DETECTING DIABETIC RETINOPATHY WITH SUPERVISED LEARNING

[1*]ADITHYA KUSUMA WHARDANA, [2]PARMA HADI RANTELINGGI

[1]Department of Informatics Engineering, Tanri Abeng University

Jl. Swadarma Raya No.58, Ulujami, Kec. Pesanggrahan, Kota Jakarta Selatan

[2]Graduate School Science and Technology, Keio University, Japan

2-15-45 Mita, Minato City, Tokyo 108-0073, Japan

e-mail: [1]adithya@tau.ac.id, [2]parma@ohtsuki.ics.keio.ac.id

[*]Corresponding author

ABSTRACT

Diabetic retinopathy is a common complication that occurs in people with diabetes mellitus. Diabetic retinopathy damage is characterized by the blood vessel system in the layer at the back of the eye, especially in tissues that respond to light. This research aims to detect diabetic retinopathy early using SVM and Random forest. SVM is a classification technique that divides the input space into two classes. Random Forest is a supervised learning algorithm that utilizes a collection of decision trees trained using the bagging method. This research uses datasets from diaretdb1 and messidor to evaluate the performance of both methods. The diaretdb1 dataset consists of 178 data points diagnosed with Proliferative Diabetic Retinopathy and Non-Diabetic Retinopathy. In addition, the messidor dataset consists of 105 data points diagnosed with Diabetic Retinopathy and Non-Diabetic Retinopathy. Experimental results on the diaretdb1 dataset showed that SVM achieved 88% accuracy, while Random Forest achieved 91% accuracy. Similarly, on the messidor dataset, SVM achieved 80% accuracy, while Random Forest achieved 85% accuracy.

Keywords: *Diabetic Retinopathy, Machine Learning, Support Vector Machine, Random Forest*

1. INTRODUCTION

Diabetic retinopathy arises as a consequence of eye conditions that occur in individuals with diabetes and is a major source of visual impairment in diabetic individuals with the condition. It is a chronic eye condition that can lead to complete visual impairment and inability to see. Diabetic retinopathy is categorized into two types: Non-Proliferative Diabetic Retinopathy (NPDR) and Proliferative Diabetic Retinopathy (PDR). NPDR is the early stage of disease progression, while PDR is a more advanced and prolonged phase. In PDR, patients may experience the growth of new blood vessels in the eye, which may eventually lead to blindness [1].

The human eye serves as the primary sensory organ responsible for vision, which plays a vital role in seeing and understanding the world around us. It is important to prioritize protecting our eyes against diseases that can affect vision. Detecting such diseases at an early stage is crucial as it allows timely intervention, promotes effective treatment, and prevents the onset of blindness to ensure optimal visual health and well-being [2].

Microvascular complications are often seen in individuals with diabetes, with approximately 40% of patients developing diabetic retinopathy. The incidence of this condition is higher in patients with type 1 diabetes compared to patients with type 2 diabetes. Diabetic retinopathy specifically refers to damage to the small blood vessels within the retina due to elevated glucose levels. DR is a progressive condition that can be classified into various stages, included in the spectrum of diabetic eye complications are background diabetic retinopathy, diabetic maculopathy, preproliferative diabetic retinopathy, and proliferative diabetic retinopathy. This classification is commonly used in clinical observations. Symptoms of diabetic retinopathy include microaneurysms, retinal hemorrhages, exudates, diabetic macular edema, venous beads, and ischemic maculopathy. These manifestations arise as a result of poorly controlled diabetes, hypertension, smoking, and anemia [3].

With the increasing prevalence of diabetic retinopathy, it is expected to be a major contributor to visual impairment and blindness in the future. Therefore, conducting research on vision health is essential for effective diabetes management. Individuals who have been diagnosed with diabetes should prioritize regular eye examinations with an

DOI: https://doi.org/10.54732/jeecs.v8i2.7

ophthalmologist to prevent the onset of blindness [4]. In recent years, machine learning and artificial intelligence have made tremendous progress in tackling complex challenges in various fields. This progress is especially noticeable in the field of scientific research. The field of medicine has undergone a major transformation due to the impact of machine learning and artificial intelligence, and this influence continues to expand [5].

Overall, recent research in Diabetic Retinopathy (DR) detection using SVM and Random Forest methods has resulted in significant progress in identifying early signs of the disease in eye images. The integration of more advanced image processing and data mining techniques along with the use of these classification methods has contributed to improved detection accuracy. Nonetheless, there are still challenges in terms of eye image variations, changes in lighting conditions, and limitations in the available data. Therefore, it is expected that future research will continue to develop more comprehensive approaches to detect DR more accurately and early.

