A Comparative Study of Deep Learning Techniques for Ocular Diseases

Amal AlShahrani [1], Ruba Hassan Balubaid [1], Rafaa Ismail Alowaybidi [1], Hadeel Abdulaziz Alnasiri [1], Jumanah Moqbel Alsehli [1], Sarah Saeed Alshehri [1]

[1] Department of Computer Science and Artificial Intelligence, College of Computers, Umm Al-Qura University, Makkah, Saudi Arabia

Abstract

This study conducts a comparative evaluation of two cutting-edge deep learning models, You Only Look Once (YOLO) and VGG16, utilizing fundus images for automated ocular disease classification. The research endeavors to discern between Normal (N), Diabetes (D), Glaucoma (G), and Cataract (C), prevalent in fundus imagery. Fundus images, being a cornerstone in ophthalmic diagnostics, pose unique challenges due to variations in image quality, pathology manifestation, and disease complexity. By rigorously comparing the performance, strengths, and limitations of YOLOv8, YOLOv5, and VGG16 on this specific dataset, this study aims to provide insights into their efficacy in accurately diagnosing ocular conditions. The outcomes of this investigation have the potential to advance the development of more precise and efficient automated diagnostic systems for ocular diseases, thereby facilitating early intervention and improving patient care in ophthalmology.

Keywords:

Deep Learning, YOLO (You Only Look Once), VGG16, Fundus Images, Ocular Disease Classification

1. Introduction

Ocular diseases such as diabetic retinopathy, glaucoma, and cataracts are leading causes of vision impairment and blindness across the globe. The world faces the challenge of detecting eye diseases, especially glaucoma. Early detection and accurate diagnosis are pivotal for effective management and treatment of these conditions. While the diagnosis of such diseases traditionally depends on the manual interpretation of fundus images by experienced ophthalmologists, this process can be subjective and prone to errors, especially in underserved areas with a shortage of specialists [1] [3].

Advancements in deep learning have revolutionized the field of medical image analysis, offering new avenues to enhance diagnostic precision and reliability. Convolutional neural networks (CNNs), a class of deep learning algorithms, are particularly well-suited for image recognition tasks and have been increasingly adopted for the automated analysis of medical imagery, including ocular fundus photographs.

This study focuses on a comparative evaluation of two prominent CNN architectures: You Only Look Once (YOLO) and VGG16. YOLO is primarily renowned for its efficiency in object detection tasks, offering a unique approach that divides images into regions and predicts bounding boxes and probabilities for each region simultaneously. This makes it particularly fast and effective in real-time applications. On the other hand, VGG16 is known for its simplicity and depth, which have proven effective in various image classification challenges, making it a benchmark in the field of visual recognition tasks [2].

By conducting a thorough comparative analysis of these two models on a dedicated dataset of fundus images, this paper aims to uncover their respective strengths and limitations in the context of ocular disease classification. The objective is to determine which model provides more accurate and reliable performance, thereby guiding future applications in automated ocular diagnostics and contributing to the broader field of medical image analysis.

2. Literature Review

In Elloumi et al. [5] discussed the use of deep learning in diagnosing eye diseases through images of the fundus of the eye. It presents a survey of various methods for detecting eye diseases based on deep learning, analyzing preprocessing steps, and neural network architectures. It also discusses the hardware and software environment required to employ deep learning architecture, the principles of experimentation involved, and the databases used in the training and testing stages. The paper identifies a significant difference in the sizes of input images used in various deep learning architectures. It discusses two categories of deep learning architectures, those based entirely and partially on deep learning, for detecting eye diseases, with several methods falling into each category, always incorporating a preprocessing step of the fundus image before deep learning processing. These methods are characterized by higher detection

Manuscript received May 5, 2025
Manuscript revised May 20, 2025
https://doi.org/10.22937/IJCSNS.2025.25.5.3

performance, thanks to their ability to tailor the network regarding the detection goal.

Ahmad et al. [7] paper aimed to classify external eye diseases using a dataset of images obtained from a digital camera. The dataset is categorized into four major classes based on the part of the eye affected by the disease, with further subdivisions for each class. The paper proposes the use of a Hierarchical multi-label classification (HMC) technique for classification, which is a widely used approach in text classification, image annotation, and bioinformatics problems, where examples can be assigned to multiple paths of the class hierarchy simultaneously. The HMC approach involves assigning each example a subset of consistent labels based on a hierarchical structure. The paper also extracts color histogram features and law texture features from the images for classification. The paper uses algorithms and hierarchical multi-label Artificial Neural Network for classification. The overall prediction accuracy achieved is 75.7142%. The paper suggests that additional external eye diseases like cysts, glaucoma, keratitis, and uveitis can be added to the classification in the future. Expanding the dataset with more examples of minor classifications is also recommended to improve classification accuracy.

