JESICA: Journal of Enhanced Studies in Informatics and Computer Applications
Journal homepage: jesica.itsk-soepraoen.ac.id
e-ISSN: 3046-6938 p-ISSN: 3046-6997
Vol. 1, No. 2, July 2024, Pp. 47-55

The Discriminant Analysis Function Was Implemented to Predict the Presence of Diabetes

Herry Prasetyo Wibowo[1], Mochammad Anshori*[2], M. Syauqi Haris[3]
1,2,3 Institut Teknologi, Sains, dan Kesehatan RS.DR. Soepraoen Kesdam V/BRW, Indonesia

***Corresponding author**
E-mail address:
moanshori@itsk-soepraoen.ac.id

Keywords:
Discriminant analysis, diabetes, machine learning, prediction, LDA, QDA

Abstract
Diabetes is a condition where blood sugar concentrations are high and there is something wrong with insulin inside the body. A hormone called insulin controls the equilibrium of blood sugar concentration in humans. Diabetes has high-risk health, such as CKD, CVD, skin disease or even blindness. The reason people suffer from diabetes is caused by bad consumption habits. Some symptoms of diabetes are frequent urination and feeling hungry too quickly. Diabetes is sometimes difficult to diagnose, which is why it is also referred to as the silent killer. A preventive way is an early prediction of diabetes disease. This is very important to do. In this study, the discriminant analysis algorithm is used along with machine learning techniques. In this study, machine learning techniques are used. Its name is discriminant analysis algorithm. Two popular versions are linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). This method is used because it is suitable for high-dimensional data and the discriminant analysis algorithm has minimal parameters. The discriminant analysis algorithm uses few parameters, and this method is appropriate for high-dimensional data. We'll compare the two approaches to find a way to demonstrate their dependability. Both approaches would be contrasted. Based on the result, QDA has the best performance. QDA can produce accuracy = 93.7%, TPR = 93.7%, precision = 94.3%, recall = 93.7% and F-measure = 93.9%. FPR of QDA is the lowest one, it is 1.02%. It means QDA has a small error in making predictions. Overall, based on the result QDA is the proven and proper method for detecting diabetes disease.

1. Introduction
A condition when high levels of blood sugar (blood glucose) caused metabolic disease is namely diabetes. Diabetes makes the human body cannot properly respond to insulin [1]. The hormone insulin is responsible for controlling the equilibrium of blood sugar levels. Hyperglycemia, or elevated blood glucose, is a sign that the insulin is not functioning correctly. The term is impaired glucose tolerance (IGT). The effect of diabetes suffering over time has serious risks such as kidney disease (CKD), heart disease (CVD), damaged blood vessels, blindness (diabetes retinopathy), and skin problems [1]–[4]. So, it is included as one of the most dangerous diseases in the world. According to the WHO, 422 million individuals in low- and middle-income countries have diabetes. It is predicted that there will be 642 million cases of diabetes worldwide in 2040 [5].

Diabetes comes in two varieties: type 1 and type 2. Type 1 occurs because the pancreas produces little insulin, and type 2 is the human body resists insulin so there is not enough insulin in the body [4],[6]. A reason that causes diabetes is bad consumption habits. As an example, Indonesian people like to consume high calories food without protein, vitamins, and fat balancing [7]. In addition, psychological factors like mental health conditions, cognitive dysfunction, personality traits, and quality of life can also contribute to it. [8]. A prolonged period of high blood glucose concentration is the primary sign of diabetes, frequent urination, blurry eyes, abnormal loss of weight and feel hungry too quickly [9]. Diabetes can be detected from urine checks, blood pressure tests, kidney tests and biopsy. But studies show that 40% of diabetic subjects do not show initial screening. Diabetes also called as a silent killers beacause the undetectable symptoms. One concern of diabetes is that it can cause several health issues or possibly early mortality if it is not identified and treated promptly. Therefore, it is necessary to have a good system to detect diabetes as a preventive method. In this research, we take advantage of machine learning classifier implementation to predict diabetes early.

In the last ten years, machine learning has been used extensively in the healthcare industry. Machine learning has a great deal of potential to raise human standards and move toward a peaceful existence due to its dependability and efficiency. There is a lot of data stored in different formats because of the current digitalization era. Machine learning

can use this data to examine patterns or hidden knowledge within a set of data. Since machine learning is a data-driven approach, learning requires data.

