman Wijaya[3]
School of Applied Science, Department of Informatics
Telkom University, Institut Teknologi Sepuluh Nopember
Bandung, Indonesia
[3]dedyrw@tass.telkomuniversity.ac.id

***Abstract*—Diabetes is one of the common disease that many people have suffered especially elderly. However, unfortunately only few of them that aware of this metabolic disease and most of them are undiagnosed. Therefore, in this research we propose low cost, non-invasive, and easy to use system that can distinguish healthy or diabetic patients so they can have early preventive action. A total of 40 e-Nose response signal from breath samples have been collected. There are seven main stages to build this system, the making of e-Nose hardware using microcontroller and gas sensors, ground-truth data acquisitions for the training set, signal processing for denoising using Discrete Wavelet Transform (DWT) and Z-score normalization, statistical features extraction, feature selection for optimization, classification, and e-Nose performance evaluation. The experimental results show that this system can distinguish healthy and diabetes patients with promising performance (95.0% of accuracy, 91.30% precision of diabetes, 94.12% precision of healthy and 0.898 kappa statistic's value) using k-NN classifier.**

Keywords—classification; diabetes; e-Nose; k-NN; microcontroller; signal processing

I. INTRODUCTION

Diabetes is one of the metabolic diseases that can affect almost all human organ systems in long-term and characterized by blood sugar levels (glucose) that far above normal. Traditionally, diabetes detection is performed invasively by taking blood samples but this method often giving negative stigma, like expensive, painful, and troublesome [1]. There are so many people that unaware with diabetes, for example in 2010 at about 7.6 million people suffered from diabetes in Indonesia with the prevalence of diabetes by 5.7% and more than 70% undiagnosed [2].

One that distinguishes diabetes patients with healthy patients is the gas content contained in breath. In a previous study, discovered that there was a relation between several biomarkers in human breath such as carbon monoxide, carbon dioxide, acetone, and volatile organic compound with diabetes and blood glucose levels [5].

One of the alternative method to detect diabetes non-invasively is using breath analysis [3]. Breath analysis is used to obtain information by analyzing volatile organic compound in breath samples. Electronic nose (E-Nose) is a tool for breath analysis that can identify, measure, and analyze the compound to gain information.

Hence, E-Nose for diabetes detection is attracting widespread interest because of its cheap value, portable, and easy to use [4]. In previous study, a breath analysis system has been developed for diabetes detection based on acetone concentration [3], but it only used ten subjects and one gas (acetone) to identify healthy, diabetic type 1, and diabetic type 2 patients.

In this study, we proposed early diabetes detection system from analysis of several biomarker contained in subject's breath sample to distinguish healthy and diabetic subjects. This study consists of seven stages, making of Electronic Nose (e-Nose), collecting ground-truth data, data preprocessing, feature extraction, feature selection, classification, and evaluation.

The data class is divided into two, healthy or diabetes so we can say that this is a binary classification problem. The data used for training as much as 40 samples, consisting of healthy and diabetes breath samples respectively as many as 20 samples taken from patients in Puskesmas Kedungdoro in Surabaya, Indonesia. Accuracy obtained by k-NN classifier is 95.0% with a kappa statistic value of 0.898 and this make our classifier had a near perfect performance.

II. METHODOLOGY

This study has a primary goal for the detection of diabetes through human breath analysis of gas signal data captured by electrochemical sensors connected to the microcontroller (e-Nose). Description of how the systems work can be seen in Figure 1.

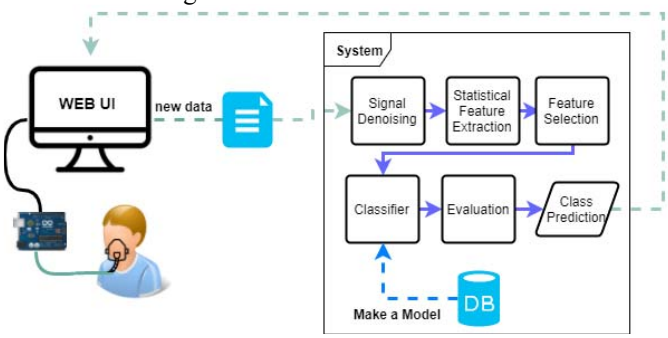


Figure 1. Blueprint of How the Diabetes Detection System Works

A. Making of E-Nose

E-Nose is a device that can identify the physical components of the odor and perform chemical susceptibility analyses to perform identification [5]. In this research, an E-Nose was designed using Arduino MEGA 2560 (microcontroller) and four analog electrochemical gas sensors along with temperature-humidity sensor. The e-Nose design can be seen in Figure 2. List of sensors used to capture the gas

concentration in the breath of diabetic patients can be seen in TABLE I.

