Available online at http://scik.org

J. Math. Comput. Sci. 11 (2021), No. 6, 7287-7301

https://doi.org/10.28919/jmcs/6535

ISSN: 1927-5307

FEATURES DEVELOPMENT IN MACHINE LEARNING MODEL FOR NON-INVASIVE BLOOD-GLUCOSE MEASUREMENT

YULI WIBAWATI*, ERFIANI, BAGUS SARTONO

Department of Statistics, IPB University, Bogor 16680, Indonesia

Abstract: Diabetes mellitus is the result of changes in the body caused by the decrease of insulin performance which is characterized by an increase in blood sugar level. Detection of blood sugar can be done with an invasive method or a non-invasive method. However, the non-invasive methods are considered better because they do not hurt the body and can check earlier, faster, and more accurately. A non-invasive blood glucose meter has been developed by a research team at IPB University. The output of the non-invasive tool is the intensity of the residual spectrum data, which will be related to the result of invasive measurement of blood glucose using some classification models, i.e. support vector machine and random forest. This research is aimed to compare the features development methods from the output of non-invasive tools and get the best features for modeling that can provide better predictions. The result of feature development shows that the best feature in the output of the non-invasive device is the trapezoidal area method at period because it has a higher accuracy value than the other four methods. The validation process shows that the random forest method has a higher accuracy value compared to the support vector machine.

Keywords: diabetes mellitus; non-invasive tool; features development; support vector machine; random forest.

2010 AMS Subject Classification: 65C60, 97C60.

*Corresponding author

E-mail address: yuliwibawati@gmail.com

Received July 20, 2021

1. INTRODUCTION

Health is an important factor in building the nation's civilization; one of the problems in the health sector is diabetes mellitus. According to the American Diabetes Association [1], diabetes mellitus is a group of metabolic diseases characterized by hyperglycemia that occurs due to abnormalities in insulin secretion, insulin action, or both. The number of cases and prevalence of diabetes mellitus in the world has continued to increase over the last few decades. Indonesia occupies the seventh position in cases of diabetes mellitus at the age of patients between 20 to 79 years. Diabetes mellitus can be prevented by implementing a healthy lifestyle and regular check-ups [9].

Examination of blood glucose levels is early detection in controlling and monitoring people with diabetes mellitus. Examination of blood glucose levels is generally carried out invasively, namely examinations that injure body parts. Currently, the Non-invasive Biomarking Team of IPB has developed a blood glucose measuring device that is non-invasive (non-invasive). The development of two tool designs has been carried out by the non-invasive team since 2016 namely the development of the missed tool design and the development of the reflected tool design. The differences in the two device designs are expected to be taken into consideration in the development of the tool by the non-invasive team.

Modeling to transform the output of non-invasive devices is very important in the development of measuring blood glucose levels. Annisa [3] has developed a support vector machine model, with data obtained directly from the output of a non-invasive tool on the residual intensity pattern. The features development carried out in previous studies used the average method, the three value method, and the standard deviation method. The highest accuracy result obtained from the development of the feature that has been carried out using the support vector machine is 42.01% for the standard deviation method. The selection of methods in previous studies has not maximized the results so that in this study tried to find another way to build the features to have a better impact.

The features development used in this research is to calculate the graphic area of the output

pattern of the non-invasive tool, residual intensity against the time domain because the output of the non-invasive tool not only produces a residual intensity pattern but there is a time-domain pattern that is very important to be used as material as the consideration in this study. The method of calculating the area of the graph was developed into two, namely, the method of calculating the area of the graph based on the period and peak. The features development method in this study is expected to produce other better features to obtain explanatory variables that can represent non-invasive output tools. In addition, this research is expected to provide better predictions on machine learning modeling from the best features formed. Based on the considerations of the previous research, the objectives of this study are to:

1. Compare methods features development to obtain explanatory variables based on the output of non-invasive blood glucose level measurement tools (the methods compared are the standard deviation method, the average method, the three-value method, the trapezoidal area method in each period and the trapezoidal area method at each top of modulation peak).
2. Compare the classification modeling technique of support vector machine (SVM) and random forest.

2. PRELIMINARIES

This part will describe in detail how the research instrument and research methodology is used to answer the two research objectives.

