PERFORMANCE EVALUATION OF LDA AND SVM FOR PREDICTING DIABETES

Abdul Azis Abdillah,[1, a)] Azwardi,[1, b)] Imam Wahyudi,[1, c)] and Samsul Arifin[2, d)]

[1)] *Alat Berat, Mechanical Engineering Dept, Politeknik Negeri Jakarta, Depok, Indonesia, 16425.*
[2)] *Statistics Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia. 11480*

[a)] *Electronic mail: abdul.azis.a@mesin.pnj.ac.id*
[b)] *Electronic mail: azwardi@mesin.pnj.ac.id*
[c)] *Electronic mail: imam.wahyudi@mesin.pnj.ac.id*
[d)] *Corresponding author: samsul.arifin@binus.edu*

Abstract. Currently, there is no cure for diabetes mellitus. Diabetes mellitus is a deadly disease that causes various complications, including heart failure, kidney disease, wounds that do not heal and so on. One solution that can be applied from this disease is to take precautions. This study tries to develop an early detection system for diabetes mellitus using artificial intelligence. In this paper, researchers combine the use of two machine learning methods: Linear Discriminant Analysis (LDA) and Support Vector Machines (SVM) to detect diabetes. This paper aims to determine the optimal classification model for detecting diabetes. In addition, this paper also describes the implementation and performance of the LDA-SVM method for detecting diabetes. In this case the dataset used to conduct the experiment is the PIMA Indian Diabetes dataset. In addition, we use a 10-fold cross validation design to get the optimal parameters and to build the best model. The conclusion of this study is that LDA-SVM successfully detected diabetes using sigma=-4.5 with an accuracy value of 77.34%, a sensitivity of 73.507% and a specificity of 79.60%.

INTRODUCTION

Diabetes is a serious chronic disease that occurs because the pancreas does not produce enough insulin (a hormone that regulates blood sugar or glucose) needed by a person's body. Diabetes is also one of the important public health problems and has even been highlighted by world leaders because the number of cases and the prevalence of diabetes have continued to increase over the last few decades [1, 2]. In Indonesia the number of diabetics in 2000 reached 8.4 million population, this is expected to continue to grow to around 21.3 million in 2030 [1, 2]. Besides, throughout the world the number of diabetics has increased substantially between 1980 and 2014, increasing from 108 million to 422 million or around four times [1, 2].

Based on data released by the Indonesian-Ministry of Health [2], Indonesia is the 4th country with the highest number of diabetes mellitus sufferers in the world in 2000 with a total of 8.4 million sufferers. In addition, it is estimated that this number will continue to increase from year to year. Until it is estimated that the number of people with diabetes in 2030 in Indonesia will be around 21.3 million.Based on the the data [2] can be seen that there is an estimated increase in diabetics in the coming year. Until now the medical world has not found a cure for sufferers. While what can be done is to prevent diabetics from complications and premature death. These include policies and direct applications in populations and certain environments (schools, homes, work environments) that contribute to the health of all people, both those with diabetes and not, such as regular exercise, healthy eating, avoiding smoking, and controlling fat levels and blood pressure. In response to this, the action is needed to reduce the increase in diabetics in the future. One way to make a diabetes detection system that can detect diabetes automatically, quickly and accurately [3, 4, 5].

To make this diabetes detection system, researchers tried to combine Linear Discriminant Analysis and Support Vector Machines to make. SVM is a machine learning method that is often used [6, 7, 8, 9]. Where SVM has a basic principle for carrying out binary classifications and subsequently developed to be able to work in multiclass cases. Besides, researchers also use the Linear Discriminant Analysis method to extract features and at the same time reduce the dimensions of the dataset used.

Several previous studies [6, 7, 8, 9] related to diabetes detection, the methods used include SVM, Bayesian Regularization Backpropagation, PCA and SVM. In this study, the authors tried to conduct experiments by combining the SVM method as a classification method and Linear Discriminant Analysis (LDA) to determine its performance. The purpose of this study is to find the best parameters in building a diabetes detection system using LDA and SVM methods.