This research paper presents a comparative analysis of two classification methods, specifically SVM and Random Forest (RF), in the context of diabetic retinopathy classification. The article is divided into three sections, with each section emphasizing on a different aspect related to this research. The initial section discusses the methods and materials used in the research, with SVM and RF as the chosen methods. The dataset used consists of fundus images obtained from patients diagnosed with diabetic retinopathy. The next section provides an extensive discussion of the dataset and methodology used during the study. The last section provides a thorough examination of the results obtained and summarizes the findings of the research conducted.

This study aims to compare SVM and Random Forest classification methods in detecting DR in eye images. We will evaluate their performance based on accuracy, precision, recall, and F1-score for better DR detection guidance. The implementation of both methods on eye image data of diabetic patients is expected to improve early detection of DR, reduce the burden on ophthalmologists, and enable the development of a model that can be widely applied in various health facilities. The results of this study are also expected to contribute to the scientific literature, encourage further research, and improve DR care more effectively.

2. RESEARCH METHODOLOGY

2.1 Dataset

The UCI Machine Learning Repository served as the data source for this research. The repository is an integrated online database, theoretical domain, and data generator specifically created to offer datasets for analysis in the field of machine learning [6].The dataset will be divided into two parts: training data and testing data. The training data will be used to train the SVM and RF models, while the testing data will be used to test the performance of the trained models. Machine learning entails a fundamental component of artificial intelligence that empowers computers to interact with new environments and retain knowledge from past experiences [7].

The diaretdb1 dataset consists of a total of 178 images, of which 40 images depict mild proliferative diabetic retinopathy, and the remaining 115 images are from healthy eyes. On the other hand, the messidor dataset contains 105 images, with 55 images representing diabetic retinopathy and 50 images representing non-diabetic retinopathy [8]. Each image in the dataset underwent examination and evaluation by a panel of four experts to identify the presence of abnormalities such as microaneurysms, hemorrhages, and hard exudates, as well as the presence of soft exudates. These images were captured using a fundus camera and have a resolution of 1500x1152 pixels in PNG format [9].

2.2 Method

In detecting DR in eye images. SVM is known for its ability to separate classes with an optimal hyperplane, while RF is an ensemble learning method that utilizes a number of decision trees. Through experiments, it is possible to compare the accuracy, precision, recall, and F1-score of the two methods which one is superior in detecting DR at different severity levels.

The characteristics of each method are justified. SVM is effective in dealing with complex and high-dimensional data. This is important in DR detection because eye image data has many features and variations. On the other hand, RF offers better ability to cope with data variation and noise as it applies ensemble learning techniques. This corresponds to the variability of eye images that may arise from differences in examination conditions and the presence of noise.

The choice of method, considering interpretability is also very important. SVM provides a hyperplane that can be described and allows for a more intuitive understanding of how class separation occurs. This is an advantage in the interpretation of DR detection results to medical professionals. On the other hand, RF consists of many decision trees, which makes interpretation more complex.

Consider the adaptability of both methods in the context of DR detection. SVM, with its various kernels that can be used, can be adapted to different data characteristics. RF can provide a more stable solution by overcoming overfitting, which often occurs in the case of DR detection due to significant variations in eye images. In justification, each method has its own challenges. SVM may tend to overfit if not well optimized, while Random Forest may produce less interpretable models. However, with advanced image processing techniques and feature integration, both SVM and

RF have the potential to overcome these challenges. Based on performance comparison, description of characteristics, explanation of interpretability, adaptability, and context considerations, we chose to apply both methods in DR detection. The integration of advanced image processing techniques also allows us to overcome challenges that may arise. Thus, this study has a solid foundation in selecting the most suitable method for DR detection with optimal accuracy and effectiveness.

2.2.1 Support Vector Machine

Support Vector Machine (SVM) is a type of machine learning algorithm that uses principles of statistical learning and structural risk minimization for pattern recognition and regression. The main goal of SVM is to find the optimal hyperplane that effectively separates positive examples from a combination of negative examples, while maximizing the margin between them.

In the SVM method, we will implement SVM with an appropriate kernel to perform eye image classification. This research uses optimized parameters to avoid overfitting and ensure good performance. As well as the classification context, the SVM searches for a hyperplane that can distinguish between positive and negative examples. This hyperplane is chosen such that the distance between the hyperplane and the closest training examples, called support vectors, is the largest. SVMs use mathematical techniques such as linear programming and quadratic optimization to find the optimal hyperplane. The algorithm utilizes kernel functions, which allow SVMs to project the data into a higher feature space, where a linear separation is possible. Commonly used kernel functions include linear, polynomial, and Gaussian kernels. The advantage of SVM lies in its ability to tackle complex classification problems and non-linear data. It is also resistant to overfitting as it maximizes the margin between classes. In addition, SVM has a high accuracy in performing classification, making it one of the popular methods in pattern recognition. SVM has a disadvantage in handling very large data sets due to its computational complexity. In addition, proper selection of parameters, such as C and gamma parameters, is essential to obtain optimal results.