Li et al. [8] discussed learning technology that was conducted to provide an in-depth assessment of the levels of computational pathology. The study includes the Ocular Image Analysis-Intelligent Ocular Disease Recognition (OIA-ODIR) dataset, which includes 10,000 images of the right and left eyes of 5,000 patients. 9 different versions of the synthetic networks Vgg-16, ResNet-18, ResNet-50, ResNeXt-50, SE-ResNet-50, SE ResNeXt-50, Inception-v4, Densenet, and CafeNe were used. The fine-line network was defined in Two sets of treatments using three different methods (SUM, PROD and CONCAT) combine features for analysis in multiple disease classifications. Through experimental verification, they found that the only element combination method performs better compared to other methods. It does not improve performance, but increasing network width can produce better results, and consolidating computer networks can help improve performance.

Dipu et al [9] addressed the challenge of early and accurate ocular disease detection through automated diagnosis using retinal fundus images. To achieve this, they implemented different deep learning models, including ResNet-34, EfficientNet, MobileNetV2, and VGG-16, on a large dataset that contains 5000 cases of color fundus photographs (CFPs). The dataset used by the authors is the Ocular Disease Intelligent Recognition (ODIR) dataset, which is publicly available, and it is split into eight different ocular disease classification categories. Results showed that the VGG-16 model achieved the highest accuracy (97.23%) among the tested models. The authors suggest further development of such systems to build a user-friendly and real-time ocular disease classification system.

Bernabe et al. [10] paper aimed to address the problem of eye diseases, specifically Glaucoma and Diabetic Retinopathy, which are significant global health issues. The purpose is to develop an intelligent pattern classification algorithm based on Convolutional Neural Networks (CNNs) as the primary algorithm to accurately detect and classify these diseases. CNN is trained using two different datasets of retinography images of Glaucoma and Diabetic Retinopathy. The training process involves K-fold Cross Validation to validate the performance of the algorithm. The accuracy percentage of the proposed classifier is reported to be 99.89%. Additionally, numerical metrics such as recall, specificity precision, and F1 score, all with values close to 1, support the suitable performance of the classifier. The paper suggests future work that could involve improving image analysis by implementing new channels for the RGB matrix and analyzing healthy images as well. Additionally, further research could focus on classifying other eye diseases and expanding the application of the proposed algorithm to a broader range of conditions.

Ling et al. [11] introduced DeepDR, a deep learning system tailored for the precise detection of diabetic retinopathy. Trained on a vast dataset of 466,247 fundus images, DeepDR demonstrates exceptional accuracy in grading retinopathy severity and identifying lesions. The study's utilization of IoU metrics for evaluating the segmentation network further emphasizes its reliability. Utilizing Python 3.7.1 and OpenCV 2 for analysis and processing, this study marks a significant advancement in diabetic retinopathy diagnosis. However, the study lacks a comparative analysis of DeepDR's performance against existing methods or systems for diabetic retinopathy detection and grading.

Shamsan et al. [12] conducted a study to explore the use of deep learning for the classification of eye diseases in color fundus photographs (CFP). Early detection and accurate classification of these diseases are crucial for preventing blindness. However, it can be challenging to differentiate between early-stage diseases. The authors proposed a hybrid approach that combines feature extraction with fusion methods to improve classification accuracy. They implemented three methods: first, they classified features extracted from separate MobileNet and DenseNet121 models with an artificial neural network (ANN) after dimensionality reduction using Principal Component Analysis (PCA). Second, they

classified fused features from both models with an ANN, again with dimensionality reduction. Finally, they classified fused features alongside handcrafted features using an ANN. The dataset used in the study is the OIH dataset, which includes 4217 CFP images of three types of eye disease and a normal class. The third method achieved the best results with an AUC of 99.23% and an accuracy of 98.5%, demonstrating the potential of combining deep learning features with handcrafted ones for accurate eye disease classification.