There are algorithms in the machine learning method that was used in previous research to predict diabetes disease. Such as support vector machine (SVM) reach accuracy about 84,10%, 100%, and 82% [2], [5], [10]. SVM reliably handles data that is separated both linearly and nonlinearly by constructing a hyperplane that divides data classes. An algorithm known as random forest (RF) uses multiple decision trees to produce decisions and conduct majority voting. RF could reach accuracy about 84%; 79%; and 98% [9], [11], [12]. Logistic model tree (LMT) is a tree model based on logistic function. This method canreach accuracy = 79.31% [13]. Fuzzy is one of soft computing method that reach accuracy about 79,8% [6]. Artificial neuro fuzzy inference system (ANFIS) reach accuracy = 84.48% [4]. Decision tree (DT) is tree based model to do prediction and reach accuracy about 91.2% [14]. It is proven that machine learning can be implemented to predict diabetes disease.

2. Research Method

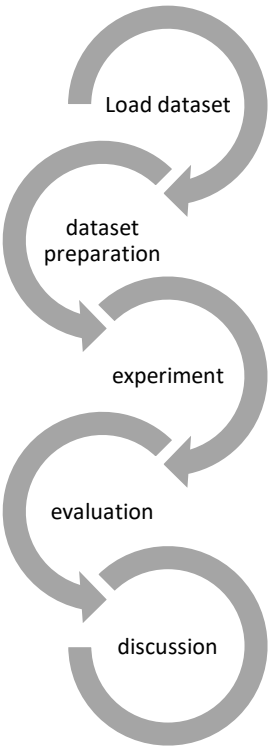

Figure 1. Research methodology

The research approach employed in this study is depicted in Figure 1. Generally, there are 4 stages that are load the obtained dataset; dataset preparation by doing some data preprocessing from raw data until already to use; experiment is implementing the machine learning algorithm, there are LDA and QDA; evaluation to measure performance of the classifier models; and discussion to gain the best model based on experimental phase. The details will be explained below.

The first stage is the dataset must be loaded. Dataset pulls from Mendeley dataset [15] and consists of 1000 row instance data. More description about dataset shown at

Table *1*. From the table there are 14 features such as id, no of patient, gender, age, urea, HBA1C, creatinine ratio (Cr), cholesterol (Chol), HDL cholesterol triglycerides (TG), LDL, VLDL, body mass index (BMI) and target class. There are three class targets, namely yes diabetic, nondiabetic, predict-diabetic. Information from patient medical records, including personal information and laboratory test results, is where the data is gathered from.

The data type of each data consists of numeric and nominal type and have various range of data. The second stage is dataset preparation. The data will be cleaned first before doing an experiment. Feature ID and No of the patient will be removed because it has no correlation to the result. By reducing these 2 features, we can also reduce the high dimensionality of the data. The gender with nominal type will be transformed into numeric. In this phase, we use ordinal encoding to convert categorical into numerical values [16]. The conversion process is by assigning a value in the form

of an integer to each category. Conversion based on the number of known categories. Since the target class does not need to be converted into numeric form, the data type in the class feature is left unchanged.

$$\text{new\_data} = \frac{\text{curr\_data} - \text{min\_data}}{\text{max\_data} - \text{min\_data}} \quad (1)$$

Table 1. The details of dataset

Feature	Range	Type
ID	1-800	Numeric
No	123 - 75435657	Numeric
Gender	F, M	Nominal
Age	20 - 79	Numeric
Urea	0.5 - 38.9	Numeric
Cr	6 - 800	Numeric
HbA1c	0.9 - 16	Numeric
Chol	0 - 10.3	Numeric
TG	0.3 - 13.8	Numeric
HDL	0.2 - 9.9	Numeric
LDL	0.3 - 9.9	Numeric
VLDL	0.1 - 35	Numeric
BMI	19 - 47.75	Numeric
Class	Y, N, P	Nominal

Based on

Table *1*, the data range is very diverse. To ensure that all data values have the same minimum and maximum bounds, the range of values must be uniformed [17]. The min-max scaler will be used for normalization once the next data type expectation is fully numeric. The formula to compute normalization using min-max scaler shown at Equation (1). To get new normalized data, this formula uses current data, minimal value and maximal value. The data value range that results from applying the min-max scaler is between 0 and 1. Another advantage of min-max scaler normalization is that it can speed up the computing process because the data range is no longer large. In some case, by implementing min-max scaler to do data normalization, it also can increase the evaluation measure such as accuracy [18].

$k1$	$k2$	$k3$	$k4$	$k5$	$k6$	$k7$	$k8$	$k9$	$k10$
$k1$	$k2$	$k3$	$k4$	$k5$	$k6$	$k7$	$k8$	$k9$	$k10$
$k1$	$k2$	$k3$	$k4$	$k5$	$k6$	$k7$	$k8$	$k9$	$k10$
$k1$	$k2$	$k3$	$k4$	$k5$	$k6$	$k7$	$k8$	$k9$	$k10$
$k1$	$k2$	$k3$	$k4$	$k5$	$k6$	$k7$	$k8$	$k9$	$k10$
$k1$	$k2$	$k3$	$k4$	$k5$	$k6$	$k7$	$k8$	$k9$	$k10$
$k1$	$k2$	$k3$	$k4$	$k5$	$k6$	$k7$	$k8$	$k9$	$k10$
$k1$	$k2$	$k3$	$k4$	$k5$	$k6$	$k7$	$k8$	$k9$	$k10$
$k1$	$k2$	$k3$	$k4$	$k5$	$k6$	$k7$	$k8$	$k9$	$k10$
$k1$	$k2$	$k3$	$k4$	$k5$	$k6$	$k7$	$k8$	$k9$	$k10$