Overall, SVM is a powerful and effective machine learning algorithm for pattern recognition and classification. By maximizing the margin between classes, SVMs are able to produce accurate decision boundaries and can be applied to various types of classification problems. One of the main advantages of SVM is its ability to build robust classification models even when faced with limited data.

The drawback of SVM becomes impractical when dealing with large datasets [10].

$$y(x,w) = \sum_{j=0}^{M-1} \omega j \emptyset j \qquad (1)$$

In this particular context, the input variables are denoted as x = (x1, x2, ..., xD)T, while the parameters are represented by w = (w0, w1, ..., wD)T. The basis function, ∅(x), is used, and M refers to the overall number of model parameters. Usually, the basis function ∅(x) is equal to 1, which signifies that w0 acts as a bias term.

When applying the SVM method to test the diaretdb1 dataset, it was seen that the accuracy achieved was 88%. The test dataset includes 89 data points diagnosed with non-diabetic retinopathy and 89 data points diagnosed with diabetic retinopathy. In addition, when evaluating the messidor dataset using the same method, the accuracy obtained was 80%. In this case, the test dataset consisted of 50 data points diagnosed with non-diabetic retinopathy and 55 data points diagnosed with diabetic retinopathy.

2.2.1 Random Forest

This method builds a Random Forest (RF) classification with a number of decision trees. Each tree will be randomly generated and used to classify eye images. The majority decision of these trees will determine the prediction class.

RF is a machine learning method that involves combining the predictions of multiple decision trees to produce a final prediction. The term "Random Forest" comes from the creation of many trees through a bootstrapping process. Each tree in RF generates its own class prediction, and the final prediction is determined by selecting the most frequent prediction among the trees [12].

RF is a highly effective machine learning algorithm, especially suitable for classifying large data sets. It stands out for its ability to handle data of various scales and delivers outstanding performance. By combining decision trees using the provided training dataset, the algorithm demonstrates its versatility. Notably, RF does not require complex tuning to achieve good accuracy. Increasing the number of trees used in the algorithm will result in higher accuracy rates. In addition, RF excels in overcoming common challenges faced in data classification, which further improves its overall performance [13].

The RF algorithm has a unique feature that allows it to increase the level of randomness during the tree growth process. Unlike other methods that prioritize the most significant features when splitting a node, RF uses a random subset of features to identify the optimal features. This approach introduces considerable variation, which often results in better model performance and more robust designs [14].

The RF algorithm uses ensemble learning, a technique that combines multiple classifiers to tackle complex problems and improve model performance. RF operates in two stages. In the first stage, N decision trees are combined to build an ensemble of decision trees. In the second stage, predictions are generated for each individual tree in the forest [15].

RF offers many benefits, such as its ability to effectively handle non-linear data, reduce the risk of overfitting, and achieve higher accuracy compared to alternative classification algorithms, but it is also important to recognize certain limitations. One such limitation is the tendency of RF to exhibit bias towards categorical variables. Also, when working with large datasets, the computation time required by this algorithm can be relatively slow. Also, Random Forest is not suitable for linear methods that involve many sparse features.

$$P(c \vee I, x) = \frac{1}{T}\sum_{t=1}^{T} p_{\tau}(c \vee I, x) \quad (2)$$

Each tree in the RF is trained using a set of randomly selected images. The purpose of this random selection is to ensure a balanced distribution across the retina affected by DR. Each individual tree is trained using an algorithm that presents a set of randomly selected candidate separators, consisting of a function parameter θ and a threshold τ.

The RF technique was used to evaluate the diaretdb1 dataset, which resulted in an accuracy of 91%. The test dataset for this evaluation included 89 cases diagnosed as non-diabetic retinopathy and 89 cases diagnosed with diabetic retinopathy. Similarly, when the RF method was applied to test the messidor dataset, it achieved 85% accuracy. In this case, the testing dataset consisted of 50 cases diagnosed as non-diabetic retinopathy and 55 cases diagnosed with diabetic retinopathy.