2. 1 Research instrument

Non-invasive blood glucose level measurement tools use the concept of spectroscopy. Spectroscopy is the science that deals with the spectrum of light. In this study, the infrared light spectrum used in non-invasive devices is with a wavelength of 1600 nm. The wavelength of 1600 nm is a wavelength that is sensitive to changes in sugar concentration [10]. The working principle of the non-invasive tool is to illuminate fingers with infrared, then infrared light rays captured by the sensor will produce a digital analog in the ADC process. The result of the ADC is a digital signal that will be programmed in the server. The server display will appear on the LCD showing

the residual light intensity against the measurement time domain.

Residual light intensity to the time domain is formed based on certain periods of time when the infrared lamp is turned on, each period is set to work in a certain modulation. Modulation is the lighting level of the lamp, with the maximum lighting level being 1023. There are 10 modulations applied, namely modulation 0, 10, 20, up to 90. Illustration for Modulation of 20 means that the lamp is working with an illumination level of 222 out of 1023 or 21,70 %. The modulation name is adjusted to make it easier to pronounce. The total period of the non-invasive device is 40 periods in which one measurement period the lamp is turned on and turned off for five times, namely off-on-off-on-off.

Table 1. Working period and modulation settings on non-invasive devices

Infrared light turned on					Illumination Level	Modulation
Periode	None	1550 nm	1600 nm	1500 nm & 1600 nm		
	1	11	21	32	0/1023 = 0%	0
	2	12	22	32	111/1023 = 10,85%	10
	3	13	23	33	222/1023 = 21,70%	20
	4	14	24	34	333/1023 = 32,55%	30
	5	15	25	35	444/1023 = 43,40%	40
	6	16	26	36	555/1023 = 54,25%	50
	7	17	27	37	666/1023 = 65,10%	60
	8	18	28	38	777/1023 = 75,95%	70
	9	19	29	39	888/1023 = 86,80%	80
	10	20	30	40	999/1023 = 97,65%	90

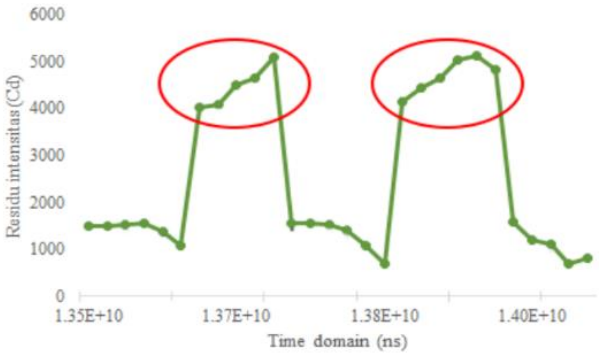

Figure 1. Graphic illustration of residual intensity of one period of non-invasive device measurement

The pattern of the residual intensity graph against the time domain has been carried out by Aurelia [4]. The results of the study shows a constant pattern at modulation 0 to modulation 40, compared to modulation 50 to 90 modulation which shows a non-constant pattern.

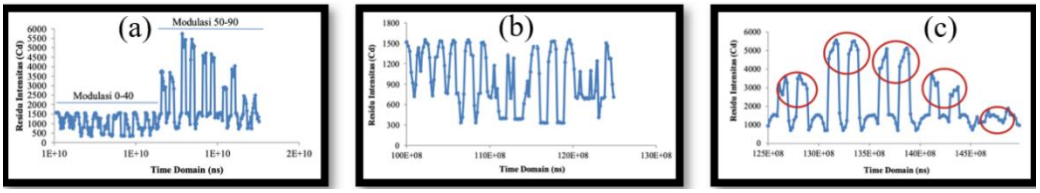

Figure 2. The residual intensity pattern against the time domain of non-invasive instruments is based on (a) a 0 to 90 modulation pattern, (b) a 0 to 40 modulation pattern, and (c) a 50 to 90 modulation pattern.

The features used in this study is a modulation of 50 to 90 modulation (period 26–30), because it produces a residual intensity pattern that is not constant, and the intensity and time domain residual patterns are read differently.

2. 2 Research methodology for the first objective

2. 2. 1 Data

The data used in this study were primary data resulting from measuring blood glucose levels using invasive and non-invasive methods. The invasive method is the result of laboratory measurements and the non-invasive method is part of the research "Development of a Non-Invasive Patient Monitoring System for Patients with High Blood Glucose Levels" by the Non-Invasive Biomarking team of Bogor Agricultural University (IPB). The data were obtained from the research team that has developed a non-invasive device design. The data collection was carried out in the period April–January 2017, hereinafter referred to as 2017 data. The respondents involved in the 2017 data collection were 118 students from several departments at IPB

The response variable used in this study was data from the measurement of the invasive method (gold standard) by using a syringe inserted into a vein as much as ±4 ml, then the blood

sample was processed in the Prodia clinical laboratory. The results of invasive measurements are blood glucose levels (mg/dl), while the explanatory variables used in this study are non-invasive measurement data. There was several methods features development to obtain explanatory variables, which was the first objective in this study.