TABLE I. LIST OF GAS SENSOR FOR THE MAKING OF E-NOSE

Sensor	Function
MQ-7	Carbon Monoxide (CO)
MQ-135	Carbon Dioxide (CO2)
MQ-138	Acetone
MiCS-5524	Volatile Organic Compound (VOC)
DHT-22	Temperature-Humidity

Figure 2. Design of E-Nose for Gas Analysis of Human Breath

Metal oxide semiconductor (MOS) sensors used is an analog sensor, so it takes a formula to convert the output into parts per million (ppm). The output generated by the MOS sensors is analog to digital conversion (ADC). To calculate sensor resistance (Rs) can be found on formula (1) or (2).

$$Rs = \frac{Vc - VRL}{VRL} \, x \, RL \qquad (1)$$

Where,

$$VRL = \frac{ADC \, x \, Vc}{1023} \qquad (2)$$

Vc is the voltage of microcontroller board, V_{RL} define the voltage of sensor in sample space and R_L define the sensor load resistance that can be measured using Ω meter. ADC is the output of gas sensor (analog to digital conversion). Rs is raw output from MOS sensors. Other than that, there are also a formula for converting the raw output to gas concentration in parts per million (ppm) by the formula (3) or (4).

$$C = \gamma \left[\frac{R_s}{R_o}\right]^{\tau}, \quad \gamma, \tau \in R^{+} \qquad (3)$$

$$C = 10^{\frac{\log(\frac{Rs}{Ro} - \gamma)}{\beta}}, \alpha, \beta, \gamma \qquad (4)$$

C, Rs, and Ro is estimated gas concentration (in ppm), actual sensor resistance and sensor resistance in the clean air that can be found on the MQ datasets for each gas sensors. Furthermore, $\gamma, \tau, \alpha,$ and β are constant value that can be determined from curve fitting using MATLAB Toolbox. Ro is determined when sensor calibration is performed in the clean air.

In the previous work, (3) was proposed method for estimating gas concentration [6] but in this research, new proposed method (4) is given because the new proposed method give lower value of standard error and better value of adjusted coefficient of determination for the estimating gas concentration.

Comparison of the two proposed methods for estimating gas sensor, standard error of the estimate and adjusted R^2 were used to prove which one of the formula is better. Below is the formula of those performance measure (5) and (6).

$$\sigma_{est} = \sqrt{\frac{\sum (Y - Y')^2}{N}} \qquad (5)$$

Where σ_{est}, Y, Y', and N are standard error of estimate, actual gas concentration, estimated gas concentration and the number of data in dataset.

$$R^2 Adj = 1 - \left[\frac{(1 - R^2)(N-1)}{N-K-1}\right] \qquad (6)$$

Where R^2, N, and K are coefficient of determination, number of data in dataset, and the number of independent regressors (i.e. the number of variables in model), respectively. On the other hand, coefficient of determination R^2 can be shown in formula (7).

$$R^2 = 1 - \frac{SSE}{SST} \qquad (7)$$

SSE define as sum of squared errors and SST define as total sum of squares.

Comparison of standard error of the estimate and adjusted coefficient of determination for both method can be seen in TABLE II. Coefficients for mathematical model used in the experiment are listed in TABLE III.