Figure 2. Illustration of k=10 cross validation

$$f(x) = \frac{1}{(2\pi)^{p/2}|C(X)|^{.5}} \exp\left(-0.5(x-\mu)^T \left(C(X)\right)^{-1}(x-\mu)\right) \quad (2)$$

$$\delta_k(x) = x^T \textstyle\sum_k^{-1} \mu_k - 0.5\mu_k^T\left(C(X)\right)^{-1}\mu_k + \log(\pi_k) \quad (3)$$

The third stage is performing the discriminant analysis classifier. Before creating the model, the dataset will be split into training and test data by implementing cross-validation with k=10. Figure 2 shows illustraton of cross-validation with k=10 [19]. Ten sections will be created from the data. There will be two segments of data: one for testing and the other for training. Based on the number of k values, this test data will shift from each segment ten times

iteratively. Cross-validation helps us to make a more general model and avoid the overfitting of the classifier [20]. Then the data is ready to carry out the experiment. A classifier that the research use is discriminant analysis. There are two common kinds of discriminant analysis to compare, there are linear and quadratic. The aim of this comparison is to get the best model for this case. Because it directs the selection of a suitable discriminant analysis method and provides a gauge for the quality of the method that is ultimately selected [21].

One technique for supervised learning is linear discriminant analysis, or LDA to do classification tasks and can be used for pre-processing data such as dimensionality reduction [22]–[25]. In various classification and pattern recognition problems, this dimension reduction, for example, makes the data matrix smaller than the original data. Supervised learning is the foundation of this data transformation for the process to consider the data class and generate low-dimensional data points that are near the class group. The mathematical function of LDA is to maximize heterogeneous data and minimize homogeneous data [23]. The equation of LDA can be shown at Equation (2). The manual calculation of LDA is based on three steps, there are calculate inner class, calculating between classes, and reconstructing data. LDA is very suitable for linearly separated data or data that only has two targets of classes. This equation requires of covariance and pooled covariance of data. This step is also applied for data dimension reduction. Then bayesian theory is implemented to calculate probability of its class. Where μ is mean, C means covariance, x is data input and k are class target.

Equation (3) shows the mathematical function of quadratic discriminant analysis (QDA). In Equation (2), the μ_k assuming the covariance matrix for remains the same for all classes. For QDA, μ_k ise used to calculates the covariance matrix for each class [26]. It is evident from Equation (3) that a logarithmic function exists. The ability of QDA to solve quadratic problems is attributed to this function. The advantage of LDA and QDA is not require some parameter to optimize their model, also known as nonparametric function. These two techniques work well in situations where the data is high-dimensional as well. [27].

$$Accuracy = \frac{TN+TP}{total\ row\ data} \times 100\% \quad (4)$$

$$TPR = \frac{TP}{TP+FN} \times 100\% \quad (5)$$

$$FPR = \frac{FP}{FP+FN} \times 100\% \quad (6)$$

$$precision = \frac{TP}{TP+FP} \times 100\% \quad (7)$$

$$recall = \frac{TP}{TP+FN} \times 100\% \quad (8)$$

$$F - measure = 2 \times \frac{recall \times precision}{recall + precision} \times 100\% \quad (9)$$

The fourth stage is evaluating the model based on some evaluation metrics. In this discussion, we used accuracy, TPR, and FPR evaluation. Accuracy to know how best the model to predict and shown in Equation (4). TPR uses to know the correct rate of classifier and the equation shown in Equation (5). Equation (6) to calculate FPR to know the wrong rate of the model predictor [28].TP = true positive, FN = false negative, TN = true negative, and FP = false negative. This variable is obtained from the confusion matrix. Precision to know how precise the model to predict and recall to knowing all the positive rate. Each formula is shown in Equation (7) and (8). Equation (9) presents the formulation for the application of F-Measure evaluation. Since we have the values of FPR and TPR, these values will be used to create the ROC (receiver operating characteristics) graph. The value of the ROC graph can be used to model errors and determine the reliability of the algorithm in carrying out classification [29]. The evaluation computation makes use of the calculated precision and recall values. Applications of the F-measure's benefits to unbalanced data are appropriate. As a result, the f-measure is also used to enrich the discriminant analysis evaluation results. These metrics will be used in the next discussion to gain the optimum classifier between LDA and QDA.