2. 2. 2 Features Development

The features development process carried out in this study is to calculate the area of the graph on the residual intensity pattern against the time domain in each period and each modulation peak. The pattern of the graph area formed from the residual intensity and time domain is like a trapezoid, so in this study, the calculation results are referred to as the area of the trapezoid. The illustration is shown in Figure 3, and the area calculation approach uses the following formula:

$XL_k = \frac{1}{2}\sum_{i=1}^{n-1}(t_{i+1} - t_i)(y_i - y_{i+1})$ k= 1, 2, …, 25 (Based on period) & k= 1, 2, …, 50 (Based on peak),

XL_k is the value of the residual graph area of the modulation peak intensity and or the *k-th* (Cdns), t_i is the *Time domain i-th* (ns), y_i is the residual light intensity value *i-th* (Cd), n is the number of time-domain boundary points at the peak and or period.

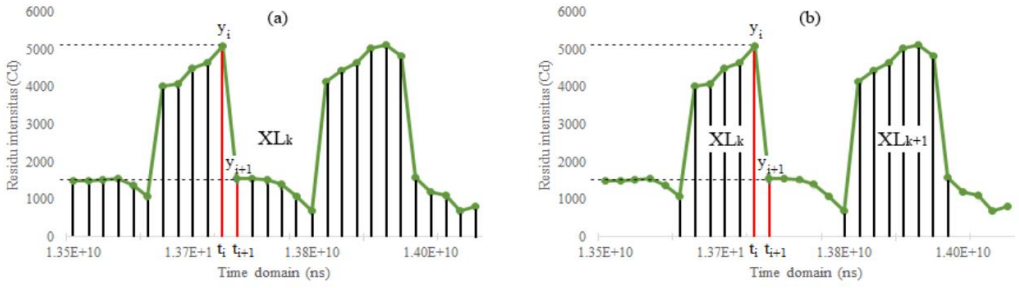

Figure 3. Illustration of area calculation under residual intensity graph based on (a) period and (b) peak.

The next process will compare the features development method that has been carried out to obtain the best features, as follows:

a) Average Method
b) Three Value Method

c) The Standard Deviation Method

d) Method of trapezoidal area period

e) Trapezoid area method at each top of modulation peak

The method in points (a), (b), and (c) has been carried out by Annisa [3] by looking at the accuracy of classification modeling using a support vector machine.

The method in point (d) and point (e) that will be used in this study is to calculate the area of the graph on the intensity residual pattern against the time domain at each period and peak. The measure of the goodness of the model used is accuracy using a support vector machine.

2. 2. 3 Measures of model goodness

Confusion Matrix is a method commonly used to perform the calculations of the accuracy value on the concept of data mining or decision support systems. The following is the result of the confusion matrix [7].

Tabel 2 Tabel *Confusionan Matrix*

Prediction	Actual	
	True	False
True	TP	FN
False	FP	TN

Information:

TP = number of correct predictions are positive (*True Positive*)

TN = number of correct predictions are negative (*True Negative*)

FP = number of correct predictions are positive (*False Positive*)

FN = number of correct predictions are negative (*False Negative*)

Accuracy is the proportion of the number of correct predictions. The equation used to calculate accuracy is shown in the equation below. The application of accuracy calculation used in this study is balanced-accuracy (ba) which functions to handle multiclass classifications with unbalanced data [7] [8].

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad b.a = 1/2(\frac{TP}{TP + FN} + \frac{TN}{TN + FP})$$

The measure of the goodness of the model used is accuracy, which is the percentage of the amount of data that is correctly predicted to the total amount of data. The classification model used in this study is a *support vector machine* with the categories used according to the [1], namely by

using low blood glucose levels (<70 mg/dL), normal (≥70 mg/dL - 99 mg/dL) and high (≥99 mg/dL).