TABLE II. COMPARISON OF ESTIMATION GAS CONCENTRATION BETWEEN (3) AND (4)

Sensor	Gas	Equation (3)		Equation (4)	
		Standard Error	Adj. R^2	Standard Error	Adj. R^2
MQ-2	Alch.	0.1001	0.9752	0.09884	0.9759
MQ-7	CO	0.0407	0.9967	0.03045	0.9982
MQ-135	CO_2	0.04861	0.9967	0.0212	1
MQ-138	Ketone	0.03376	0.8408	0.01878	0.9507

TABLE III. COEFFICEINT USED FOR (3) AND (4)

Sensor	Gas	Equation (3)		Equation (4)		
		γ	τ	α	β	γ
MQ-2	Alcohol	16.11	-0.32	11.7	-0.22	-0.71
MQ-7	CO	34.29	-0.76	47.6	-0.85	0.07
MQ-135	CO_2	5.42	-0.35	5.89	-0.17	-1.55
MQ-138	Ketone	6.206	-0.75	-0.00026	1.40	0.38

B. Ground-Truth Data Acquisition

Ground-Truth data acquisition is done to get data learning (training set) that will be feed into classifier. There are 40 ground-truth data stored in database with 20 patients data for each class (healthy and diabetic). Patients with blood glucose

level (BGL) below 120 mg/dL considered to be healthy person and patients with BGL over 150 mg/dL considered to be diabetes. The measured patients whose breath was take were randomized blood glucose, not a fasting patient. Procedure for capturing ground-truth data can be seen on Figure 3.

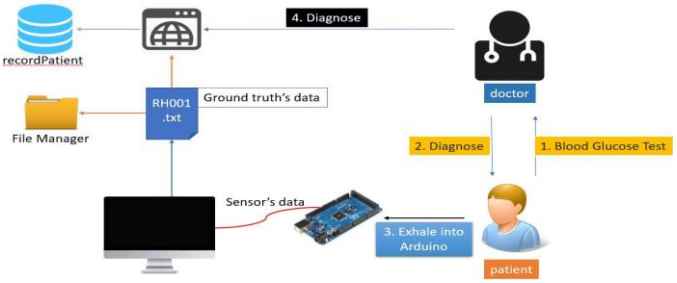


Figure 3. Collecting Ground-Truth Procedure

Each patient who was selected to be the ground-truth data sampled his breath for about 150 seconds using e-Nose connected to laptop. The graph of the time-based gas concentration recording for carbon monoxide, carbon dioxide, and ketones in the breath of healthy patient can be seen on Figure 4 (a), (b), and (c).

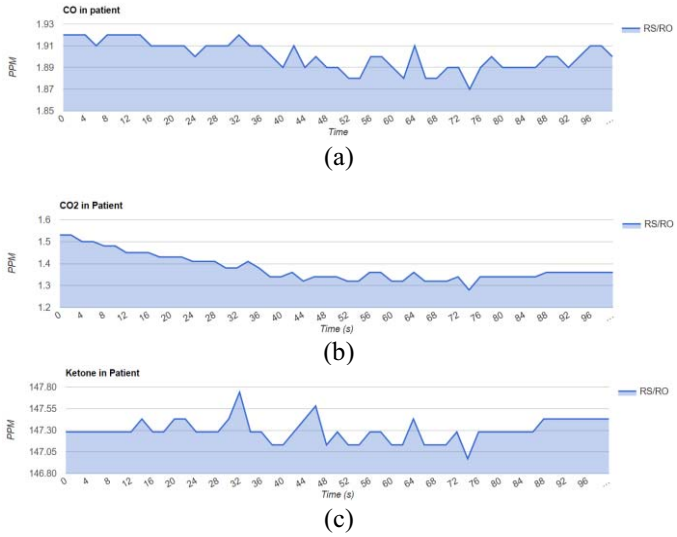
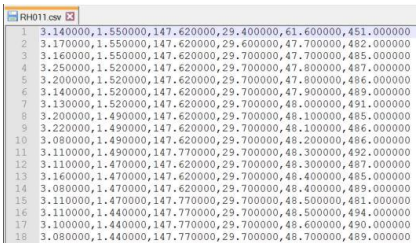


Figure 4 (a) CO Concentration, (b) CO_2 Concentration, (c) Ketone Concentration in Healthy Person Over Time from Gas Sensors

Sensor response data in ppm are sent in packets via USB Port and received by laptop or desktop in the format of comma separated value (csv), one packet (one line) every 3 second and written into the file manager. After that, the patient's csv file is uploaded into web system to store the data into database. Both csv file and web system shown in Figure 5 and Figure 6.