3. Results and Discussion

According to the experiment, the result and discussion by observing LDA and QDA will be explained here. First, we show the materials that used in this study and discuss the details. Then elucidate the results of the experiment to obtain and clarify the best method based on the comparison.

The distribution of the target dataset is shown in Figure 3. There 5.3% or 53 data are predicted to have diabetes, 10.3% as non-diabetes about 103 data, and the remaining 844 data are yes-diabetes about 84.4%. It shows if yes-diabetes is dominance than other classes. The imbalance of the class is indicated by the display. Therefore, using the f-measure evaluation in this study makes perfect sense.

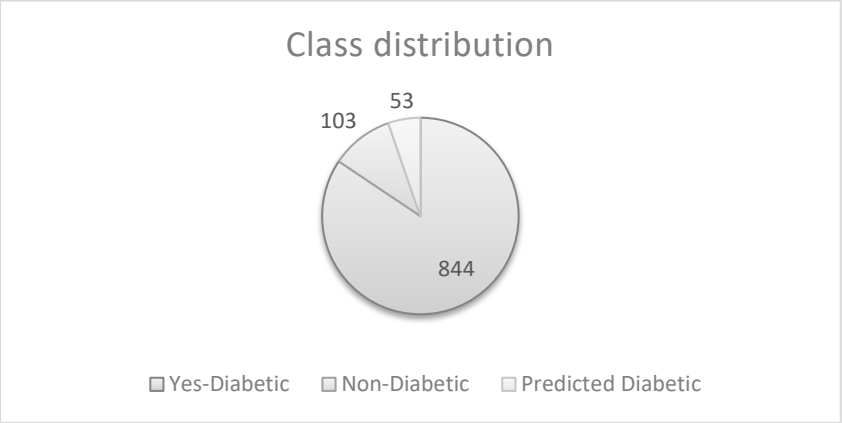

Figure 3. Class distribution

The total attribute in this experiment is 12 features after removing 2 features. Gender features have been transformed into numeric by changing female to 1, and male to 2. The name of the method is ordinal coding. A new dataset details shown in Table 2. With 12 variables, the current dataset has a dimension of 12D. The more complex the data is to visually represent, the higher its dimensions. In addition, it necessitates high computing. Processing the data thus becomes a challenge in and of itself. Thus, another goal of this study is to demonstrate the applicability of discriminant analysis to the situation of diabetes. The current dataset with features in the form of numbers is ready for learning because the similarities between LDA and QDA are based on functions that use numerical data.

Table 2. Dataset form after cleaned

Feature	Range	Type
Gender	1, 2	Numeric
Age	20 - 79	Numeric
Urea	0.5 - 38.9	Numeric
Cr	6 - 800	Numeric
HbA1c	0.9 - 16	Numeric
Chol	0 - 10.3	Numeric
TG	0.3 - 13.8	Numeric
HDL	0.2 - 9.9	Numeric
LDL	0.3 - 9.9	Numeric
VLDL	0.1 - 35	Numeric
BMI	19 - 47.75	Numeric
Class	Y, N, P	Nominal

The data is ready to carry out the model. In this experiment, we split data into 10 segments, and each segment is the test data. It happened iteratively 10 times. So, it is called k-10 cross-validation. Following the application of both discriminant analysis algorithms, the comparison outcomes will be displayed. Then LDA and QDA are implemented to get the optimum model between its discriminant analysis. The result shown in Table 3. In experiment we use a tool to namely Weka [30].

Table 3. Evaluation metric result

Evaluation	LDA	QDA
Accuracy	89,4	93,7
TPR	89,4	93,7
FPR	1,8	1,02
Precision	89,6	94,3
Recall	89,4	93,7
F-measure	89,4	93,9
ROC	91,9	97,7

Accuracy, false positive rate (FPR), true positive rate (TPR), recall and precision are shown in Table 3 above. It is evident that QDA typically offers greater value than LDA. In this instance, LDA which is typically effective for a wide range of problems is insufficient, and QDA produces superior evaluation outcomes. We can see if the accuracy of LDA

= 89.4%, which is lower than QDA that reaches 93.7%. TPR has the same result with its accuracy. Precision of LDA = 89.64, it is lower than QDA = 94.3%. The recall value of LDA reaches 89.4% less than QDA = 93.7%. The FPR has a contrary result, FPR of LDA = 1.8% is higher than QDA = 1.02%. The higher the FPR value, the worse the model because there are many errors in the prediction. F-measure performed in LDA = 89.4%, and QDA = 93.9%. This shows that the f-measure percentage of QDA is better than LDA. Now ROC values are discussed. ROC of LDA is 91,9%, and QDA is 97,7%. In this measure, QDA still get the better percentage than LDA. Overall based on evaluation measures, it means QDA could give a better result than LDA. Almost all the evaluation shows if LDA evaluation is lower than QDA.

