Analysis of Eye Disease Classification by Comparison of the Random Forest Method and K-Nearest Neighbor Method

Dwiny Meidelfi [a,1,*], Hendrick[b], Fanni Sukma[a], Srintika Yuni Kharisma[a]

[a] Department of InformationTechnology, Politeknik Negeri Padang, West Sumatera, Indonesia
[b] Department of Electrical Engineering, Politeknik Negeri Padang, West Sumatera, Indonesia

[1]dwinymeidelfi@pnp.ac.id
* corresponding author

ARTICLE INFO

Article history
Received May 10, 2023
Revised June 15, 2023
Accepted July 30, 2023

Keywords
Classification
Eye Disease
Random Forest
K-Nearest Neighbor
Python

ABSTRACT

Eye disease is a serious issue all over the world, and image-based classification systems play an important role in the early detection and management of eye disease. This research compares the performance between Random Forest (RF) and K-Nearest Neighbor (KNN) classification models in identifying eye disorders using image datasets divided into four classes: "normal," "glaucoma," "cataract," and "diabetic retinopathy." The dataset is converted into a feature vector and then divided into training data and test data subsets. The analysis results show that the RF model achieved an accuracy level of 80%, whereas the KNN model achieved 70%. Based on these findings, it is possible to conclude that the RF model outperforms the other models in categorizing the types of eye illnesses in the dataset. A Python-based website was also built utilizing the Flask framework to build an interactive and real-time eye illness diagnosis system. Users can upload photos of their retinas to this website and quickly receive eye disease detection results. The adoption of this technology has a tremendous impact, making eye disease detection solutions more accessible. Furthermore, this solution plays an important role in the early detection and effective management of eye health cases.

1. Introduction

The eye as one of the main human senses plays a crucial role in carrying out daily activities and maintaining the quality of life is a very important basic concept [1]. The eye's primary function is as a light detection mechanism, allowing people to discern between light and darkness and forming visual perception, which allows us to view our surroundings. The retina is one of the most important components of the eye, as it converts light impulses into nerve signals, which are then interpreted by the brain as the images we see [2]. This process is at the heart of human visual abilities.

However, when there is an eye disease or illness, it might interfere with vision function and make it difficult to carry out normal tasks. Eye health issues, such as cataracts, glaucoma, and diabetic retinopathy, can range from mild to severe. Cataracts, for example, occur when the lens of the eye gets clouded, interfering with light entry and resulting in blurred vision. This can be caused by a variety of factors such as oxidation, ultraviolet radiation exposure, heredity, and nutrition [3].

Glaucoma is a series of progressive optic nerve disorders that, if left untreated, can cause eye damage and perhaps eyesight loss. The main risk factor is high ocular pressure, however, other

factors can play a role in the development of this condition [4]. Diabetic retinopathy is a diabetes complication that can result in permanent visual loss if not treated properly. A detailed eye examination, such as with an ophthalmoscope, is usually required to discover this condition [5].

The Ministry of Health's 2014-2016 Rapid Assessment of Avoidable Blindness (RAAB) Blindness Survey indicated substantial levels of blindness among people over the age of 50, with cataracts being the leading cause. This highlights the need for eye disease prevention and early detection to reduce blindness rates.

In the context of the complexity of this eye health problem, machine learning technologies such as Random Forest and K-Nearest Neighbor are emerging as effective solutions. Random Forest has the potential to develop excellent prediction models in identifying various types of ocular images due to its capacity to merge several decision trees. K-Nearest Neighbor, on the other hand, uses a technique based on data closeness to classify ocular pictures, allowing for early detection by comparing test data with existing training data. In this study, the performance of two methods for detecting eye diseases through eye image processing will be compared.

This project will also entail the development of a website that will allow users to upload retinal images and receive predictions regarding eye diseases using machine learning technologies. Thus, it is intended that this study would contribute to improving the quality of people's eye health, lowering the risk of preventable blindness, and identifying the most appropriate methodologies for classifying eye diseases. It is envisaged that by comparing the accuracy results of these two methodologies, the most efficient approach to overcoming complicated challenges linked to eye health would be determined.