RH011.csv

```
3.140000,1.550000,147.620000,29.400000,61.600000,451.000000
3.170000,1.550000,147.620000,29.600000,47.700000,482.000000
3.160000,1.550000,147.620000,29.700000,47.700000,485.000000
3.250000,1.520000,147.620000,29.700000,47.800000,487.000000
3.200000,1.520000,147.620000,29.700000,47.800000,486.000000
3.140000,1.520000,147.620000,29.700000,47.900000,489.000000
3.130000,1.520000,147.620000,29.700000,48.000000,491.000000
3.200000,1.490000,147.620000,29.700000,48.100000,485.000000
3.220000,1.490000,147.620000,29.700000,48.100000,486.000000
3.080000,1.490000,147.620000,29.700000,48.200000,486.000000
3.110000,1.490000,147.770000,29.700000,48.300000,492.000000
3.110000,1.470000,147.620000,29.700000,48.300000,487.000000
3.160000,1.470000,147.620000,29.700000,48.400000,485.000000
3.080000,1.470000,147.620000,29.700000,48.400000,489.000000
3.110000,1.470000,147.770000,29.700000,48.500000,481.000000
3.110000,1.440000,147.770000,29.700000,48.500000,494.000000
3.100000,1.440000,147.770000,29.700000,48.600000,490.000000
3.080000,1.440000,147.770000,29.700000,48.700000,489.000000
```

Figure 5. Sensor Response in .csv File for 54 seconds

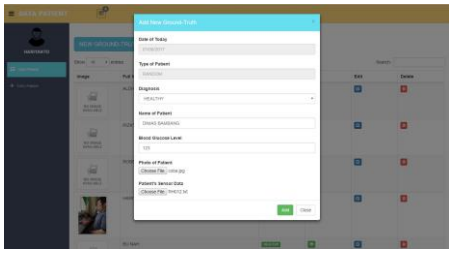


Figure 6. Sensor Response's File being Uploaded into Web System to Store the Data into Database

After being used by one person then the sensor box is opened to release the previous patient's breath and be fanned for about 1 minute.

C. *Preprocessing*

There are two phases on the preprocessing stage, the first one is signal denoising and the second is feature scaling (normalization). Signal denoising is important and can help to make accuracy and sensitivity of e-Nose better [7]. In this research, approximate coefficients of Discrete Wavelet Transform (DWT) with db6 base wavelet and decomposition level 1 is being used to denoise the signal data from each patient. The using of approximate coefficients of DWT with db6 mother wavelet and decomposition level 1 is based on the previous work [8]. In the previous work, it is confirm that wavelet transform successfully performs denoising in e-Nose application [9]. Denote the comparison between sensor response for carbon monoxide before and after denoising in Figure 7, the signal after denoising is seen more smooth (red lines draw diabetic patients and blue lines draw healthy patients).

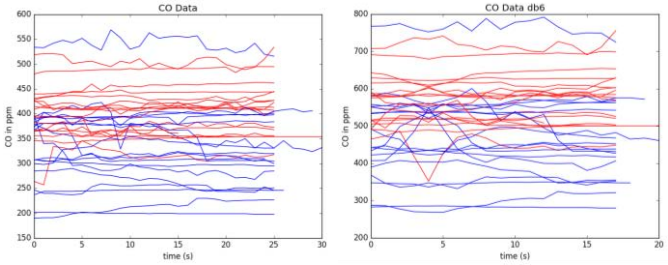


Figure 7. Left: CO Before Denoising. Right: CO After Denoising

Second phase is to do feature scaling using Z-Score Normalization (Standard Score) that will be useful for the features will be rescaled so they will have properties like a

standard normal distribution [10]. Below is the formula to compute z-score (8).

$$z = \frac{x-\mu}{\sigma} \quad (8)$$

Where z, x, μ, and σ are standard scores (can be called z score), actual data, mean of data, and standard deviation from the mean, respectively. Figure 8 is the result of normalization from denoise ppm CO concentration over a period.