Arif et al. [13] conducted a study on classifying eye diseases in fundus images using the CNN based EfficientNet-B0 architecture. They categorized fundus images into normal, cataract, and glaucoma classes, achieving an accuracy of 79.22% with precision, recall, and F1-score values exceeding 78%. The research demonstrated the potential of EfficientNet in enhancing the diagnostic process for eye conditions, showcasing improved performance metrics compared to previous studies. By leveraging deep learning and advanced image processing techniques, Arif et al. (2023) highlighted the efficiency and accuracy of CNN models in classifying eye diseases based on fundus images. These findings underscore the significance of utilizing advanced architectures like EfficientNet in medical imaging applications, paving the way for more accurate and efficient diagnosis of eye diseases with implications for enhancing patient care and treatment outcomes.

Babaqi et al. [14] conducted a study on eye disease classification using deep learning techniques, focusing on the differentiation of normal eyes from those affected by diabetic retinopathy, cataracts, and glaucoma. The research utilized convolutional neural networks (CNNs) and transfer learning to achieve high accuracy rates in multi-class classification tasks. Transfer learning, a method where a model developed for one task is repurposed for another, played a significant role in optimizing the classification of eye diseases. The dataset consisted of approximately 4200 colored images of normal eyes, cataracts, diabetic retinopathy, and glaucoma, which were preprocessed and divided into training, testing, and validation subsets for model evaluation. The study demonstrated that transfer learning outperformed the traditional CNN approach, achieving a 94% accuracy rate compared to 84%. Evaluation metrics such as precision, recall, and F1-score were employed to assess the model's performance, highlighting the impact of transfer learning on enhancing CNN accuracy.

Afsana et al. [15] introduce a novel approach to automate the detection and classification of eye diseases from fundus images, eliminating the need for time-consuming manual evaluation by experts. Their method employs a deep Convolutional Neural Network (CNN)-based ensemble model with 20 layers, incorporating various activation, optimization, and loss functions. By utilizing pre-processing techniques like contrast-limited adaptive histogram equalization (CLAHE) and a Gaussian filter, image quality is improved and noise is reduced. Augmentation techniques during training prevent overfitting, ensuring model robustness. The CNN model is compared with pre-trained models (VGG16, DenseNet201, and ResNet50), demonstrating superior performance. Experimental results on the ODIR dataset validate their approach, marking a significant advancement in automated eye disease detection.

Current research on ocular disease classification using retinal images often relies on deep-learning models such as YOLO and VGG16. However, these models are typically utilized in isolation, which restricts their capacity to produce a comprehensive evaluation of ocular health. Furthermore, there is a lack of thorough comparison regarding the efficiency and accuracy of these models in disease classification [4].

Our research effectively addresses these limitations by conducting an extensive comparative study between YOLO and VGG16. In addition to comparing these two models, we further explore the topic by incorporating two different versions of YOLO, namely versions 8 and 5. This comprehensive approach enables us to examine the influence of YOLO architecture choice on performance in ocular disease diagnosis. This thorough analysis holds significant importance in the field for various reasons.

First and foremost, it furnishes valuable insights into the strengths and weaknesses of each model when utilized for ocular disease classification. By discerning which model excels at identifying specific types of ocular conditions, researchers can customize their deep-learning approach to enhance disease detection [6]. Additionally, the comparison of these models' efficiency is of utmost importance. In real-world scenarios, processing speed can be a crucial factor, particularly in time-sensitive settings. Understanding which model provides the optimal balance between accuracy and efficiency can be pivotal for clinical utilization.

Ultimately, this study sets the stage for future exploration of the potential advantages of integrating these models. Through the amalgamation of YOLO and VGG16, a more resilient and precise automated diagnosis system for ocular diseases could be formulated. Such a system has the potential to substantially enhance early detection rates and, in turn, improve patient outcomes. In summary, our

comparative study not only illuminates the distinct capabilities of these models in ocular disease diagnosis but also establishes a foundation for future progress in leveraging deep learning algorithms for a more thorough evaluation of eye health.

3. Dataset collection

The "Eye-Disease Image Dataset" on Roboflow is a curated collection of 2555 total retinal images used for the classification of ocular conditions. The dataset has been structured into three subsets to support the development and validation of machine learning models, (88%) for training and with a validation set (8%), and finally evaluated on a separate test set (4%) to assess their performance.

These images typically represent a variety of ocular conditions, potentially including but not limited to normal retinal images, diabetic retinopathy, glaucoma, and cataracts. The images may vary in terms of presentation and severity of the conditions depicted, offering a diverse range of cases for comprehensive model training.