2. Literature Review

Classification

Classification is a process or method of grouping objects or data into different types or classes based on the characteristics or characteristics of each object or data [6]. Classification is used to manage patterns and correlations between types or classes, as well as to develop mathematical or statistical models that may predict the type or class corresponding to a particular object or data.

Eye Diseases

The eye is one of the five senses that has a very vital role in human life, namely as an organ of vision [7]. When an eye disorder or disease occurs, the impact is very upsetting, and if not handled seriously, it can have major consequences on a person's quality of life. As a result, keeping eye health is critical in your everyday routine. Cataracts, glaucoma, and diabetic retinopathy are examples of eye diseases.

Python

Python is a simple but very flexible programming language [8], which is explained in its documents. Python is a dynamic programming language that is widely utilized in the development of applications in a variety of fields. This feature enables the creation of programs in multiple approaches at the same time. Graphical interfaces, for example, can be constructed using an object-oriented method, whilst data processing can be done using a functional or procedural approach.

Flask

Flask is a web framework created using the Python programming language and is included in the microframework category [9]. The main function of Flask is as a basic structure for developing web applications. With Flask, developers can create well-structured websites and manage the functionality of those websites more easily [10].

Accuracy

Classification accuracy is the ability of a model or algorithm to make correct predictions on the given data, measured in percentage form. The higher the accuracy, the better the model is at correctly classifying data [11].

Performance is a form of action, deed, or work, which has been achieved or carried out. Classification performance can be evaluated by calculating the performance values of accuracy, precision, recall, and F1Score [12].

Accuracy = ((TP+TN))/((Total sample)) (1)

Precision = (TP/(TP+FP)) (2)

Recall = (TP/(TP+FN)) (3)

F1 Score = 2 . (precise .recall)/(presisi+recall)) (4)

Description:

TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

3. Method

3.1. Random Forest

Random Forest (RF) is a machine learning method that builds classifier classes using supervised concepts. This algorithm integrates predictions from different decision trees [13]. The Random Forest approach can increase accuracy outcomes since it generates child nodes for each node at random [14]. This method is used to generate decision trees with root nodes, internal nodes, and leaf nodes by selecting attributes and data at random. random according to the existing provisions. The decision tree calculates entropy as a measure of attribute impurity first and then measures the resulting information [14].

To calculate the entropy value, use the formula in the equation below [15].

$$Entropy\ (Y) = -\sum p(c|Y) \log 2\ p(c|Y) \quad (1)$$

Description:

Y = the case set

p(c|Y) = the proportion of Y value towards class C.

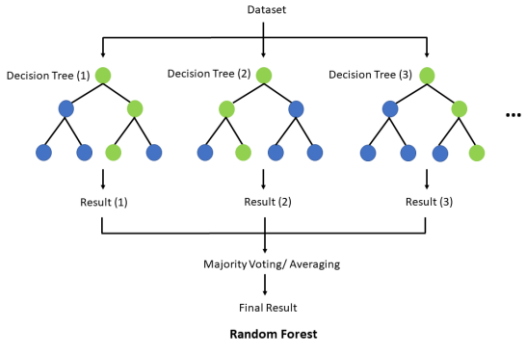

Figure 1. Random Forest (Tibco)

3.2.K-Nearest Neighbor

K-Nearest Neighbors (K-NN), a simple and easy-to-implement algorithm in machine learning that is beneficial for tackling classification and regression problems [17]. K-NN searches for groupings of k objects from the training data that are most comparable to the objects in the test data [18]. This is akin to solving a new patient's problem based on the patient's prior experience. The Euclidean algorithm is used to compute the distance between neighbors, with the Euclidean distance formula calculated as the square root of the total of the attribute differences between two data records. K-NN is a classification method that classifies new test data using previously categorized training data.

