Classification of Glaucoma Using Invariant Moment Methods on K-Nearest Neighbor and Random Forest Models

Athalla Rizky Arsyan, Wikky Fawwaz Al Maki*

Informatics, School of Computing, Telkom University, Bandung, Indonesia
Email: [1]athallara@student.telkomuniversity.ac.id, [2,*]wikkyfawwaz@telkomuniversity.ac.id
Correspondence Author Email: wikkyfawwaz@telkomuniversity.ac.id

Abstract−One of the cardiovascular diseases that can interfere with eye vision is glaucoma. This disease is caused by high pressure on the inside of the eyeball to cause blindness slowly. In general, screening or early diagnosis can help prevent glaucoma, specifically by analyzing several eye components affected by pressure, including the optical disc, optical cup, and blood vessels. Thus, by blending machine learning algorithms and computer vision technology, glaucoma classification and identification can be accelerated and improved. This study applied the Invariant Moment method to extract the optical cup and blood vessel segmentation's shape, scale, and rotation features. To obtain segmentation results for these two objects, we threshold two image datasets, DrishtiGS-1 and REFUGE, and implemented the approach described in this study to analyze system performance on these datasets. For the classification method used in this study, we proposed KNN and RF models to evaluate the suitability of the methods we used on the REFUGE and DrishtiGS-1 datasets and demonstrated that both models could be used to identify glaucoma through the use of fundus images. When the datasets were merged, we obtained 81.86% and 75.86% of accuracy when using blood vessel and optical cup segmentation results, respectively.

Keywords: Glaucoma; Invariant Moment; KNN; Random Forest; Machine Learning

1. INTRODUCTION

Glaucoma, which is typically characterized by optic nerve damage, can be related to increased pressure in the eyeball. Patients with glaucoma can experience headaches, eye pain, redness, nausea, and vomiting. Glaucoma sufferers may also experience visual abnormalities that develop to progressive blindness if the disease is not treated appropriately. Visual disturbances in patients with this disease cannot be cured in any way, even with surgery, because the damage occurs in the optic nerve. The thing that can damage the optic nerve in this disease is the pressure that prevent the entry of the nutrient fluid needed in the eye components to cause nerve cell death in the eye.

Until now, the only option to prevent the occurrence of this disease has been through routine screening or early identification of the retina of the eye, this is essential because, in the early stages of glaucoma, patients commonly show no symptoms. One of the most often used methods for detecting retinal diseases is to evaluate the fundus image (fundus copy), which would be obtained using the ophthalmoscope method to examine the back and interior of the eye. This procedure is time-consuming because the doctor should inspect the fundus of the eye through the ophthalmoscope's small lens to diagnose abnormalities in the eye's retina. As a solution, numerous studies have attempted to automate the diagnostic process by combining computer vision and machine learning methods. Typically, evaluation of the optical disc, optical cup, and blood vessel anatomy can be employed to determine cardiovascular disease in the eye's retina [1][2].

The easiest thing to analyze in detecting glaucoma is to measure the diameter ratio between the optical disc and the optical cup (cup-to-disc ratio). This ratio measurement method is widely used because glaucoma patients' pressure has a significant effect on the shape of the two objects. The greater the pressure, the more significant the change in the object's diameter, particularly in the optical cup diameter ratio. The research [3], attempted to build a system for early detection of glaucoma by calculating the ratio of the optical disc and optical cup obtained through experiments on the combination of red and green channels on several labeled image samples and succeeded in demonstrating that both channels were successful—proven in obtaining the segmentation of the optical disc and optical cup in order to calculate the ratio to detect glaucoma.

Karmawat R et al. (2019) proposes a glaucoma diagnosis system based on the fuzzy c-means (FCM) algorithm for segmenting the optical cup into normal and glaucoma categories. Then the results of the segmentation will be extracted using the CDR, ISNT rule, and DDLS methods. HOG and 2D-DWT. This study obtained system performance accuracy of 80.20% and 82.20% for the Support Vector Machine (SVM) model and ensemble classifier, respectively [4].