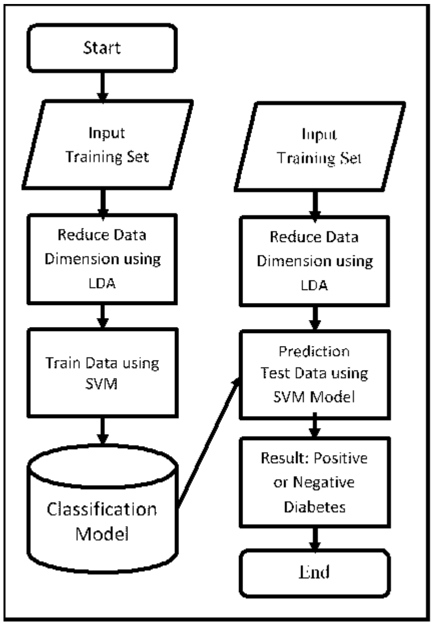
Start
Input Training Set
Reduce Data Dimension using LDA
Train Data using SVM
Classification Model
Input Training Set
Reduce Data Dimension using LDA
Prediction Test Data using SVM Model
Result: Positive or Negative Diabetes
End

FIGURE 1. Research Framework

RESEARCH METHODOLOGY

In this section will be described about the methods and methods of evaluation used in this research. Framework of this research can be seen in figure 1.

Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) [10, 11, 12] is known as a technique used in reducing dimensionality in the pre-processing step for pattern classification and machine learning applications. The method used is to project the dataset to a lower-dimensional space with good class separation to avoid overfitting ("curse dimension") and also reduce computing costs. The purpose of this LDA can be said to project feature space (n-dimensional sample dataset) to smaller subspaces k (where $k \leq n - 1$) while maintaining discriminatory class information. In general, reducing dimensions not only helps reduce computing costs for certain classification tasks but can also help to avoid overfitting by minimizing errors in parameter estimation ("curse of dimensions").

Support Vector Machines

SVM is a method used to separate two data sets from two different classes in linearly separable data [6, 7, 9, 13, 14, 15]. The basic concept of Support Vector Machines is simply determining the best hyperplane that functions as a separator for classes in linear input space with maximum margins. Besides working in linear data separation, SVM was also developed to work in non-linear cases. In the case of non-linear SVM includes the concept of kernel tricks to map non-linear data so that it becomes data that can be separated linearly at higher dimensions.

Dataset and Evaluation Model

The dataset used is the Pima dataset [9, 16, 17, 18, 19]. This dataset comes from the National Institute of Diabetes and Digestive and Kidney Diseases. The purpose of using the dataset in this study is to predict diagnostically whether

TABLE I. Accuracy, Sensitivity, Specificity And Auroc Values For All Experiments

Experiments	1	2	3	4	5
Sigma	-3.8	-4.2	-4.4	-4.5	-4.6
Accuracy	77.21	77.08	77.08	77.34	77.34
Sensitivity	73.51	73.13	73.13	73.51	73.13
Specificity	79.2	79.2	79.2	79.4	79.6
AUROC	0.764	0.762	0.762	0.765	0.764

a patient has diabetes. Specifically, the patients referred to in this study are women who are at least 21 years old with and have several characteristics as follows pregnancy, glucose level, blood pressure, skin thickness, insulin, BMI, Diabetes Function, Function, Age, and Class. Evaluation of experiments used in this study are Confusion Matrix [20, 21, 22] and Receiver Operating Characteristic Curve (ROC) [22].

The Confusion Matrix is a performance measure for machine learning classification problems where the output can be two or more classes. The Confusion Matrix is a table with 4 different combinations of predicted and actual values. There are four terms that represent the results of the classification process in the confusion matrix, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), where 1) TP (True Positive): Data on a person suffering from diabetes and predictably also suffering from diabetes. 2) TN (True Negative): Based on the data, a person does not have diabetes and also does not predict diabetes. 3) FP (False Positive): Data on a person does not have diabetes but predictably has diabetes 4) FN (False Negative): Data on someone suffering from diabetes but predictably suffering from diabetes.