Euclidean distance calculation [19].

$$euc = \sqrt{\sum_{i=1}^{n}(pi - qi)^2} \quad (1)$$

Description:

pi = data sample/data training

pi = test data / data testing

i = data variable

n = data dimension

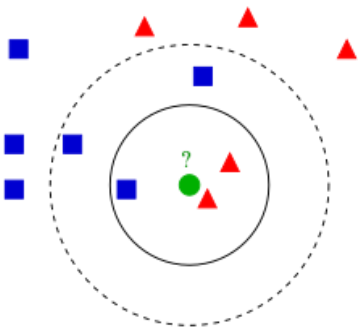

Figure 2. K-Nearest Neighbor (Wikipedia)

3.3. Dataset

The dataset was collected through the official Kaggle website, which is a collection of images exhibiting various forms of eye disorders, such as Cataracts, Diabetic Retinopathy, Glaucoma, and Normal. This dataset contains 1038 images of Cataract type eye disorders, 1098 images of Diabetic Retinopathy, 1007 images of Glaucoma, and 1074 normal images. This research has the ability to

provide in-depth insight into numerous types of eye disorders and enable the creation of a better understanding of the field of ophthalmology by employing this diversified dataset. In this study, 80% of the datasets were used for training and 20% for testing (844 images).

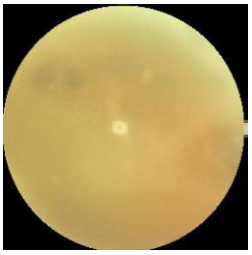

Figure 3. Cataract

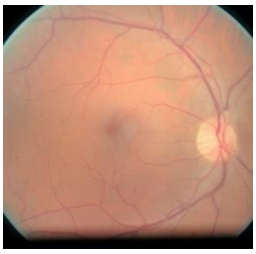

Figure 4. Diabetic Retinopathy

Figure 5. Glaucoma

Figure 6. Normal

3.4. Systems Analysis

Based on input eye images, this new system can diagnose and predict many types of eye disorders using two classification algorithms, Random Forest (RF) and K-Nearest Neighbor (KNN). The RF model is set up to make judgments using 100 trees, with training data accounting for 80% of the dataset and testing data accounting for the remaining 20%. The KNN model, on the other hand, uses three nearest neighbors in the classification process and uses an 80% training data distribution, with the remaining 20% used as testing data.

4. Results

4.1. Login

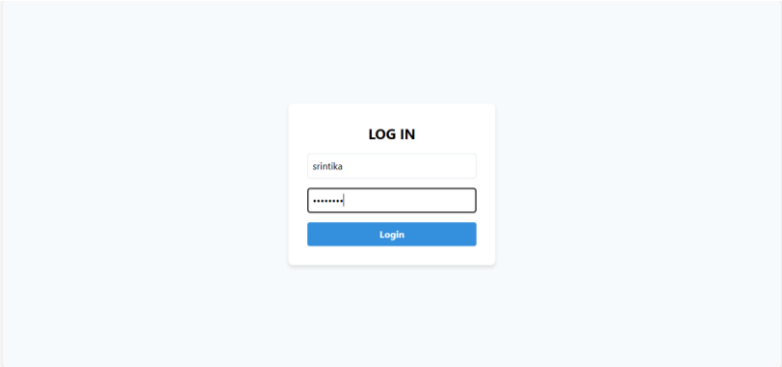

Figure 7. Login

Users will be prompted to enter a username and password. If this combination is verified as correct by the system data, the user will be sent to the home page. Users are provided access to the linked content or features by going through this login process, making the login step a gateway that assures only authorized users can explore the main page and everything it has to offer.

4.2. Main page

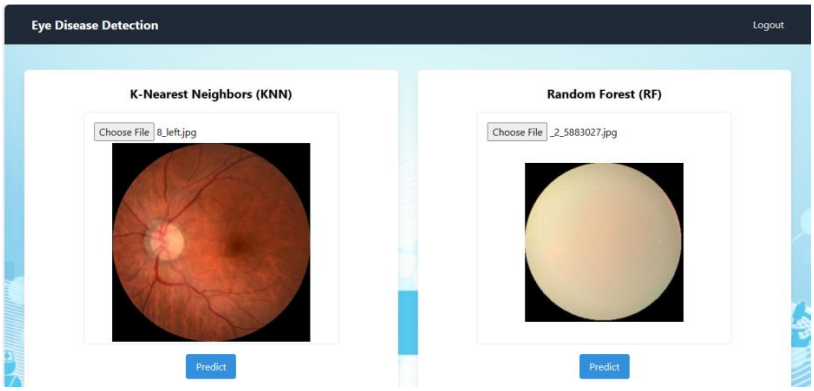

Figure 8. Upload Image

On this page, the user is expressly prompted to insert an image of the retina of the eye to be evaluated for prediction purposes. The system will process the image after the user successfully uploads it. If the image is correctly predicted, the prediction results in the image file name and the image itself being displayed to the user. This gives providers visual information about the prediction findings, as well as precise details about the photographs they provided and the associated prediction.