Salam et al. (2017) proposed a glaucoma detection system which by calculating the ratio of the optical disc to the optical cup. The procedure of optical disc segmentation starts with the selection of the HSV plane on the image. It will be segmented using Otsu Thresholding, and then the image's edges will be retrieved and smoothed using the ellipse fitting method. Meanwhile, the contrast in the green channel is enhanced to obtain optical cup segmentation, and blood vessels are removed first using opening and closing morphological operations, followed by the canny edge detector method to acquire the optical cup's surface. This study succeeded in obtaining an average accumulation of accuracy, specificity, and sensitivity of 88.5%, 85.25%, and 93.75, respectively [5].

Dutta et al. (2018) proposed a system that created to automates the optical cup area segmentation process in RGB-channel images in order to predict glaucoma. This system's classification approach is pixel-based feature

classification, which analyses the feature set of each pixel with a numerical value in the image. While the optical cup segmentation process is carried out by calculating the average threshold value in the image, the accuracy attained in this system is 83.168% whenever 84 photos from 101 images in the dataset were collected from Arvind Eye Hospital in Madurai with IIIT Hyderabad are classified [6].

Joshi et al. (2018) proposed a system by focusing on extracting features from the optical disc segmentation obtained using the level set method and the Local Binary Pattern (LBP) method, which leverages the average, median, standard deviation, entropy, and skewness characteristics. Besides that, the kurtosis will be examined using a Support Vector Machine model. The final result of the performance evaluation of the system obtained accuracy, sensitivity, specificity, and F1-score values of 78.57%, 83.33%, 75%, and 76.92 %, respectively [7].

According to previous research, the optical disc and optical cup play an essential role in diagnosing glaucoma through fundus images. However, only a few studies use blood vessel segmentation to detect glaucoma. Blood vessels in the retina also have an important role in detecting the disease because the pressure on the eyeball can also affect the shape, structure, and thickness of the blood vessels. Even so, this study employs Invariant Moment to extract form, dimension, and rotational features from the optical cup and blood vessel segmentation of the eye's retina. Then, the method's feature extraction vectors will be examined using multiple machine learning techniques to decide which classifier model produces the most significant system performance on the used dataset. The primary reason we use the Random Forest model is that it is well-suited for classifying data with a significant level of randomness, as our feature extraction method generates multiple features with random values. While K-Nearest Neighbor was chosen as another comparison model because the results of image segmentation will have adjacent pixel distances, this method is suitable for calculating the image matrix value in the surrounding neighbors.

2. RESEARCH METHODOLOGY

2.1 System Design

We have designed a system that will classify glaucoma into two categories, normal retina and glaucoma, leveraging feature extraction and the Invariant Moment method. The optical cup and blood vessel segmentation results are the primary targets for machine learning algorithms. Prior to initiating the classification process, image pre-processing is needed to acquire the segmentation results for the two objects. After the invariant moment method is used to extract features, the feature matrix array will be classified using the Random Forest and K-Nearest Neighbor models with two categories, normal and glaucomatous retina, to analyze the performance of the system. Each model classifier will be validated using 5-Fold Cross-Validation. **Figure 1.** provides an overview of the process in general.
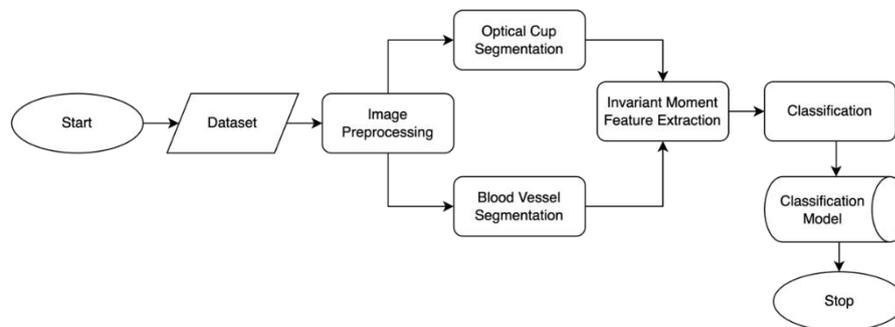
Figure 1. System Design

This research implemented two available datasets, DrishtiGS-1 [8][9], which contains 31 normal retinal images and 70 glaucoma retinal images, and the REFUGE [10][11], which has 360 normal retinal images and 40 glaucoma retinal images. Additionally, images from the Drishti-GS1 dataset were collected through a 30° field of view with a resolution of 2896x1944 pixels in patients aged 40–80 years at India's Aravind Eye Hospital [12]. And the REFUGE dataset image was captured with a Zeiss Viscuam 500 camera at a resolution of 2124x2056 pixels and 96 dpi [12]. **Table 1.** contains the total number of images in each dataset.