3. RESULTS AND DISCUSSION

The initial step in this process is to input the raw data into the Input Dataset. Next, the Training Dataset is used as a subset to train the machine model in classifying the severity of absence of DR and presence of proliferative DR. The testing dataset serves as a subset to assess the performance of the model trained on the patient's unseen retinal images using both methods. SVM Model Input consists of the set of datasets used to train the SVM method. The SVM Model Train Dataset, a subset of the dataset, is used to train the SVM model to classify and predict the severity of DR on retinal images. The SVM Evaluation Model is used to measure the accuracy in predicting and classifying the severity of DR in retinal images using the SVM method. Similarly, the RF Model Input consists of a set of datasets used to train the RF model. The Train RF Model Dataset, which is part of the provided datasets, is used to train the RF model in classifying and predicting the severity of diabetic retinopathy in retinal images. RF Model Evaluation assesses the accuracy in predicting and classifying the severity of DR in retinal images using the RF method The flow of the research conducted is located in Figure 1. In this study, two datasets, diaretdb1 and messidor, were used. The datasets are then used to define X and Y variables. X represents a Ratio variable, while Y represents a Nominal or label variable. The label data consists of two different categories: No DR and Proliferative DR. The data was then divided into training data and test data for further analysis and evaluation.

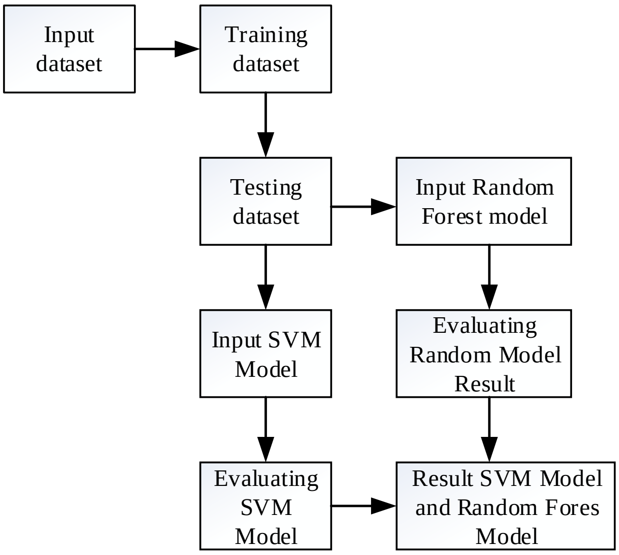
Input dataset
Training dataset
Testing dataset
Input Random Forest model
Input SVM Model
Evaluating Random Model Result
Evaluating SVM Model
Result SVM Model and Random Fores Model

Figure 1. Flowchart

Table 1. Accuracy Result

Method	**Dataset Accuracy Results**	
	Diaretdb1	**Messidor**
SVM	88%	80%
Random Forest	91%	85%

The results of the SVM and RF performance analysis will be analyzed to understand their ability to classify eye images of diabetic patients. We will look at the extent to which each method can identify DR at an early stage or higher severity using SVM and RF methods can be seen in Table 1 This table provides information regarding the accuracy achieved for the diaretdb1 and messidor datasets using each method.

With reference to Table 1, it can be concluded that RF shows superior performance in terms of accuracy compared to SVM. This shows that when applied to the same dataset, RF outperforms SVM as a classification method.

4. CONCLUSION

In the diabetic retinopathy study using the diaretdb1 dataset, the classification results showed that the SVM method achieved an accuracy of 88%, while the RF method had a higher accuracy of 91%. The same thing happened on the messidor dataset, where SVM obtained an accuracy of 80% and RF outperformed with an accuracy of 85%. These results show a significant difference between the two methods. In the context of DR research, this finding highlights that the RF algorithm provides more accurate results than the SVM method, suggesting that RF is able to recognize better patterns and characteristics in DR data compared to SVM.

This study provides deep insights into the performance of SVM and RF in detecting DR. This approach allows us to provide better guidance on the potential of both methods in supporting DR early detection and management efforts with a more complete understanding. In choosing the right algorithm, it is important to consider the research context and the dataset used. In this case, the RF method has been shown to provide better performance in classifying DR on the diaretdb1 and messidor datasets. Therefore, in the context of this research, RF is a better and attractive option to use.

It should be noted that the superiority of RF in this study does not mean that this method is always better than SVM in all situations. Each method has advantages and disadvantages that need to be considered according to the characteristics of the data and the research objectives. These results show that in the case of DR classification, RF has provided more accurate results and can be considered an attractive alternative to use.

Overall, this study shows that both SVM and RF have the potential to be effective tools in DR detection. The choice of method depends on the characteristics of the data, the complexity of the problem, as well as the need for interpretability. The integration of advanced image processing techniques can also improve the performance of both methods. This study provides valuable guidance for medical professionals and researchers to select an appropriate method for early detection and better accuracy in addressing DR.