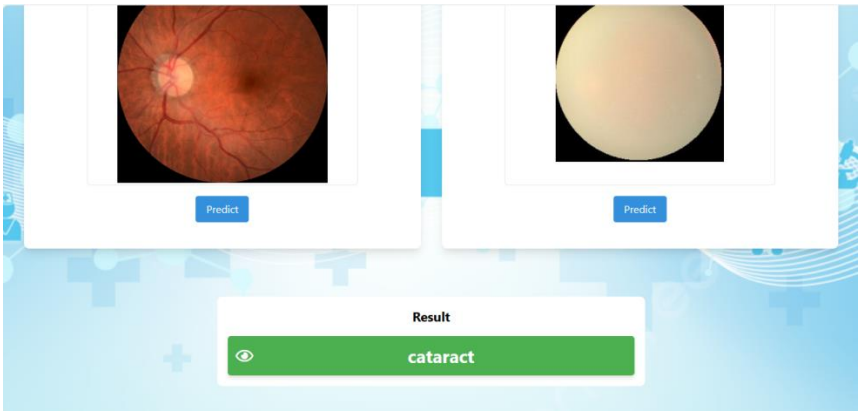

Figure 9. Prediction

Users will see information or labels that describe the analysis results of uploaded retinal images. With this step, users will immediately gain insight into the predicted conditions or characteristics of the image.

4.3. Discussion

4.3.1. Random Forest Classification

Classification Report:

	precision	recall	f1-score	support
normal	0.76	0.84	0.80	245
diabetic_retinopathy	1.00	1.00	1.00	220
glaucoma	0.68	0.62	0.65	183
cataract	0.74	0.70	0.72	196
accuracy			0.80	844
macro avg	0.79	0.79	0.79	844
weighted avg	0.80	0.80	0.80	844

The classification report includes an evaluation of a model's performance in four different classes: "normal," "diabetic_retinopathy," "glaucoma," and "cataract." Each class is analyzed using three main metrics, namely precision, recall, and F1-score. Precision indicates the extent to which the model's positive predictions are correct, recall measures how well the model identifies positive data, and the F1 score is a combination of the two. The report also lists the amount of data support within each class.

Furthermore, the entire model evaluation results are provided in the form of accuracy, which shows the overall percentage of data properly predicted by the model. In this scenario, the model attained an accuracy of 80%, proving its ability to appropriately categorize the data. The indicators above are also aggregated into a macro average (macro avg) and a weighted average (weighted avg), offering a full picture of model performance across all classes. Overall, this report provides extensive insight into the ability of the model to distinguish and categorize data into four different types.

4.3.2. K-Nearest Neighbor Classification

Classification Report:

	precision	recall	f1-score	support
normal	0.61	0.67	0.64	245
diabetic_retinopathy	0.97	0.97	0.97	220
glaucoma	0.60	0.43	0.50	183
cataract	0.60	0.69	0.64	196
accuracy			0.70	844
macro avg	0.70	0.69	0.69	844
weighted avg	0.70	0.70	0.70	844

In the classification report, the performance evaluation of a model on four classes (“normal,” “diabetic_retinopathy,” “glaucoma,” and “cataract”) is presented through several important metrics. Precision, which measures how precisely a model's positive predictions are, varies between 0.60 and 0.97 for different classes. Recall, which shows how well the model identifies positive data, ranged between 0.43 and 0.97. The F1 score (F1-score), which combines both precision and recall, is in the range of 0.50 to 0.97.