The dataset classifies patients into four labels, including:

- Normal (N)
- Glaucoma (G)
- Cataract (C)
- Diabetes (D)

4. Methodology

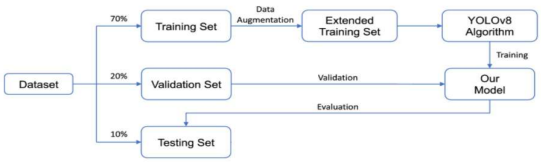


Figure 1: Methodology.

4.1 Data Preprocessing:

A. Data Acquisition:

A subset of the "Eye-Disease Image Dataset" from Roboflow is utilized, comprising:

- Training Set: 70% of the dataset (448 images).
- Validation Set: 20% of the dataset (128 images).
- Test Set: 10% of the dataset (64 images).

B. Label Encoding:

The dataset categorizes patients into four labels: Normal (N), Diabetes (D), Glaucoma (G), and Cataract (C).

C. Data Augmentation:

Rotation:
Images in the training set are rotated by various degrees (e.g., 90°, 180°, 270°) to introduce diversity and enhance model robustness. Rotation ensures that the model learns to classify ocular conditions regardless of their orientation within the image.

Resizing to 640x640 pixels:
All images are resized to a standard dimension of 640x640 pixels. Resizing ensures uniformity in input size, facilitating efficient model training and inference. This standard size is chosen to balance computational efficiency with sufficient resolution for accurate classification.

4.2 Model Selection:

In our effort to effectively classify ocular conditions, we have carefully selected a range of models designed to address different aspects of the task at hand.

1. YOLOv5: As part of our comprehensive evaluation, we have included YOLOv5, a variant of YOLO. By comparing its performance with YOLOv8, we aim to gain insights into the specific strengths and weaknesses of each model, ultimately refining our strategy.

2. YOLOv8: To explore the limitations of YOLOv5, we have included YOLOv8 in our comparative study. This model is an excellent choice due to its seamless integration of object detection and classification capabilities [14]. Its efficiency in simultaneously detecting and categorizing ocular conditions within images positions it as a cornerstone of our approach.

3. VGG16: Adding to our ensemble is VGG16, a well-regarded convolutional neural network architecture specializing in image classification [15]. Leveraging its established reputation, VGG16 serves as a benchmark against which we measure the efficacy of YOLOv8 and YOLOv5. Through this comparative analysis, we aim to discern the most effective approach to ocular condition classification.

4.3 Model Training:

A. YOLOv8

We have leveraged Roboflow, a platform designed to streamline the management and preprocessing of image datasets for machine learning projects. It provides a wide range of tools and functionalities to simplify tasks such as data annotation, augmentation, and integration into machine learning workflows.

In our implementation of YOLOv8, we have categorized datasets into three subsets: training, validation, and testing, utilizing RoboFlow. The YOLOv8 model, the latest iteration (2.0) developed by Ultralytics, has demonstrated substantial effectiveness in the classification of ocular conditions. After a comprehensive evaluation of its performance, we have achieved an impressive validation accuracy of 92.5%. This accuracy is determined by the ratio of correctly labeled images to the total number of images within all validation set samples.

Analysis of training loss:

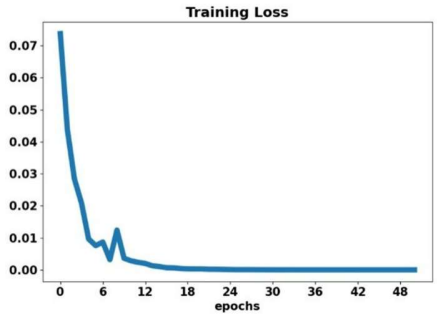


Figure 2: Training loss for YOLOv8.

The above Figure 2 depicting the training loss, provides insight into the model's progression. Initially, there is a rapid decrease, followed by a stabilization, indicating effective learning. The consistently high validation accuracy suggests that the model generalizes well. The primary objective of minimizing loss and maximizing accuracy has been successfully achieved in this instance.

Validation Accuracy:

The validation accuracy graph in Figure 4 shows how well the model performs on new data. It initially fluctuates but stabilizes later, reaching peak accuracy after about 36 epochs. This accuracy refers to the model's ability to predict the highest probability class label accurately. Overall, stability and high validation accuracy indicate the model's good performance.