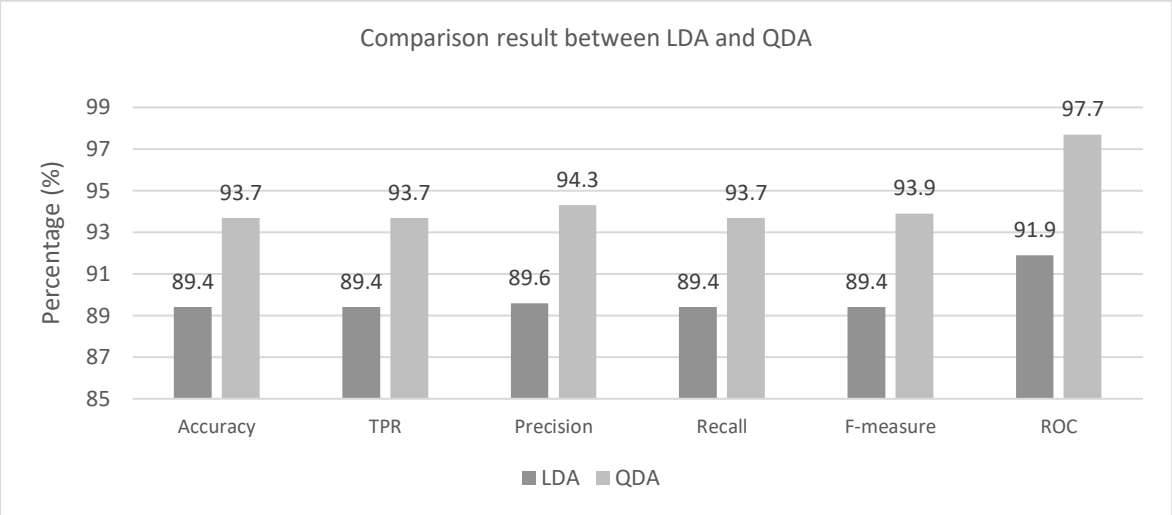

Figure 4. Comparison evaluation measure of LDA and QDA

Figure 4 is a comparison result between LDA and QDA. The evaluation such as accuracy, TPR, precision, recall and f-measure percentage from QDA is higher than the LDA result. QDA can produce more than about 90% result, while LDA yields a lower score of about 80%. LDA accuracy = 89.4%, TPR = 89.4%, recall = 89.4% and f-measure = 89.4%. The difference result between LDA and QDA based on accuracy, TPR and recall are 4.3%. A difference in precision is 4.7%, LDA = 89.4% and QDA = 93.7%. The result shows a big difference score between the method. Based on this figure, we can get if QDA can give better results. It is proven if the model of QDA is more precise and accurate then LDA. When it comes to harmonic evaluation, QDA can also yield a higher f-measure percentage than LDA. Next FPR comparison is discussed.

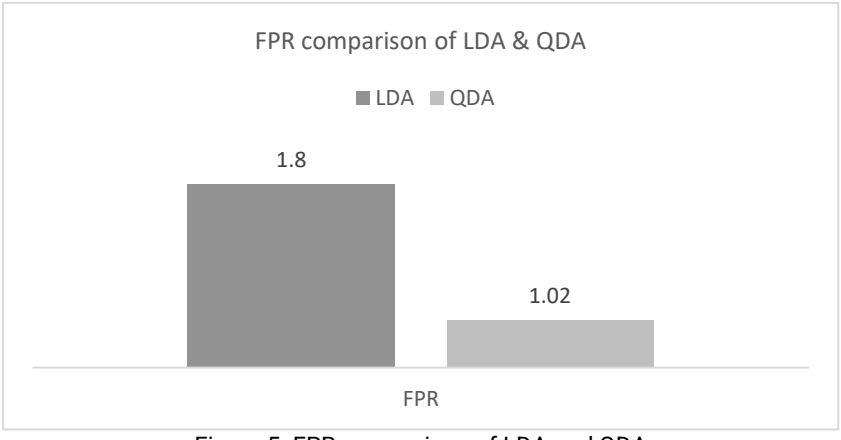

Figure 5. FPR comparison of LDA and QDA

The comparison of the FPR results is shown in Figure 5. FPR used to know how many negative events among the negative ones are mistakenly labeled as positive. From the figure of FPR comparison, LDA give higher value than QDA. FPR of LDA = 1.8%, QDA = 1.02%. The FPR difference is 0.78%. A tiny value that is near to zero is an excellent FPR. The model is more optimal the closer it gets to zero. It means LDA gives more errors while predicting the data. It shows that LDA performance has much miss classified on the model. In the health sector, it is necessary to use a model that has

an accurate level or a small error in prediction because it is related to health and even life. Based on the figure, we can get the result if QDA can perform better to predict diabetes disease.

```
  a   b   c   <-- classified as
 85  12   6 |   a = N
 14  13  26 |   b = P
 26  22 796 |   c = Y
```