2. 3 Research methodology for the second objective

2. 3. 1 Data

The data used in this study were primary data resulting from measuring blood glucose levels using invasive and non-invasive methods. The invasive method is the result of laboratory measurements, and the non-invasive method is part of the research "Development of a Non-Invasive Patient Monitoring System for Patients with High Blood Glucose Levels" by the Non-Invasive Biomarking team of Bogor Agricultural University (IPB). Data collection was carried out in the period of July 2019. The data set is then referred to be the 2019 dataset. The respondents involved in the 2019 data collection were 74 general public aged 21 to 87 years.

The response variable used in this study was the data from the measurement of the invasive method (*gold standard*) by using a syringe inserted into a vein as much as ±4 ml, then the blood sample was processed in the Prodia clinical laboratory. The results of invasive measurements are blood glucose levels (mg/dl), while the explanatory variables used in this study are non-invasive measurement data using the best method features development produced in the first objective of this study.

2. 3. 2 Features Development

The features development process used in this second research methodology is the best feature produced from the first objective, namely the area of the peak trapezoid. This feature development is the best because it involves the residual value of the intensity of the time domain.

The illustration is shown in Figure 3a, and the area calculation approach uses the following formula:

$$XL_k = \frac{1}{2}\sum_{i=1}^{n-1}(t_{i+1} - t_i)(y_i - y_{i+1}) \qquad k = 1, 2, \dots, 25 \text{ (Based on period)}$$

XL_k s the residual value of the k-th modulation peak intensity graph (Cdns), t_i is the i-th time domain *(ns)*, y_i is the i-th residual value of light intensity *(Cd) n* is the number of time domain boundary points at the period (Figure 3a).

2. 3. 3 Comparison of support vector machine and random forest classification methods

Machine Learning classification methods have long been used in the field of data mining. According [6], classification is the process of finding a collection of data patterns with other data to be used to predict data that does not yet have a certain data class. Support vector machine is a machine learning method that works on the principle of Structural Risk Minimization (SRM) which aims to determine the best hyperplane to separate data classes. Random forest was first introduced by [4], his research showed the advantages of random forest.

Validation is done by cross validation method. Cross validation is a method for predicting the accuracy of test data. The cross validation method used is k-fold cross validation, namely, a technique that can repeat training and test data with k epetitions and $1/k$ division of the data set, where $1/k$ will be used as test data (Kohavi, 2014). Comparing classification methods based on the output of non-invasive blood glucose levels in 2019 data is carried out by a support vector machine and random forest. The distribution of training data and test data used in this study was 90%:10% with k=10. According to the [2] the categories used in the comparison of this classification method use normal blood glucose levels (<100 mg/dL), pre-diabetes (≥100 mg/dL - 126 mg/dL) and diabetes (≥126 mg/dL).

The measure of the goodness of the model used is accuracy, which is the percentage of the amount of data that is correctly predicted to the total amount of data. The classification model used in this study is a support vector machine and a random forest.

3. MAIN RESULTS

This section will describe in detail how the results and discussion obtained from the two research methodologies have been described and answer the results of the research objectives.

3. 1 Results and discussion for the first objective

3. 1. 1 Descriptive Analysis

Descriptive analysis used on the response variable is the data from the measurement of invasive blood glucose levels. According to the [1] there are 3 categories of blood glucose levels, as shown in Table 3. Of the 118 respondents, 6.8% of respondents had low blood glucose levels, 86.4% of respondents had normal glucose levels, and 6.8% of respondents had high blood glucose levels. Descriptively, it can be said that most of the respondents had normal blood glucose levels.

Table 3 Descriptive response variables in 2017 observation data data

Year	Low (<70mg/dl)	Normal (≥70-99mg/dl)	High (≥99mg/dl)	Numbers of respondents	Min	Median	Max	Mode
2017	8 (6,8%)	102 (86,4%)	8 (6,8%)	118	67	80	276	78

Figure 5 is a repeat visualization of the results of non-invasive measurements on 2017 data. The pattern formed on respondents who have the same blood glucose levels and who have different blood glucose levels, has almost the same pattern.

Figure 5 Comparison of non-invasive charts of two 2017 respondents based on (a) different blood glucose levels and (b) the same blood glucose levels

The descriptive analysis used as the explanatory variable is the result of non-invasive measurement of blood glucose levels. Table 4 shows that the median and average values have a range that is not too much different. The skewness value is 0.0608, which means that the value is

symmetrically distributed, with the same distance between the right and left side of the distribution.