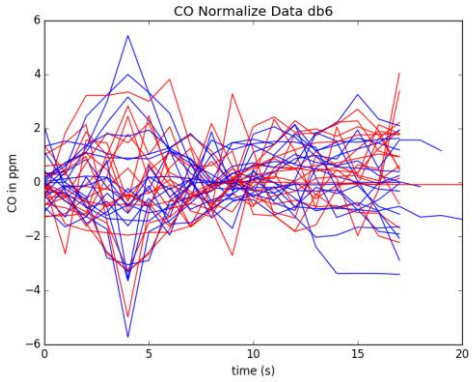


Figure 8. Normalization of Denoise CO Signal (blue: healthy, red: diabetic)

D. Statistical Features

The correspond gas concentration sensor response over time is a stationary signal. Stationary signals have statistical parameters that being constant over time [11]. In this experiment, we calculate four statistical parameters which is average, standard deviation, maximum value, and minimum value. Equation (9), (10), (11), (12) were used to calculate those four statistical parameters.

$$\mu = \frac{1}{n}\, x\, \sum_{t=0}^{n} y(t) \quad (9)$$

Where μ is the average value of signal, n is the length of signal, and y(t) is the signal value corresponding to time (t).

$$\sigma = \sqrt{\frac{1}{N}\, x\, \sum_{t=0}^{N} (y(t) - \mu)^2} \quad (10)$$

Where σ is the standard deviation of signal y, N is the length of signal, y(t) is the signal value towards time (t) and μ is the signal y average value.

$$\bigvee = max(y(t)) \quad (11)$$

$$S = min(y(t)) \quad (12)$$

Where $\bigvee$ and S are the maximum and minimum value from the signal y(t).

Therefore, because of there are four gas sensors and one temperature-humidity sensors, each patient will have 24 attributes corresponding to the statistical features of each gas concentration measured by the sensors, respectively there is 16 values from gas sensors and 8 values from temperature-humidity sensor.

ID	avg CO	std CO	max CO	min CO	…	avgV OC	stdV OC	max VOC	minV OC	Class

Figure 9. Ground-Truth's Vector Attributes Stored in Database

E. Feature Selection

Feature selection technique was performed to overcome the feature redundancy problem because of high correlated features. The main function of this technique is to find the optimal combination of gas sensor array or optimal attribute from the statistical features calculated [12]. In the previous work, Fast Correlation-Based Filter (FCBF) was being performed to find best suites features for e-Nose. Previous research proposed to use selected 7 sensor array that has 16% higher respect of General Resolution Factor (GRF) value to the use of 11 sensors, although there was an overlapping selectivity.

In this work, weighting by Chi Squared Statistic was performed to calculate the relevance of attributes by computing the value chi-squared statistic of each ground-truth's attribute with respect to the class attribute. To compute the chi-squared statistic, we can use (13).

$$X^2 = \sum \frac{(O-E)^2}{E} \quad (13)$$

Where X^2 is the chi-squared statistic, O is the observed frequency, and E is the expected frequency. After running this formula, then it is concluded to take best 6 attributes considered to distinguish healthy and diabetes patient. The selected attribute with the chi-squared statistic value can be shown in descending from the highest value.

TABLE IV. SELECTED ATTRIBUTES FROM X^2

Attributes	Chi-Squared Statistic Value
AVG_HUMID	32
AVG_CO	10.7
STD_CO	10.5
STD_CO2	9.5
STD_KETONE	9
STD_VOC	7

According to the table, the statistical features from temperature are excluded because of very small value of chi-squared statistic which can be interpreted that it has small correlation respected to the class attribute. So, after running the feature selection process 24 attributes of each ground-truth data can be reduced to just 5 attributes. The scheme of the ground-truth data that being stored in database after feature selection can be seen on Figure 10.

ID	avg CO	stdCO^2	stdKetone	std CO	avgHumid	stdVOC	Class

Figure 10. Ground-Truth's Vector Attributes Stored in Database after Feature Selection

F. Classification

There have been several previous studies using machine learning algorithms for diabetes classification. One of the using *Support Vector Machine* (SVM) for four levels of diabetes and get an accuracy of 68.66% [13].

In this research, we use k-NN for detection of diabetes from a new patient. There are two classes (conditions) that will be diagnose by the classifier, that which person is healthy or diabetes. K-NN is a similarity measurement technique that can be used for classification problem that based on learning from corresponding neighbor by comparing and calculating vote from given test case and training samples that are similar [14].