This study suggests the potential of SVM and RF in supporting early detection and management of DR. The choice of method depends on the characteristics of the data, interpretation, and context of application. Through this research, we aspire to provide more focused guidance in the application of classification techniques to improve DR detection outcomes more effectively.

Further research is recommended to improve and refine existing algorithms or explore alternative methods with the aim of achieving more optimized and superior results. By investigating different algorithms or approaches, there is potential to improve the accuracy and performance of the classification process. Additionally, exploring the feasibility of combining multiple algorithms or using advanced techniques may yield better results. The continued pursuit of research and innovation in this domain can contribute to the advancement of this field and potentially offer more effective solutions.

REFERENCES

[1] K.K.M. Rahman, M. Nasor and A. Imran, (2022*). "Automatic Screening of Diabetic Retinopathy Using Fundus Images and Machine Learning Algorithms"*. Diagnostics, vol. 12, no. 9, pp.1132-1141.

[2] B. Sumathy, S. Gupta, S.S. Hishan, B. Raj, K. Gulati, G. Dhiman, (2022). "*Prediction of Diabetic Retinopathy Using Health Records With Machine Learning Classifiers and Data Science*". International Journal of Reliable and Quality E-Healthcare, vol. 11, no. 2, pp.1-16.

[3] P.M. Karpecki, (2015). Kanski's Clinical Ophthalmology: A Systematic Approach, Brad Bowling, LWW, Lippincott Williams & Wilkins, Philadelphia, Pennsylvania, LWM.

[4] Omer Faruk Gurcan, Omer Faruk Beyca, Onur Dogan. (2021). "*A Comprehensive Study of Machine Learning Methods on Diabetic Retinopathy Classification"*, International Journal of Computational Intelligence Systems, vol. 14, no. 2, pp.1132-1141.

[5] J.A.M. Sidey-Gibbons, and C.J. Sidey-Gibbons, (2019). Machine learning in medicine: A practical introduction. BMC Med Res. Methodol. 64.

[6] A. Roihan, P.A. Sunarya, and A.S. Rafika, (2020). "*Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper"*, IJCIT (Indonesian Journal on Computer and Information Technology), vol. 5, no. 1, pp.75-82.

[7] H. Al Azies, and G. Anuraga, (2021). "*Klasifikasi Daerah Tertinggal di Indonesia Menggunakan Algoritma SVM dan K-NN"*. Jurnal ILMU DASAR, vol. 22, no. 1, pp.31-38

[8] L. Wang, H. Liu, Y. Lu, H. Chen, J. Zhang and J. Pu. (2019). "*A Coarse-to-Fine Deep Learning Framework for Optic Disc Segmentation in Fundus Images*", Biomedical Signal Processing and Control, Vol. 51, pp. 82-89.

[9] A.K. Whardana, and H. Tjandrasa, (2014). "*Segmentasi Microaneurysm Pada Citra Fundus Retina Untuk Deteksi Dini Diabetic Retinopathy*" Scan, vol. 9, no. 3, pp.49-54.

[10] R.H.B. Saputra, and R. Mubarok, (2022). "*Implementasi Algoritma Random Forest Untuk Mendiagnosis Kejadian Berulang (Kekambuhan) Pada Kanker Payudara Berbasis Web"*, OKTAL : Jurnal Ilmu Komputer dan Sains, vol. 1, no. 6, pp.564-572.

[11] F.A. Novianti, and S.W. Purnami, (2012). "*Analisis Diagnosis Pasien Kanker Payudara Menggunakan Regresi Logistik dan Support Vector Machine (SVM) Berdasarkan Hasil Mamografi"*. Jurnal Sains Dan Seni ITS, vol. 1, no. 1, pp.147-152.

[12] P. Saimadhu, (2017). How the Random Forest Algorithm Works in Machine Learning, Published on May 22.

[13] B. Lowe, and A. Kulkarni. (2015). "*Multispectral Image Analysis Using Random Forest"*, International Journal on Soft Computing (IJSC), vol. 6, no. 1, pp.1-14.

[14] R. Casanova, S. Saldana, E.Y. Chew, R.P. Danis, C.M. Greven, W.T. Ambrosius, (2014). "*Application of Random Forests Methods to Diabetic Retinopathy Classification Analyses"*, Plos One, vol. 9, no. 6, doi:10.1371/journal.pone.0098587

[15] M. Gandhi, and R. Dhanasekaran, (2013). Diagnosis of Diabetic Retinopathy Using Morphological Process and SVM Classifier, *in Proceedings of IEEE International conference on Communication and Signal Processing*, pp.873-877, doi: 10.1109/iccsp.2013.6577181.