Table 4 Descriptive of the explanatory variables in the 2017 observation data

Tahun	Min	Median	Max	Mean	Range	Skewness
2017	27996308307	2.67679E+11	6.39416E+11	2.42612E+11	6.1142E+11	0,060834789

3. 1. 2 Comparing methods to obtain explanatory variables

Some of the methods features development used to obtain explanatory variables are 2017 data. The goal is to produce data that represents the output of non-invasive blood glucose levels. The method to obtain explanatory variables in previous research has been carried out by Annisa [3], namely by using the average method, the three-value method and the standard deviation method. In this study, the method used to obtain the explanatory variable is using the area of the trapezoid at the period and peak of each modulation. All methods to obtain explanatory variables were tested with accuracy values in the support vector machine classification modeling.

Table 5. Accuracy values of several methods for explanatory variables

Method	**Support vector machine accuracy**	**Random forest accuracy**
Average Method	32.34 %	—
Three Value Method	35.68 %	—
The Standard Deviation Method	42.01 %	—
Method of trapezoidal area period	86.92 %	86.92 %
Trapezoid area method at each top of modulation peak	85.92 %	87.68 %

The method used to obtain the explanatory variables in previous studies only involved the residual intensity value, while the method used in this study was a calculation involving the residual intensity value with the time domain. Table 5 shows that the method that has the highest accuracy is a calculation involving the residual intensity value with the time domain, namely the Trapezoid area method at period. Research in the next stage for the classification process uses the

period trapezoidal area method.

3. 2 Results and discussion for the second objective

3. 2. 1 Descriptive Analysis

Descriptive analysis used as a response variable is the result of invasive blood glucose measurement. According to [2] it is known that there are three categories of blood glucose levels, as shown in Table 6. Of the 74 respondents, 45,9% of respondents had normal blood glucose levels, 14,9% of respondents were in the pre-diabetes category, and 39,2% of respondents were in the category of diabetes. Descriptively, it can be said that most of the respondents had normal blood glucose levels.

Table 6 Descriptive response variables on observational data in 2019

Year	Normal (<100mg/dl)	Pra-diabetes (≥100-126mg/dl)	Diabetes (≥126mg/dl)	Numbers of respondents	Min	Median	Max	Mode
2019	34 (45,9%)	11 (14,9%)	29 (39,2%)	74	69	105	614	87

The graph in Figure 6 is a repeat visualization of the results of non-invasive measurements on 2017 data. The patterns formed in respondents who have the same blood glucose levels and those who have different blood glucose levels have almost the same pattern. The graph that is formed on respondents who have the same blood glucose levels tends to have a rare range that is not too different from the range of respondents who have different blood glucose levels.

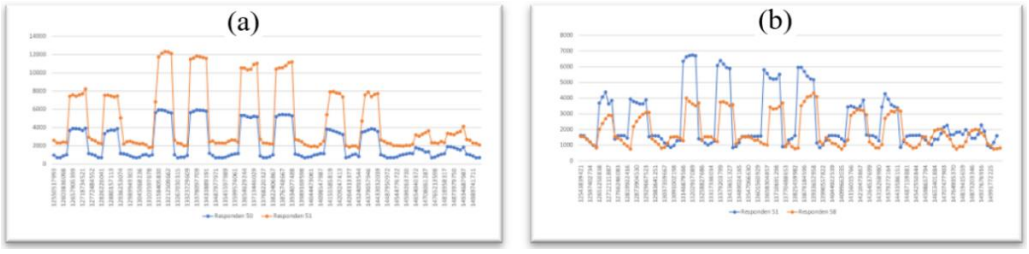

Figure 6 Comparison of non-invasive charts of two 2017 respondents based on (a) different blood glucose levels and (b) the same blood glucose levels

The correlation coefficient between invasive blood glucose levels and the period area method in non-invasive measurements is 0,036355. This correlation value means that there is a low relationship between invasive blood glucose levels and period area in non-invasive measurements. It can be seen in Figure 7 that a large area value does not indicate a high blood glucose level because each area value in the explanatory variable has a different level of lighting and modulation.