When looking at overall accuracy, the model managed to correctly predict around 70% of the total data (844 data) evaluated. The "macro avg" and "weighted avg" metrics provide an average representation of key metrics across classes. Macro average (macro avg) reflects the overall average of metrics in each class without taking into account the data distribution. Meanwhile, the weighted average (weighted avg) gives weight to each class based on the distribution of the data.

Overall, this report shows that the model does an outstanding task of categorizing various classes, particularly "diabetic_retinopathy." However, there was some variation in performance among classes, with some classes, such as "glaucoma," having low recall. Even while the model's accuracy exceeds 70%, there is still potential for improvement, particularly in recognizing classes with lower recall.

4.3.3. Comparison of Classification Reports

1. General Accuracy

The Random Forest method achieves an accuracy level of 80%, while the KNN method achieves an accuracy of 70%. This shows that Random Forest has better performance in correctly predicting the presence of eye disease in the given dataset.

2. Precision and Recall

Precision in each class ranges from 0.68 to 1.00 in the Random Forest Method, while recall ranges from 0.62 to 1.00. This demonstrates that the model is capable of correctly classifying and identifying positive data in almost all classifications.

The KNN method, on the other hand, has greater fluctuation in precision and recall. Precision varied from 0.60 to 0.97, while recall was 0.43 to 0.97. This demonstrates that the KNN approach performs better in identifying and classifying data across several classes.

3. F1-Score

The F1 score combines precision and recall, providing a holistic picture of model performance in the face of trade-offs between the two metrics [20]. In both methods, the F1 score reflects the level of agreement between precision and recall in each class. Random Forest has a more consistent range of F1 scores, while KNN has greater variation.

4. Best and Worst Performing Classes

The class "diabetic_retinopathy" achieves perfect precision and recall (1.00) in the Random Forest Method, demonstrating outstanding performance in recognizing and categorizing this class. The "glaucoma" class, on the other hand, showed poorer precision and recall (0.68 and 0.62, respectively), indicating difficulties in data classification in this class.

The "diabetic_retinopathy" class exhibits good precision and recall in the KNN approach as well. The "glaucoma" class, on the other hand, had a low recall (0.43), showing difficulties in detecting positive data for this class.

5. Accuracy vs. Accuracy Considerations

The Random Forest method is more accurate, but the KNN method has more variability in precision, recall, and F1 score. This demonstrates that, while KNN may not always give accurate predictions, its performance in recognizing diverse classes is substantially more constant.

A comparison of these two methods reveals that none is perfect in every situation. In general, the Random Forest approach offers more constant performance and a better level of accuracy. However, the KNN method has significant advantages, particularly in giving more consistent predictions for most classes. The approach chosen is determined by the features of the dataset, the application requirements, and the trade-off between accuracy and consistency in classification.

5. Conclusion

In this study, two extensively used methods for classification, Random Forest (RF) and K-Nearest Neighbor (KNN) were compared in the context of eye disease detection using image datasets with four different classes. According to the research findings, Random Forest has an accuracy rate of around 80% with 100 trees, whereas KNN has an accuracy rate of around 70% with K=3. In other words, in this dataset, Random Forest succeeds at classifying eye disorders. However, KNN has several advantages, particularly in making more consistent predictions across multiple classes. In conclusion, Random Forest performs better when dealing with the complexities of image fluctuations associated with eye illness. The findings of this study suggest that Random Forest is a promising method for classifying eye diseases based on picture data; however, more research is needed to gain a greater understanding of the subject matter.

References

[1] S. N. Sari, B. S. Ginting, and N. Novriyenni, “Design of a walking aid for the blind using Arduino-based Fuzzy Logic,” *JTIK (Jurnal Tek. Inform. Kaputama)*, vol. 6, no. 2, pp. 528–543, 2022.

[2] D. A. Dharmawan, “Retinal Vessel Segmentation to Support Foveal Avascular Zone Detection,”

Telemat. J. Inform. dan Teknol. Inf., vol. 20, no. 1, pp. 41–50, 2023.

[3] A. U. Detty, I. Artini, and V. R. Yulian, "Characteristics of Risk Factors for Cataract Sufferers," *J. Ilm. Kesehat. Sandi Husada*, vol. 10, no. 1, pp. 12–17, 2021.