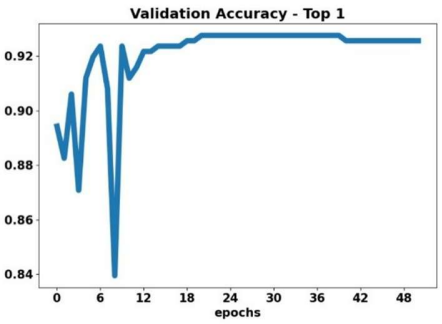


Figure 3: Validation Accuracy for YOLOv8

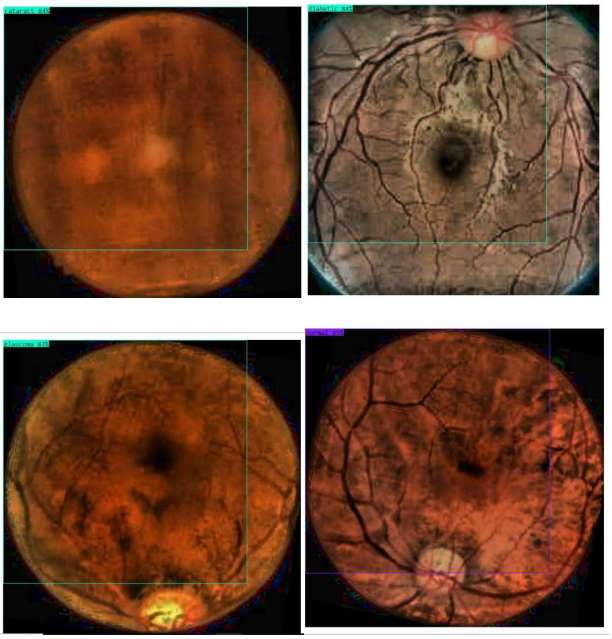


Figure 4: Implementation classification.

B. YOLOv5

Similarly, to train YOLOv5, we split the dataset into three sets: 70% for training, 20% for validation, and 10% for testing. We also pre-processed the data using RoboFlow, which included orientation and resizing the images to 640x640. This was necessary to ensure a fair comparison between our two versions of the YOLO model.

We trained a pre-existing YOLOv5 model for classification tasks, which was also developed by

Ultralytics. We experimented with different numbers of epochs and found that training for 50 epochs produced the best results. However, due to the limitations of the YOLOv5 model architecture compared to YOLOv8, the accuracy of the model remains relatively low. compared to YOLOv8, the accuracy of the model remains relatively low.

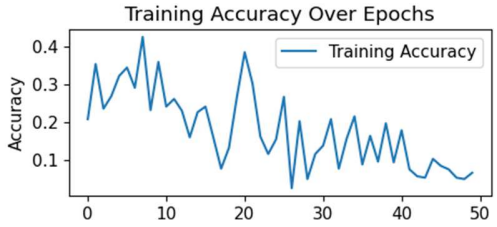


Figure 5: Training accuracy for YOLOv5

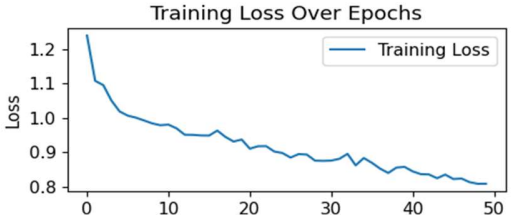


Figure 6: Training loss for YOLOv5

The accuracy of the model reached its peak at the end of the 10th epoch, after which further improvements were observed with continued training beyond this point. As evidenced by the graph illustrating the training loss, it is apparent that the training loss consistently decreased with each successive epoch, ultimately achieving its minimum value at epoch 50.

In terms of overall accuracy across the three data sets--training, validation, and testing--the model attained a rate of 63%. This accuracy is comparatively low when benchmarked against the performance metrics of YOLOv8. Nonetheless, the observed performance is both justifiable and anticipated, in accordance with the documentation provided by the development team at Ultralytics. The documentation indicates that each new iteration of YOLO is engineered to exhibit enhanced performance over its predecessors.

Table 1 below demonstrates the different hyperparameters used to train the two versions of YOLO. Although the hyperparameters used in the two versions of the model were similar, the differences and advances in the YOLOv8 architecture were the primary reasons for the enhancement in the model's performance.