Table 1. Datasets

Dataset	Normal	Glaucoma	Total
DrishtiGS-1	31	70	101
REFUGE	360	40	400

2.2 Optical Cup Segmentation

Physically, glaucoma can be diagnosed through the shape of the optical cup because generally, the diameter of the optical cup will enlarge when there is the pressure caused by fluid on the inside of the eyeball. Therefore, we use the optical cup as one of the objects for glaucoma detection. To segment the optical cup, we apply the adaptive

thresholding method, which involves measuring the mean and standard deviation of pixels with a minimum and maximum threshold value in order to distinguish the object from the background in the image. In general, the following formula could be used to calculate the threshold value [13].

$$f(y,x) = \begin{cases} 1, if\ g(x,y) > Thr \\ 0, if\ g(x,y) \leq Thr \end{cases}, \tag{1}$$

where *f(y,x)* returns a value of 1 for pixels with a threshold value greater than a predefined limit and a value of 0 otherwise. The variable *Thr* represents the threshold value used to segment the optical cup, *g(x,y)* is green channel matrix from the image, and *f(y,x)* is the result of segmented image using thresholding method. We define the threshold value for the optical cup in the image using the green channel. For this reason, the formula for calculating the threshold value for the optical cup is defined using the following formula [14].

$$T = (0,5 * c) + (2 * \sigma CStdDev) + (2 * \sigma GreenStdDev) + (\mu GreenMean), \tag{2}$$

where T, c, *σGreenStdDev, σCStdDev,* and *μGreenMean* are the segmentation result of the optical cup, width of the gaussian window, standard deviation of normalized green channel, standard deviation of the gaussian window, and the mean value of normalized green channel, respectively. **Figure 2.** Shows the processed green channel of the image and segmented optical cup.
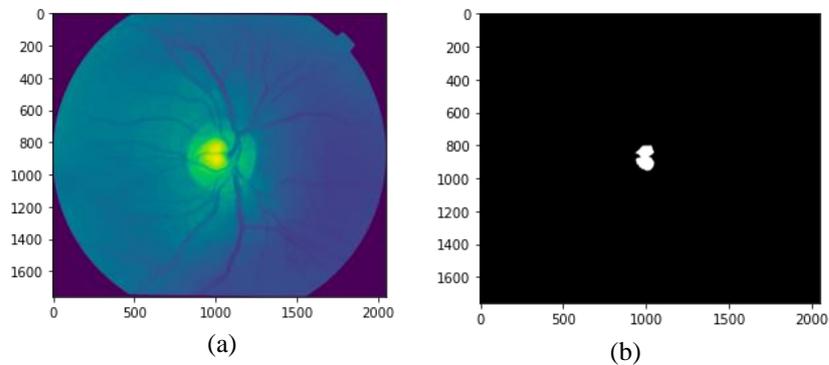
**Figure 2.** Green Channel from Images (a), Optical Cup Segmentation (b)