Figure 6. Confusion matrix of LDA model

Figure 6 above show the confusion matrix result based on LDA prediction model. The predictions, 796 predicted data for actual diabetes, 13 predicted possible diabetes, and 85 predicted no diabetes at all. The 106 remaining data points were included in incorrect predictions. The error in the data classification is roughly 10.6%. Because of the significant number of prediction errors, the model is unsuitable for use.

```
  a   b   c   <-- classified as
 94   1   8 |   a = N
  0  43  10 |   b = P
 34  10 800 |   c = Y
```

Figure 7. Confusion matrix of QDA model

Figure 7 contains the confusion matrix of the QDA classifier. According to the confusion matrix, there were 94 nondiabetic predictions, 43 predicted diabetes, and 800 positive diabetes. The accuracy reaches 93.7%, an indication that most of the classes were predicted correctly. The 6.4% remain said the wrong classify. This error value is lower than that of LDA. Thus, it is evident that QDA is superior to LDA.

Since we get the TPR and FPR values refer to the Table 3, ROC values can be produced. The value of ROC has been shown in Table 3 before. Based on ROC before, here we will discuss the AUC (area under curve) graph. There are x and y axes in this two-dimensional graph. TPR is the x-axis and FPR is the y-axis.

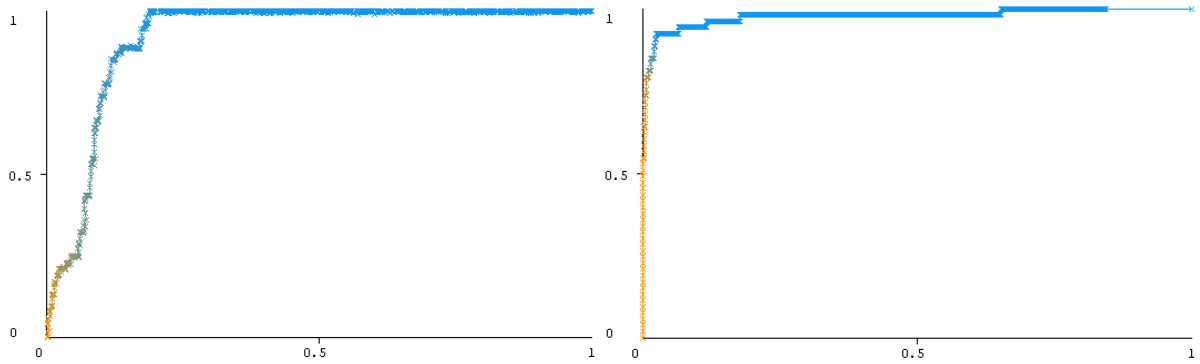

Figure 8. ROC curve of LDA (left) and QDA (right)

The area under the curve (AUC) for LDA is shown in Figure 8 (left) above. This graph can be said to be good because ROC = 0.919 which shows that the value is close to 1. There is only a difference about 0.081. A good ROC value is one that is close to 1. If it is close to 0, then the model is very bad. The threshold value for ROC is 0.5 as the threshold point. So, the ROC from LDA can still be said to be good.

Figure 8 (right) above is a visualization of the ROC curve of the QDA model. It can be seen from this visualization that the line almost fills the graph perfectly. The graphics have come a long way since the start. As seen on the y-axis, this suggests that the model has a high TPR value. The value of ROC for QDA is 0.977. This value is more than 0.5 and very close to 1, The difference is merely 0.023. The model is fitter than the LDA previously discussed. In this case, the performance of QDA is close to its exact value due to the extremely slight difference and proves that QDA is suitable for predicting diabetes disease.

In this research, based on all experiments that have been carried out from data preprocessing to testing, the evaluation metric of LDA gives a high value but didn't touch the QDA score. The performance of the LDA model approaches a value of more than 80% when examining the evaluation metrics that have been generated. It is proven that LDA can be used to classify diabetes disease causes the evaluation scores are high. But QDA performs more properly than LDA. QDA can be fitter to do classification on this data case. This is indicated by the high evaluation measure value, around more than 90%. Because QDA provides the most optimum performance, it can be said that the dataset is spread nonlinearly, and quadratic discriminant analysis is the best learning to do prediction diabetes disease.