The distance method for similarity measurement that being used for this research is Canberra distance that can be shown in (14).

$$d_{CAD}(x,y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|} \quad (14)$$

Where d_{CAD} is computed Canberra distance, n is the length of vector attributes, x is the vector attribute one. And y is the vector attribute two. Canberra distance itself is a weighted version of Manhattan distance [15].

G. Evaluation Method

Evaluation method is used to estimate the performance of a model that being made by machine learning algorithm. There are several tools to measure those performance, such as confusion matrix on TABLE V and from confusion matrix after performing *leave-one-out* we can calculate accuracy, precision, recall, and kappa statistic to measure the classifier performance, respectively (15), (16), (17), and (18).

TABLE V. CONFUSION MATRIX FOR DETECTION OF DIABETES

	Predicted: Healthy	**Predicted: Diabetes**
Actual: Healthy	TruePositive(TP)	FalseNegative(FN)
Actual: Diabetes	FalsePoisitve(FP)	TrueNegative(TN)

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (15)$$

$$Precision\ of\ Diabetes = \frac{TN}{Predicted\ Diabetes} \quad (16)$$

$$Recall\ of\ Diabetes = \frac{TN}{Actual\ Diabetes} \quad (17)$$

$$Kappa = \frac{P(c)-P(r)}{1-P(r)} \quad (18)$$

Where P(c) is the performance of classifier (accuracy) and P(r) is the performance of random classifier.

III. RESULT AND DISCUSSION

A. Comparison of Result from Classifier

N is the number of top attributes.

TABLE VI. SCENARIO 1

Classifier: k-NN		
K = 8		
Distance method: Canberra		
Normalize: YES		
Top N	**Accuracy**	**Kappa**
1	95.0 %	0.898
2	95.0 %	0.898
3	90.0 %	0.794
4	90.0 %	0.796
5	85.0 %	0.685
6	80.0 %	0.583

TABLE VII. SCENARIO 2

Classifier: k-NN		
K = 8		
Distance method: Canberra		
Normalize: NO		
Top N	**Accuracy**	**Kappa**
1	92.5 %	0.846
2	72.5 %	0.412
3	60.0 %	0.14
4	60.0 %	0.13
5	65.0 %	0.239
6	57.5 %	0.061

TABLE VIII. SCENARIO 3

Classifier: Support vector machine		
Kernel: dot		
Normalize: YES		
Top N	**Accuracy**	**Kappa**
1	92.5 %	0.846
2	92.5 %	0.848
3	90.0 %	0.796
4	90.0 %	0.796
5	90.0 %	0.796
6	90.0 %	0.794

TABLE IX. SCENARIO 4

Classifier: Neural network		
Normalize: YES		
Top N	**Accuracy**	**Kappa**
1	92.5 %	0.846
2	90.0 %	0.796
3	87.5 %	0.746
4	87.5 %	0.746
5	92.5 %	0.848
6	90.0 %	0.796

Overall, k-NN give the highest value of accuracy and kappa statistic and stable against changes in the number of attributes. Although SVM and NN also give reliable performance but it's hard to implement those two classifiers into the web system and take longer time to execute.

B. Usability Testing

Usability testing is a test to measure how easy to use the system by testing it with real users [16]. There are two peoples who used the web-based system to do detection of diabetes so that early detection of preventive action can be done.

1. The first person has blood glucose level in 127 mg/dl so he must be a healthy person. After his data was put into the web system, the classifier said that he is a healthy person with the precision value of 89.47 %.

Figure 11. Usability Test 1

2. Second person has blood glucose level in 213 md/dl, so she must be had a diabetes. After his data was put into the web system, the classifier said that she had a diabetes with the precision value of 95.25 %.

Figure 12. Usability Test 2

IV. CONCLUSION

Based on the implementation of several machine learning algorithms for detection of diabetes and preprocessing method, some performance-based results were gathered. The highest accuracy of the system is obtained by k-Nearest Neighbors with the value of 95.0%. The preprocessing stage like feature scaling considered to be the reason of the high accuracy, because when k-NN was performed without preprocessing phase, the accuracy falls. The k-NN was proven reliable to handle the data despite with different number of attributes prove by the value of kappa statistic that considered classifier to be *near perfect performance*. Support Vector Machine and Neural Network also give good results but k-NN is preferable because of easy to implement to the web system and faster.