2.3 Blood Vessel Segmentation

The dominant characteristic for blood vessel segmentation is the green channel in fundus images with RGB channels. As a result, we pre-processed the green channel using the Contrast Limited AHE (CLAHE) approach to produce the segmentation results. This step is essential because it efficiently differentiates blood vessels from other eye components by turning the image to grayscale. This technique will adaptively minimize the slopes of the transformation function [15]. To repair the image's interior and exterior, we apply Top-Hat and Bottom-Hat morphological techniques to fill in the image's hollow pixels. The image will then be segmented using the thresholding approach to obtain the portion of the blood vessels visible in the image. **Figure 3.** Shows processed of contrast enhancement using CLAHE, morphological operation using Top-Hat and Bottom-Hat, and segmented blood vessels.
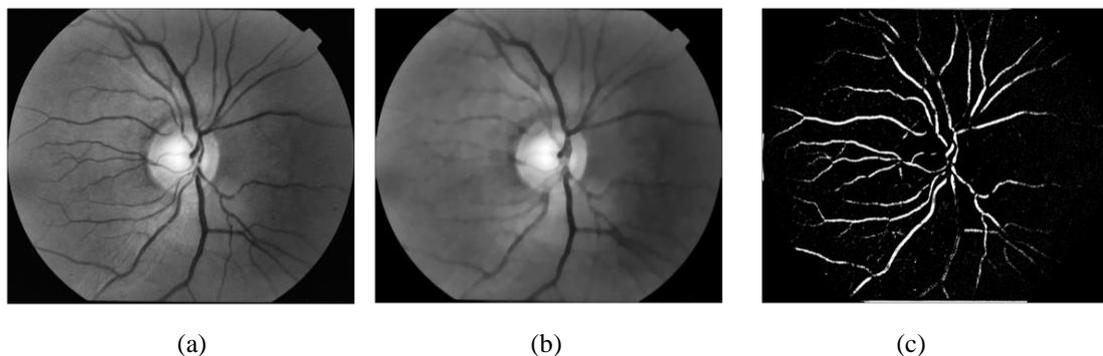
Figure 3. CLAHE Enhanced (a), Morphological Operation (b), Blood Vessel Segmentation (c)

Then, in order to eliminate noise from the image, we perform morphological erosion techniques to remove small portions of the image that are not required. Furthermore, this noise removal methodology uses object detection techniques to detect and identify small contours inside the image, as well as thresholding and masking to select these small contours. **Figure 4.** Shows small contour from image and removal of that contour.

(a)   (b)
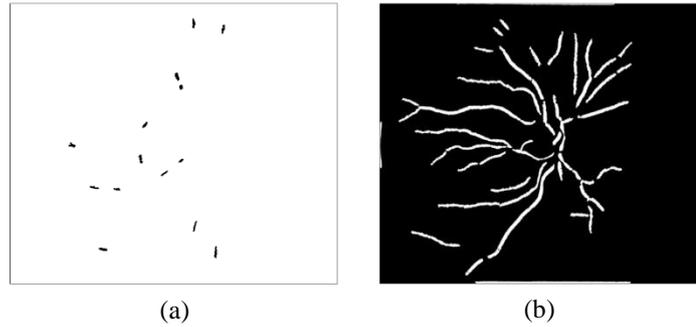
Figure 4. Small Contour Detection (a), Noise Removal (b)

2.3 Invariant Moment Feature Extraction

Invariant Moment or also known as Hu Moment is a method that can be used to extract features from shape, rotation, scale, orientation using seven hu moment centers. In this study, we performed feature extraction of the optical cup and blood vessels on the DrishtiGS-1 and REFUGE datasets to analyze whether these datasets are suitable for diagnosing glaucoma using feature extraction from Invariant Moment. This approach takes as input the results of the previously performed optical cup and blood vessel segmentation. It extracts seven vector matrices from the calculation results for the center of the moment hu and its derivatives. The following formula can be used to define the formula for the seven center of Hu Moment [16].

$$\mathfrak{f}_1 = \mu_{20} + \mu_{02} \tag{3}$$

$$\mathfrak{f}_2 = (\mu_{20} + \mu_{02})^2 + 4\mu_{11}^2 \tag{4}$$

$$\mathfrak{f}_3 = (\mu_{30} - 3\mu_{12})^2 + (3\mu_{21} - \mu_{03})^2 \tag{5}$$

$$\mathfrak{f}_4 = (\mu_{30} + \mu_{12})^2 + (\mu_{21} + \mu_{03})^2 \tag{6}$$

$$\mathfrak{f}_5 = (\mu_{30} - 3\mu_{12})(\mu_{30} + \mu_{12})\{(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2\} + (3\mu_{21} - \mu_{03})(\mu_{21} + \mu_{03}) + \{3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2\} \tag{7}$$

$$\mathfrak{f}_6 = (\mu_{20} - 3\eta_{02})\{(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2\} + 4\mu_{11}(\mu_{30} + \mu_{12})(\mu_{21} + \mu_{03}) \tag{8}$$

$$\mathfrak{f}_7 = (3\mu_{21} - \mu_{30})(\mu_{30} - \mu_{12})\{(\mu + \mu_{12})^2 - 3(\mu_{12} + \mu_{30})^2\} + (3\mu_{21} - \eta_{03})(3\mu_{21} + \mu_{03})\{3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2\} \tag{9}$$