The confusion matrix above is used to find the value of Accuracy, sensitivity, specificity, and AUROC. Accuracy in classification is the percentage of accuracy in the classification of sick patients and healthy patients who are classified correctly [22]. In the case of evaluating a patient's disease diagnosis system, sensitivity is the ability of a test to diagnose all sick patients diagnosed with the illness. Meanwhile, specificity is the ability of testing to diagnose all non-sick patients who are diagnosed with no illness. Meanwhile, the Receiver Operating Characteristic Curve (ROC) is used to evaluate the experimental results visually. ROC is a two-dimensional plot with 1-Specificity on the X-axis and Sensitivity on the Y-axis. In order to get the best model, we used 10-fold cross-validation design to get the optimal parameter.

RESULTS AND DISCUSSION

The results of all of the experiments conducted can be seen in Table I. The experimental results contained in Table I include the sigma parameters used during the experiment and the performance of the LDA-SVM method namely accuracy, sensitivity, specificity and AUROC.

Based on the experimental results contained in Table 1, it can be seen that the highest Accuracy value is achieved when the sigma parameters used are -4.5 and -4.6 with an accuracy value of 77,344. For the highest sensitivity obtained, the parameters of -3.8 and -4.5 with the highest value of 73.507. Meanwhile, for the Specificity value, the highest value is 79.6 with the parameter used is -4.6.

Figure 2-6 are the ROC curve of all experiments. The derivative of the ROC curve is the AUROC contained in Table 1. Figure 2 shows the ROC results from the first experiment with the optimal parameter, which is -3.8. After calculating the area under the curve (AUROC), the value is 0.764. In Figure 3, the experimental results with the highest accuracy are obtained when the sigma parameter used is -4.2 with an AUROC value of 0.762. The AUROC results were 0.002 smaller when compared to the AUROC results in the first experiment. In the third experiment, the best sigma parameter obtained was -4.4, while the AUROC value obtained was the same as AUROC in the second experiment, which was 0.762. The fourth experiment produced the ROC curve shown in Figure 5, in this experiment the AUROC value was 0.765. The AUROC value in experiment 4 exceeds the entire AUROC value of all experiments. Figure 6 is the last experiment with the optimal parameter value obtained is -4.6, with the AUROC value obtained is the same as the AUROC value in experiment 1 which is equal to 0.764.

It can be seen in Table 1, that the largest AUROC value is obtained when using the sigma -4.5 parameter with a value of 0.765. So based on four performance references namely Accuracy, Sensitivity, Specificity and AUROC it can be concluded that the best parameter of all experiments is when the model uses the sigma -4.5 parameter.