For the future, the diabetes detection system will be developed so that it can also detect prediabetes patients which has blood glucose between 120 mg/dL to 150 mg/dL whose current system cannot detect due to lack of the ground truth data. Also, we will collect sample data not only from random patients but also from patients who fasted before taking blood sample. We also want to estimate blood glucose level based on those gas concentrations.

ACKNOWLEDGMENT

Authors would like to thank Institut Teknologi Sepuluh Nopember for supporting the research and also Puskesmas Kedungdoro and RSAL Dr.Ramelan to let authors take sample data from patients.

REFERENCES

[1] C. Turner, "Potential of breath and skin analysis for monitoring blood glucose concentration in diabetes," *Expert Rev. Mol. Diagn.*, vol. 11, no. 5, pp. 497–503, 2011.

[2] P. Soewondo, A. Ferrario, and D. Tahapary, "Challenges in diabetes management in Indonesia: a literature review," *Global. Health*, vol. 9, no. 1, p. 63, 2013.

[3] L. S. and S. M., "Non- invasive diabetes detection and classification using breath analysis," *Commun. Signal Process. (ICCSP), 2015 Int. Conf.*, pp. 955–958, 2015.

[4] K. Yan and D. Zhang, "A novel breath analysis system for diabetes diagnosis," *ICCH 2012 Proc. - Int. Conf. Comput. Healthc.*, no. 61101150, pp. 166–170, 2012.

[5] A. D. Wilson and M. Baietto, "Advances in electronic-nose technologies developed for biomedical applications," *Sensors*, vol. 11, no. 1, pp. 1105–1176, 2011.

[6] D. R. Wijaya, R. Sarno, and E. Zulaika, "Gas concentration analysis of resistive gas sensor array," *2016 Int. Symp. Electron. Smart Devices*, pp. 337–342, 2016.

[7] H. Kim *et al.*, "Electronic-nose for detecting environmental pollutants: Signal processing and analog front-end design," *Analog Integr. Circuits Signal Process.*, vol. 70, no. 1, pp. 15–32, 2012.

[8] X. Guo *et al.*, "A novel feature extraction approach using window function capturing and QPSO-SVM for enhancing electronic nose performance," *Sensors (Switzerland)*, vol. 15, no. 7, pp. 15198–15217, 2015.

[9] J. Feng, F. Tian, J. Yan, Q. He, Y. Shen, and L. Pan, "A background elimination method based on wavelet transform in wound infection detection by electronic nose," *Sensors Actuators, B Chem.*, vol. 157, no. 2, pp. 395–400, 2011.

[10] R. Bott, "About Feature Scaling and Normalization - and the effect of standardization for machine learning algorithms," *Igarss 2014*, 2014. [Online]. Available: http://sebastianraschka.com/Articles/2014_about_feature_scaling.html. [Accessed: 01-Jul-2017].

[11] A. DLI, "Stationary Signals," 2009. [Online]. Available: http://www.azimadli.com/vibman/stationarysignals.htm. [Accessed: 01-Jul-2017].

[12] D. R. Wijaya, R. Sarno, and E. Zulaika, "Sensor array optimization for mobile electronic nose: Wavelet transform and filter based feature selection approach," *Int. Rev. Comput. Softw.*, vol. 11, no. 8, pp. 659–671, 2016.

[13] D. Guo, D. Zhang, L. Zhang, and G. Lu, "Non-invasive blood glucose monitoring for diabetics by means of breath signal analysis," *Sensors Actuators, B Chem.*, vol. 173, pp. 106–113, 2012.

[14] H. Djelouat, A. Ait Si Ali, A. Amira, and F. Bensaali, "Compressive sensing based electronic nose platform," *Digit. Signal Process.*, vol. 60, pp. 350–359, 2017.

[15] Math.NET, "Distance Metrics." [Online]. Available: https://numerics.mathdotnet.com/distance.html. [Accessed: 01-Jul-2017].

[16] Experienceux, "What is usability testing?," 2014. [Online]. Available: http://www.experienceux.co.uk/faqs/what-is-usability-testing/. [Accessed: 01-Jul-2017].