2.4 Classification

The classification method used in this study is machine learning, with the K-Nearest Neighbor with K=5 and Random Forest models. We will perform 5-fold cross-validation on each model during the classification process, and the dataset will be divided into two segments, with 80% training data and 20% testing data. Overall, the parameter that we used default parameters configuration from scikit-learn in this study [17]. The KNN model was adopted for this research because it is straightforward and simple to implement but has large computing capabilities [18]. Meanwhile, due to the dataset's imbalanced label quantity, we attempted to use the RF model, which has been shown to be extremely good at categorizing unbalanced data [19].

2.4 Performance Analysis

This study will evaluate the performance results of the created system model by considering four metrics: precision, recall, F1-score, and accuracy. Precision is used to determine the proportion of positive expected data that occurred. In this context, the term "positive" refers to glaucoma. A recall is used to determine the number of verifiable positive data that are properly identified. The F1-score is computed by combining precision and recall. Accuracy is defined as the percentage of correct predictions. The following expression would be used to evaluate the performance metrics:

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)} \tag{10}$$

$$Recall = \frac{True\ Positive}{(True\ Positive + False\ Negative)} \tag{11}$$

$$F1 - Score = \frac{(Precision * Recall)}{(Precision + Recall)} \tag{12}$$

$$Accuracy = \frac{(True\ Positive + False\ Positive)}{(True\ Positive + True\ Negative + False\ Positive + False\ Negative)} \tag{13}$$

3. RESULT AND DISCUSSION

This research used feature extraction from the Invariant Moment method to diagnose glaucoma in two distinct datasets. The optical cup and blood vessels were the only parts of the object evaluated. The final result is that the K-Nearest Neighbor (KNN) model performs better in REFUGE dataset than the Random Forest (RF). Whenever RF model also performs better in DrishtiGs-1 dataset. The validity of all classifier models used in this investigation was determined using K-Fold Cross Validation with K = 5.

The results of the analysis on the REFUGE dataset using optical cup segmentation are almost comparable: accuracy of 90% in the KNN model. Furthermore, the RF model achieves 89,50% accuracy. The conclusion that can be made is that by studying the optical cup in the fundus image, this dataset may be combined with the Invariant Moment method. In **Table 2** contains detailed information about the performance analysis's results.

Table 2. REFUGE Dataset and Optical Cup Segmentation System Performance

Classifier	Precision	Recall	F1-Score	Accuracy
RF	87,82%	92,50%	91,10%	89,50%
KNN	87,82%	92,50%	90,10%	**90,00%**

When the REFUGE dataset was examined using the KNN model and blood vessel segmentation, the highest accuracy was obtained with identical values when performance was measured using the blood vessel, particularly regarding 89,50% of accuracy. Followed by the RF model's accuracy of 89.25%. In **Table 3.** Contains detailed information about the performance analysis's results.

Table 3. REFUGE Dataset and Blood Vessel Segmentation System Performance

Classifier	Precision	Recall	F1-Score	Accuracy
RF	87,50%	87,50%	87,50%	89,25%
KNN	87,82%	92,50%	90.10%	**89,50%**

We conclude from the results in **Tables 2** and **Table 3** that both models are capable of high accuracy performance on the REFUGE dataset when optical cup and blood vessel segmentation are being used as inputs. Both models may perform well due to the enormous number of images in this dataset. Other performance parameters such as precision, recall, and F1-score are also relatively high, suggesting that these two models are quite effective for classifying glaucoma.

The best accuracy in RF models for the DrishtiGS-1 dataset using optical cup segmentation were only 58,33%. Following that KNN achieved accuracy levels of 56,48%. As a result, the conclusion that can be reached is that this dataset is not well suitable for usage with the Invariant Moment method, which requires optical cup segmentation as input. In **Table 4.** contains detailed information about the performance analysis's results.