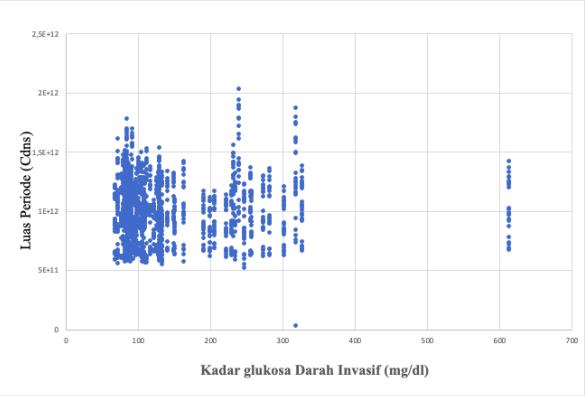

Figure 7 Period area graph on non-invasive data for each invasive data

3. 2. 2 Comparing the classification method of support vector machine and random forest

The data used in the classification was observational data for 2019, because the data collection techniques used by the respondents were more representative and vary based on age. The 2019 observations for blood glucose data in each category were more balanced, because most machine learning algorithms could work well when the data set of each category was balanced.

The classification modeling process is carried out on the best features produced by the trapezoidal area method for each modulation period. The modeling produced an accuracy value for the support vector machine of 45.87% with the optimal selection of parameters, namely k = 10, c = 1, and gamma = 0.5. While the accuracy value for random forest is 45.98% with optimal parameter selection, namely $k = 10$, $c = 1$, and $gamma = 0{,}5$. While the accuracy value for random forest is 48.73% with optimal parameter selection, namely $k = 10$, $ntree = 1$, and $mtry = 14$.

CONCLUSION

The first conclusion in features development to obtain the best explanatory variables from the output of non-invasive data tools in 2017 is to use the trapezoidal area method at each modulation period. The results of the classification modeling used the support vector machine (SVM) with the trapezoidal area method at each modulation period resulted in an accuracy value of 86.92%.

The second conclusion is on classification modeling using the trapezoidal area method at each period of the 2019 data non-invasion blood glucose measurement. The results obtained are that the classification modeling using the support vector machine (SVM) produces an accuracy value of 45.98%. The accuracy value obtained is smaller than the classification modeling using random forest (RF) which produces an accuracy value of 48.73%.

CONFLICT OF INTERESTS

The author(s) declare that there is no conflict of interests.

REFERENCES

[1] American Diabetes Association. 2010. Position statement: Standards of Medical Care in Diabetes 2010. Diab Care. Suppl.1 (2010), 33.

[2] American Diabetes Association (USA). Factors Affecting Blood Glucose [Internet]. [diakses 2018 Mei 18]. Tersedia pada: http://www.diabetes.org/living-with-diabetes/treatment-and-care/bloodglucose-control/factors-affecting-bloodglucose.html?referrer=https://www.google.co.id/ (2018).

[3] S. Annisa, Klasifikasi kadar glukosa darah keluaran alat non-invasif menggunakan metode support vector machine. Tesis. Pascasarjana Institut Pertanian Bogor (IPB). Bogor, 2019.

[4] K. Aurelia, Pendugaan kadar glukosa darah non-invasif menggunakan regresi kuadrat terkecil parsial dengan beberapa pendekatan peringkasan [skripsi]. Bogor: Institut Pertanian Bogor. 2020.

[5] L. Breiman, Random Forests. Statistics Departement, University of California [internet]. [diacu 2016 Maret 05]. Tersedia dari: https://www.stat.barkeley.edu/~breiman/randomforest2001.pdf. (2001).

[6] J. Han, M. Kamber, P. Jian, Data mining concepts and techniques, Elsevier, Amsterdam, (2012).

[7] J.D. Kelleher, B. Mac Namee, A.D. Arcy, Fundamentals of machine learning for predictive data analytics algorithms, Worked Examples, and Case Studies. The MIT Press, London, 2015.

[8] K.H. Brodersen, C.S. Ong, K.E. Stephan, J.M. Buhmann, The balanced accuracy and its posterior distribution, in: 2010 20th International Conference on Pattern Recognition, IEEE, Istanbul, Turkey, 2010: pp. 3121–3124.

[9] [IDF] International Diabetes Federation. 2019. IDF Diabetes Atlas Ninth Edition [Internet]. [diakses 2020 Jun 14]. Tersedia pada : https://diabetesatlas.org/en/

[10] J. Prabowo, Y. Suryana, R. Ferbyarto, I.M. Astawa, Sistem Instrumentasi Alat Ukur Kadar Gula Darah Non Invasive Berbasiskan Arduino. https://jurnal.umj.ac.id/index.php/semnastek/article/view/698. (2016).