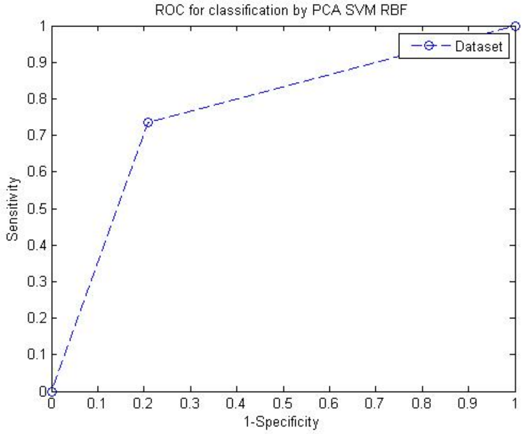

FIGURE 2. Research Framework

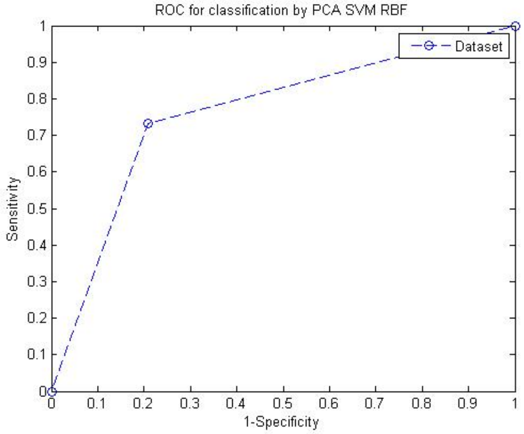

FIGURE 3. Research Framework

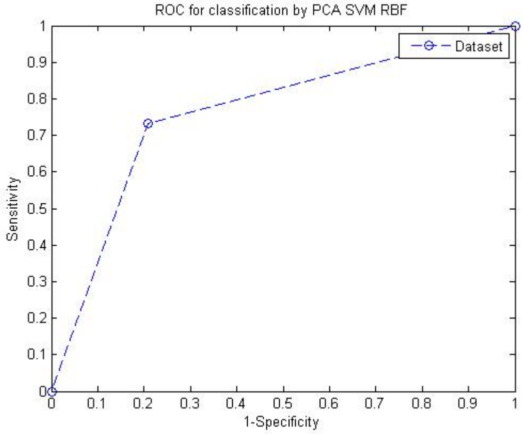

FIGURE 4. Research Framework

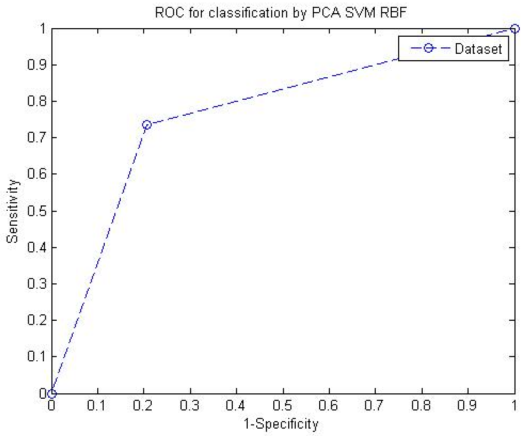

FIGURE 5. Research Framework

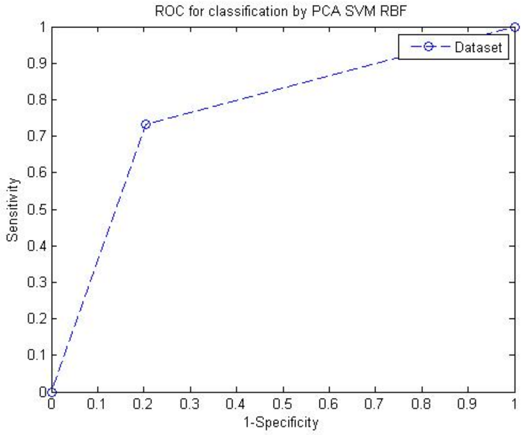

FIGURE 6. Research Framework

CONCLUSION

Linear Discriminant Analysis Method and Support Vector Machines have been used to analyze and build a diabetes detection system. The dataset used in this study is the PIMA dataset which is a dataset commonly used in research on diabetes. In this experiment, the best performance was obtained when using the sigma -4.5 parameter with the values of accuracy, sensitivity, specificity and AUROC in sequence respectively 77.34, 73.51, 79.40, and 0.77. In this study, researchers only focused on using the Radial Basis Function Kernel (RBF) parameter to test the dataset. If we look back at the SVM concept, we will find several other parameters that can be used for experiments, including Linear Kernel, Polynomial Kernel, Gaussian Kernel, Laplace RBF Kernel, Sigmoid Kernel, Anove RBF Kernel, etc. In further research, the author will try to compare several types of kernels to see each performance.

ACKNOWLEDGMENTS

The authors wish to thank Politeknik Negeri Jakarta for providing funding support.