Table 4. DrishtiGS-1 and Optical Cup Segmentation System Performance

Classifier	Precision	Recall	F1-Score	Accuracy
RF	37,14%	42,86%	38,78%	**58,33%**
KNN	77,14%	61,90%	51,43%	56,48%

Concerning the input in the form of blood vessels in the DrishtiGS-1 dataset, we obtained the highest accuracy of 67,48% using the RF model, followed by KNN model with accuracy of 64.33%. In **Table 5.** contains detailed information about the performance analysis's results.

Table 5. DrishtiGS-1 and Blood Vessel Segmentation System Performance

Classifier	Precision	Recall	F1-Score	Accuracy
RF	51,02%	52,38%	51,24%	**67,48%**
KNN	61,43%	61,90%	59,18%	64,33%

The results in **Table 4** and **Table 5** indicate that the two models have a low level of accuracy, which may be due to the small amount of data in the DrishtiGS-1 dataset and the comparison of unbalanced label numbers, which results in inaccurate data used for training the KNN and RF models. As an additional experiment, we tried to combine the two datasets so that the total label of normal retina was 390 images and glaucoma retinal was 110 images. Then the image dataset that we have combined is also analyzed using optical cup and blood vessel segmentation. While the results of the analysis of optical cup segmentation and the Invariant Moment method produce better results on the RF model, with accuracy of 75.86%. Followed by the KNN models which has an accuracy of 75.65%. In **Table 6** contains detailed information about the performance analysis's results.

Table 6. Combined Dataset and Optical Cup Segmentation System Performance

Classifier	Precision	Recall	F1-Score	Accuracy
RF	71,72%	76,24%	71,42%	**75,86%**

	KNN	80,79%	80,20%	75,29%	75,65%

The blood vessel segmentation and the Invariant Moment technique achieve the maximum performance, with 81.44%, 75.51%, 78.22%, and 73.80% for accuracy, precision, recall, and F1-score, respectively, for RF model. Following that, KNN model has an accuracy of 80.43%. In **Table 7** contains detailed information about the performance analysis's results.

Table 7. Combined Dataset and Blood Vessel Segmentation System Performance

Classifier	Precision	Recall	F1-Score	Accuracy
RF	75,51%	78,22%	73,80%	**81,44%**
KNN	78,37%	80,20%	77,61%	80,43%

When the datasets in **Table 6** and **Table 7** are combined, the accuracy improves significantly when compared to employing the DrishtiGS-1 dataset individually. This means that when data is merged, the data used for training becomes exponentially increasing, which provides for optimization of the data's training process. Despite the increase in accuracy, it still can't beat the accuracy when using only the REFUGE dataset. In addition, there is a possibility of a decrease in accuracy caused by an imbalance of labels in the dataset used. It is recommended to try adding an oversampling method; one example is the Synthetic Minority Oversampling Technique (SMOTE) algorithm which will synthesize minority data so that the number of labels on the data becomes balanced [20].

4. CONCLUSION

Based on the results of experiments that have been carried out, it was concluded that the KNN and RF model with the Invariant Moment method could provide better system performance in classifying glaucoma. Although previous research has emphasized the importance of calculating the ratio of the optical disc to the optical cup (cup-to-disc ratio), this research has demonstrated that system performance can be achieved exclusively by using blood vessels or an optical cup. Data from the overall experimental results prove that the KNN and RF model can outperform in detecting glaucoma in the REFUGE and DrishtiGS-1 datasets. The main conclusion is that the two models we use are remarkably optimal for the REFUGE dataset, as evidenced by the fact that both models achieved an accuracy of greater than 89 % when optical cup and blood vessel segmentation were used as input for the RF and KNN models, respectively. Both models, however, are inappropriate to be used with the DrishtiGS-1 dataset; we speculate that the decreased inaccuracy is due to a lack of input data and an imbalance of data labels in the dataset, likely to result in an inefficient training process on the data. When data sets are combined, the RF model achieves higher accuracy than the KNN model, attributed to the fact that the data used becomes more random; thus, the RF model is more suitable for classification cases that involve random data. While the accuracy of the KNN model is not as good as that of the RF model when the datasets are combined, the accuracy results obtained are also not far off. Therefore, in future studies, when using optical cup segmentation, it is recommended to add other feature extraction methods, try to use other segmentation object, or implement other segmentation algorithms to get a more accurate shape of the object. And also, the use of deep learning algorithms can be taken into consideration or tweaking the machine learning and feature extraction method input parameters to get even better performance results.