4. Conclusion

Machine learning approaches have been used wide area including the health field. To prevent and detect diabetes early on, machine learning techniques have been used to forecast the condition. This research uses discriminant analysis algorithm, such as LDA and QDA algorithms as comparison algorithms. The outcome of this study is LDA has less performance than QDA. LDA results are lower than QDA, whereas QDA yields accuracy = 93.7%, TPR = 93.7%, precision = 94,3%, recall = 93,7%, and F-measure = 93.9%. QDA produces the lowest FPR about 1.02%, which is less than LDA. It means LDA has more misclassifying than QDA. The ROC value shows the performance reliability of the classifier. ROC of QDA = 0,977. The value closest to 1 indicates that the model performance is very capable. It is proven that discriminant analysis can predict diabetes disease. Especially QDA because the data is spread nonlinearly and can solve high-dimensional data, so QDA result is better than LDA. Based on the result, it shows if the quadratic function is proper to do prediction in this data. For further work, the implementation of LDA is recommended to reduce the dimensions and then classify them. The primary characteristic of the dataset is its high dimension, which is a result of the abundance of features it contains

References

[1] K. Lakhwani, S. Bhargava, K. K. Hiran, M. M. Bundele, and D. Somwanshi, "Prediction of the Onset of Diabetes Using Artificial Neural Network and Pima Indians Diabetes Dataset," *2020 5th IEEE Int. Conf. Recent Adv. Innov. Eng. ICRAIE 2020 - Proceeding*, vol. 2020, 2020, doi: 10.1109/ICRAIE51050.2020.9358308.

[2] H. Abbas, L. Alic, M. Rios, M. Abdul-Ghani, and K. Qaraqe, "Predicting diabetes in healthy population through machine learning," *Proc. - IEEE Symp. Comput. Med. Syst.*, vol. 2019-June, pp. 567–570, 2019, doi: 10.1109/CBMS.2019.00117.

[3] K. Vijiyakumar, B. Lavanya, I. Nirmala, and S. Sofia Caroline, "Random forest algorithm for the prediction of diabetes," *2019 IEEE Int. Conf. Syst. Comput. Autom. Networking, ICSCAN 2019*, pp. 1–5, 2019, doi: 10.1109/ICSCAN.2019.8878802.

[4] L. Priyadarshini and L. Shrinivasan, "Design of an ANFIS based Decision Support System for Diabetes Diagnosis," *Proc. 2020 IEEE Int. Conf. Commun. Signal Process. ICCSP 2020*, pp. 1486–1489, 2020, doi: 10.1109/ICCSP48568.2020.9182163.

[5] G. A. Pethunachiyar, "Classification of diabetes patients using kernel based support vector machines," *2020 Int. Conf. Comput. Commun. Informatics, ICCCI 2020*, pp. 22–25, 2020, doi: 10.1109/ICCCI48352.2020.9104185.

[6] I. O. Lixandru-Petre, "A fuzzy system approach for diabetes classification," *2020 8th E-Health Bioeng. Conf. EHB 2020*, 2020, doi: 10.1109/EHB50910.2020.9279882.

[7] Y. Sinatrya and L. A. Wulandhari, "Deteksi Diabetes Melitus Untuk Wanita Dan Penyusunan Menu Sehat Dengan Pendekatan Adaptive Neuro Fuzzy Inference System (Anfis) Dan Algoritma Genetika (Ga)," *J. Tek. Inform.*, vol. 12, no. 1, pp. 39–58, 2019, doi: 10.15408/jti.v12i1.9578.

[8] N. A. Bhat, K. P. Muliyala, and S. Kumar, "Psychological Aspects of Diabetes," *Eur. Med. J.*, no. November, pp. 90–98, 2020, doi: https://doi.org/10.33590/emjdiabet/20-00174.

[9] R. Katarya and S. Jain, "Comparison of different machine learning models for diabetes detection," *Proc. 2020 IEEE Int. Conf. Adv. Dev. Electr. Electron. Eng. ICADEE 2020*, no. Icadee, pp. 0–4, 2020, doi: 10.1109/ICADEE51157.2020.9368899.

[10] N. Mohan and V. Jain, "Performance Analysis of Support Vector Machine in Diabetes Prediction," *Proc. 4th Int. Conf. Electron. Commun. Aerosp. Technol. ICECA 2020*, pp. 2020–2022, 2020, doi: 10.1109/ICECA49313.2020.9297411.

[11] S. Benbelkacem and B. Atmani, "Random forests for diabetes diagnosis," *2019 Int. Conf. Comput. Inf. Sci. ICCIS 2019*, pp. 1–4, 2019, doi: 10.1109/ICCISci.2019.8716405.

[12] A. S. Alanazi and M. A. Mezher, "Using Machine Learning Algorithms for Prediction of Diabetes Mellitus," *2020 Int. Conf. Comput. Inf. Technol. ICCIT 2020*, vol. 02, pp. 55–57, 2020, doi: 10.1109/ICCIT-144147971.2020.9213708.

[13] D. Vigneswari, N. K. Kumar, V. Ganesh Raj, A. Gugan, and S. R. Vikash, "Machine Learning Tree Classifiers in Predicting Diabetes Mellitus," *2019 5th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2019*, pp. 84–87, 2019, doi: 10.1109/ICACCS.2019.8728388.