[4] M. A. K. Faadhil, P. R. A. Sangging, and R. Himayani, "Relationship between Glaucoma and Hypertension," *Med. Prof. J. Lampung*, vol. 13, no. 4.1, pp. 36–41, 2023.

[5] Y. Triyani, "Classification of diabetic retinopathy on fundus images based on deep learning," *ABEC Indones.*, vol. 9, pp. 1007–1018, 2021.

[6] N. I. Widiastuti, E. Rainarli, and K. E. Dewi, "Peringkasan dan Support Vector Machine pada Klasifikasi Dokumen," *J. Infotel*, vol. 9, no. 4, pp. 416–421, 2017.

[7] G. Mercy, "Analisis Penyakit Disleksia Pada Anak Balita Akibat Kurangnya Asupan Vitamin A," 2019.

[8] D. A. Budi, "Perancangan Sistem Login Pada Aplikasi Berbasis GUI Menggunakan QTDesigner Python," *J. SIMADA (Sistem Inf. dan Manaj. Basis Data)*, vol. 4, no. 2, pp. 92–100, 2021.

[9] P. A. Nugroho, I. Fenriana, and R. Arijanto, "Implementasi Deep Learning Menggunakan Convolutional Neural Network (Cnn) Pada Ekspresi Manusia," *Algor*, vol. 2, no. 1, pp. 12–20, 2020.

[10] A. Putri and R. M. Awangga, *Membangun Frontend dan Backend Packages dengan Golang" Studi Kasus Sistem Administrasi"*. Penerbit Buku Pedia, 2023.

[11] Y. T. Widayati, Y. Prihati, and S. Widjaja, "Analisis Dan Komparasi Algoritma Na Ve Bayes Dan C4. 5 Untuk Klasifikasi Loyalitas Pelanggan Mnc Play Kota Semarang," *J. Transform.*, vol. 18, no. 2, pp. 161–172, 2021.

[12] M. R. A. Nasution and M. Hayaty, "Perbandingan Akurasi dan Waktu Proses Algoritma K-NN dan SVM dalam Analisis Sentimen Twitter," *J. Inf.*, vol. 6, no. 2, pp. 226–235, 2019.

[13] P. Rosyani, S. Saprudin, and R. Amalia, "Klasifikasi Citra Menggunakan Metode Random Forest dan Sequential Minimal Optimization (SMO)," *J. Sist. dan Teknol. Inf.*, vol. 9, no. 2, p. 132, 2021, doi: 10.26418/justin.v9i2.44120.

[14] V. W. Siburian and I. E. Mulyana, "Prediksi Harga Ponsel Menggunakan Metode Random Forest," in *Annual Research Seminar (ARS)*, 2019, vol. 4, no. 1, pp. 144–147.

[15] Z. Nurfadilla, "Implementasi Data Mining untuk Memprediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Random Forest," *AGENTS J. Artif. Intell. Data Sci.*, vol. 2, no. 2, pp. 35–42, 2022.

[16] C. Kurniawan and H. Irsyad, "Perbandingan Metode K-Nearest Neighbor Dan Naïve Bayes untuk Klasifikasi Gender Berdasarkan Mata," *J. Algoritm.*, vol. 2, no. 2, pp. 82–91, 2022.

[17] D. Prasetyawan and R. Gatra, "Algoritma K-nearest neighbor untuk memprediksi prestasi mahasiswa berdasarkan latar Belakang pendidikan dan ekonomi," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 7, no. 1, pp. 56–67, 2022.

[18] A. M. Zuhdi, E. Utami, and S. Raharjo, "Analisis sentiment twitter terhadap capres Indonesia 2019 dengan metode K-NN," *J. Inf. J. Penelit. dan Pengabdi. Masy.*, vol. 5, no. 2, pp. 1–7, 2019.

[19] D. Cahyanti, A. Rahmayani, and S. A. Husniar, "Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 39–43, 2020.

[20] I. I. Sholikhah, A. T. J. Harjanta, and K. Latifah, "Machine Learning Untuk Deteksi Berita Hoax Menggunakan BERT," in *Prosiding Seminar Nasional Informatika*, 2023, vol. 1, no. 1, pp. 524–531.