REFERENCES

[1]   M. M. Sathik, A. Padma, and M. Sivajothi, "Glaucoma Detection Using Little Wood Paley Decomposition On Local Derivative Structure Of Fundus Images," *European Journal of Molecular & Clinical Medicine*, vol. 7, no. 10, pp. 2988–2998, 2021.

[2]   N. Singh and V. Rao J, "Retinal Vessel Analysis for Detection of Glaucoma," pp. 2321–9653, Jan. 2021.

[3]   S. C. Patel and M. I. Patel, "Analysis of CDR of fundus images for glaucoma detection," in *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*, 2018, pp. 1071–1074.

[4]   R. Karmawat, N. Gour, and P. Khanna, "Glaucoma Detection using Fuzzy C-means Optic Cup Segmentation and Feature Classification," in *2019 IEEE Conference on Information and Communication Technology*, 2019, pp. 1–5.

[5]   A. A. Salam *et al.*, "Benchmark data set for glaucoma detection with annotated cup to disc ratio," in *2017 International Conference on Signals and Systems (ICSigSys)*, 2017, pp. 227–233.

[6]   K. Dutta, R. Mukherjee, S. Kundu, T. Biswas, and A. Sen, "Automatic evaluation and predictive analysis of optic nerve head for the detection of glaucoma," in *2018 2nd International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech)*, 2018, pp. 1–7.

[7]   A. Joshi *et al.*, "Glaucoma Screening Through Level Set for Optic Disc Segmentation and Textural Features for Classification," in *2018 International Conference on Intelligent and Advanced System (ICIAS)*, 2018, pp. 1–6.

[8]   J. Sivaswamy, S. Krishnadas, A. Chakravarty, G. Joshi, A. S. Tabish, and others, "A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis," *JSM Biomedical Imaging Data Papers*, vol. 2, no. 1, p. 1004, 2015.

[9]   J. Sivaswamy, S. R. Krishnadas, G. D. Joshi, M. Jain, and A. U. S. Tabish, "Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation," in *2014 IEEE 11th international symposium on biomedical imaging (ISBI)*, 2014, pp. 53–56.

[10] J. I. Orlando *et al.*, "REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs," *Medical Image Analysis*, vol. 59, p. 101570, 2020, doi: https://doi.org/10.1016/j.media.2019.101570.

[11] F. Li *et al.*, "Development and clinical deployment of a smartphone-based visual field deep learning system for glaucoma detection," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–8, 2020.

[12] A. Y. D. Pinto, "Machine learning for glaucoma assessment using fundus images," 2019.

[13] R. V. Nahari, A. Jauhari, R. Hidayat, and R. Alfita, "Image Segmentation of Cows using Thresholding and K-Means Method," *International Journal of Advanced Engineering, Management and Science*, vol. 3, no. 9, p. 239912, 2017.

[14] K. Manju, R. S. Sabeenian, and A. Surendar, "A review on optic disc and cup segmentation," *Biomedical and Pharmacology Journal*, vol. 10, no. 1, pp. 373–379, 2017.

[15] Y. Liu *et al.*, "Segmentation and Automatic Identification of Vasculature in Coronary Angiograms," *Computational and Mathematical Methods in Medicine*, vol. 2021, 2021.

[16] Z. W. Htet, V. D. Koldaev, Y. O. Teplova, E. A. Kremer, and P. A. Fedorov, "The evaluation of computational complexity of moment invariants in image processing," in *2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, 2018, pp. 1844–1848.

[17] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[18] Y. Tan, "An improved KNN text classification algorithm based on K-medoids and rough set," in *2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, 2018, vol. 1, pp. 109–113.

[19] H. Luo, X. Pan, Q. Wang, S. Ye, and Y. Qian, "Logistic regression and random forest for effective imbalanced classification," in *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, 2019, vol. 1, pp. 916–917.

[20] T. E. Tallo and A. Musdholifah, "The implementation of genetic algorithm in smote (synthetic minority oversampling technique) for handling imbalanced dataset problem," in *2018 4th international conference on science and technology (ICST)*, 2018, pp. 1–4.