[14] A. M. Posonia, S. Vigneshwari, and D. J. Rani, "Machine learning based diabetes prediction using decision tree J48," *Proc. 3rd Int. Conf. Intell. Sustain. Syst. ICISS 2020*, pp. 498–502, 2020, doi: 10.1109/ICISS49785.2020.9316001.

[15] A. Rashid, "Diabetes Dataset," vol. 1, 2020, doi: 10.17632/WJ9RWKP9C2.1.

[16] Kedar Potdar, Taher S. Pardawala, and Chinmay D. Pai, "A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers," *Int. J. Comput. Appl.*, vol. 175, no. 4, p. 375, 2017, doi: 10.5120/ijca2017915495.

[17] M. Anshori, N. Rikatsih, M. S. Haris, T. Kesehatan, I. Rs, and S. Kesdam, "PREDIKSI PASIEN DENGAN PENYAKIT KARDIOVASKULAR MENGGUNAKAN RANDOM FOREST," *TEKTRIKA*, vol. 7, no. 2, pp. 58–64, 2023.

[18] H. Henderi, "Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer," *Int. J. Informatics Inf. Syst.*, vol. 4, no. 1, pp. 13–20, 2021, doi: 10.47738/ijiis.v4i1.73.

[19] M. Anshori, M. S. Haris, and W. Teja Kusuma, "Penerapan Backpropagation Neural Network (BPNN) Untuk Prediksi Kecanduan Smartphone Pada Remaja," *Cices*, vol. 9, no. 2, pp. 192–202, 2023, doi: 10.33050/cices.v9i2.2701.

[20] M. Anshori, "Prediction Result of Dota 2 Games Using Improved SVM Classifier Based on Particle Swarm Optimization," *2018 Int. Conf. Sustain. Inf. Eng. Technol.*, pp. 121–126, 2018, doi: 10.1109/SIET.2018.8693204.

[21] S. Guo and H. Tracey, "Discriminant Analysis for Radar Signal Classification," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 56, no. 4, pp. 3134–3148,

2020, doi: 10.1109/TAES.2020.2965787.
[22] M. Anshori, F. Mahmudy, and A. A. Supianto, "Preprocessing Approach for Tuberculosis DNA Classification using Support Vector Machines (SVM)," *J. Inf. Technol. Comput. Sci.*, vol. 4, no. 3, pp. 233–240, 2019, doi: https://doi.org/10.25126/jitecs.201943113.
[23] G. B. G. Pereira, L. P. Fernandes, J. M. R. D. S. Neto, H. D. D. M. Braz, and L. D. S. Sauer, "A comparative study of linear discriminant analysis and an artificial neural network performances in breast cancer diagnosis," *2020 Ieee Andescon, Andescon 2020*, 2020, doi: 10.1109/ANDESCON50619.2020.9272057.
[24] J. Ghosh and S. B. Shuvo, "Improving Classification Model's Performance Using Linear Discriminant Analysis on Linear Data," *2019 10th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2019*, pp. 8–12, 2019, doi: 10.1109/ICCCNT45670.2019.8944632.
[25] A. Setya Budi, N. Merlina, M. Arie Hasan, D. Riana, and S. Hadianti, "Classification of Lycopersicon Esculentum Fruit Based on Color Features with Linear Discriminant Analysis (LDA) Method," *Proc. 2019 4th Int. Conf. Informatics Comput. ICIC 2019*, pp. 5–10, 2019, doi: 10.1109/ICIC47613.2019.8985787.
[26] M. R. Wasef and N. Rafla, "HLS implementation of linear discriminant analysis classifier," *Proc. - IEEE Int. Symp. Circuits Syst.*, vol. 2020-Octob, no. 1, pp. 2–5, 2020, doi: 10.1109/iscas45731.2020.9181270.
[27] Y. Wu, Y. Qin, and M. Zhu, "Quadratic Discriminant Analysis For High-Dimensional Data," *Stat. Sin.*, vol. 29, pp. 939–960, 2019, doi: 10.5705/ss.202016.0034.
[28] T. M. Rausch, N. D. Derra, and L. Wolf, "Predicting online shopping cart abandonment with machine learning approaches," *Int. J. Mark. Res.*, vol. 64, no. 1, pp. 89–112, 2022, doi: 10.1177/1470785320972526.
[29] M. Anshori and M. S. Haris, "Predicting Heart Disease using Logistic Regression," *Knowl. Eng. Data Sci.*, vol. 5, no. 2, p. 188, 2022, doi: 10.17977/um018v5i22022p188-196.
[30] E. Frank, M. A. Hall, and I. H. Witten, "The WEKA workbench," *Data Min.*, pp. 553–571, 2017, doi: 10.1016/b978-0-12-804291-5.00024-6.