REFERENCES

1. W. H. Organization *et al.*, "Global report on diabetes: executive summary," Tech. Rep. (World Health Organization, 2016).
2. I. M. of Health, "Hari diabetes sedunia," Tech. Rep. (Indonesian Ministry of Health, 2019).
3. S. Arifin and I. B. Muktyas, "Generate a system of linear equation through unimodular matrix using python and latex," in *AIP Conference Proceedings 2331* (AIP, 2021) p. 020005.
4. I. B. Muktyas and S. Arifin, "Digital image encryption algorithm through unimodular matrix and logistic map using python," in *AIP Conference Proceedings 2331* (AIP, 2021) p. 020006.
5. S. Arifin and I. B. Muktyas, "Product of two groups integers modulo m,n and their factor groups using python," in *J. Phys.: Conf. Ser. 1778* (IOP, 2021) p. 012026.
6. A. A. Abdillah *et al.*, "Support vector machines untuk menyelesaikan masalah klasifikasi pada pengenalan pola," Jurnal Poli-Teknologi **18** (2019).
7. A. A. Abdillah and S. Suwarno, "Diagnosis of diabetes using support vector machines with radial basis function kernels," International Journal of Technology **7** (2016).
8. S. Suwarno and A. Abdillah, "Penerapan algoritma bayesian regularization backpropagation untuk memprediksi penyakit diabetes," Indonesian Journal of Mathematics and Natural Sciences **39**, 150–158 (2016).
9. A. A. Abdillah *et al.*, "Pembelajaran mesin menggunakan principal component analysis dan support vector machines untuk mendeteksi diabetes," Jurnal Matematika dan Sains **24** (2019).
10. R. Ren, K. Han, P. Zhao, J. Shi, L. Zhao, D. Gao, Z. Zhang, and Z. Yang, "Identification of asphalt fingerprints based on atr-ftir spectroscopy and principal component-linear discriminant analysis," Construction and Building Materials **198**, 662–668 (2019).
11. M. J. Brusco, C. M. Voorhees, R. J. Calantone, M. K. Brady, and D. Steinley, "Integrating linear discriminant analysis, polynomial basis expansion, and genetic search for two-group classification," Communications in Statistics-Simulation and Computation **48**, 1623–1636 (2019).
12. E. I. Altman, G. Marco, and F. Varetto, "Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the italian experience)," Journal of banking & finance **18**, 505–529 (1994).
13. C. M. Biship, "Pattern recognition and machine learning (information science and statistics)," (2007).
14. B. Schölkopf, A. J. Smola, F. Bach, *et al.*, *Learning with kernels: support vector machines, regularization, optimization, and beyond* (MIT press, 2002).
15. N. Cristianini, J. Shawe-Taylor, *et al.*, *An introduction to support vector machines and other kernel-based learning methods* (Cambridge university press, 2000).
16. M. Kriještorac, A. Halilović, and J. Kevric, "The impact of predictor variables for detection of diabetes mellitus type-2 for pima indians," in *International symposium on innovative and interdisciplinary applications of advanced technologies* (Springer, 2019) pp. 388–405.
17. L. Nass, S. Swift, and A. Al Dallal, "Indepth analysis of medical dataset mining: a comparitive analysis on a diabetes dataset before and after preprocessing," KnE Social Sciences , 45–63 (2019).
18. T. K. Chan and C. S. Chin, "Health stages diagnostics of underwater thruster using sound features with imbalanced dataset," Neural Computing and Applications **31**, 5767–5782 (2019).
19. A. Asuncion and D. Newman, "Uci machine learning repository," (2007).
20. Z. Rustam, A. Kamalia, R. Hidayat, F. Subroto, and A. Suryansyah, "Comparison of fuzzy c-means, fuzzy kernel c-means, and fuzzy kernel robust c-means to classify thalassemia data," Update **1**, 1 (2019).
21. A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," Pattern Recognition **91**, 216–231 (2019).
22. F. Gorunescu, *Data Mining: Concepts, models and techniques*, Vol. 12 (Springer Science & Business Media, 2011).