Table 1: Hyperparameters for YOLO

Hyperparameters	YOLOv5	YOLOv8
Input image size	5350	5350
Epochs	50	100
Batch size	16	16
Optimizer	Adam	Adamw
Initial Learning rate	0.01	0.01
Final Learning rate	0.01	0.01

When comparing the hyperparameters of YOLOv5 and YOLOv8, both models share similar settings for input image size, epochs, batch size, initial learning rate, and final learning rate. However, an important distinction lies in the choice of optimizer. While YOLOv5 utilizes the Adam optimizer, YOLOv8 implements AdamW as its optimizer. The use of AdamW in YOLOv8 can provide benefits such as improved generalization and robustness due to its modified weight decay handling. Despite the similarities in most hyperparameters, the enhancements and additional layers in the YOLOv8 model architecture compared to YOLOv5 contribute significantly to its superior performance.

C. VGG16 model:

VGG16 is a convolutional neural network (CNN) architecture proposed by the Visual Engineering Group at the University of Oxford. It has gained popularity due to its simplicity and effectiveness in image classification tasks. The “16” in its name refers to the total number of weight layers it has. VGG16 consists of 16 weight layers, including 13 convolutional layers and 3 fully connected layers. Convolutional layers use small 3x3 filters with one step and zero padding to maintain spatial resolution. VGG16 is deeper than other convolutional networks due to its use of multiple layers stacked on top of each other.

We implemented VGG16 in Colab Pro using a T4 GPU and Keras as the framework. In the first experiment, we trained the model using the default settings: 10 epochs, 32 batch size, and a learning rate of 0.001, resulting in an accuracy of 71.06%. To improve the performance and ensure the avoidance of overfitting, we reduced the number of epochs to 5, which led to a significant increase in accuracy to 76.09%. In the final experiment, aiming for further improvement, we decreased the learning rate to 0.0001, along with freezing the base layers. This adjustment yielded the highest accuracy for the VGG16 model, reaching 78.80%.

The empirical evidence presented in the initial graph showcasing the training and validation loss trends across 10 epochs provides valuable insights into the model's performance dynamics. Initially, a consistent decrease in training loss along with a parallel decline in validation loss suggests a progressive learning pattern. However, a notable deviation occurs after epoch 4, where the validation loss sharply rises while the training loss continues to decrease. This divergence signifies a classic symptom of overfitting, a phenomenon where the model excessively tunes itself to the nuances of the training data, consequently impairing its ability to generalize patterns effectively.

Figure 7: Loss graph

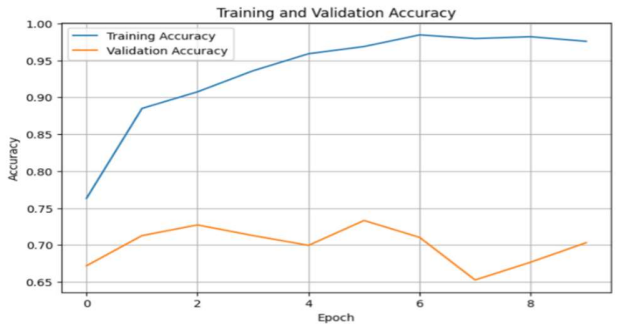


Figure 8: Accuracy graph

Correspondingly, the accuracy graph complements this narrative by showcasing a consistent trend where the training accuracy consistently outperforms the validation accuracy. Such a discrepancy signifies a potential overreliance on specific training data characteristics without a corresponding ability to generalize. Despite a marginal convergence between the two accuracies over successive epochs, the ideal scenario would entail a closer alignment, with the validation accuracy slightly trailing behind. To enhance the model's generalization capacity and mitigate overfitting risks, strategies such as early stopping around epoch 3 or 4, model complexity reduction, regularization implementation, or data augmentation merit consideration.

5. Results & Discussion

For the VGG16 result, we have:

Table 2: VGG16 result table

Epoch	Batch size	Learning rate	Accuracy
10	32	0.001	71.06%
5	32	0.001	76.09%
10	32	0.0001	79.89%

Analyzing the provided training configurations reveals key insights into their performance. Comparing the impact of epochs, we observe that increasing the number of epochs generally leads to improved accuracy, as evidenced by the higher accuracy achieved in configurations with more epochs. However, simply increasing epochs is not the sole factor influencing accuracy, as demonstrated by configuration 3, where a smaller learning rate yields the highest accuracy despite the same number of epochs as configuration 1. This highlights the significance of the learning rate for model convergence and accuracy. Specifically, configuration 3, with a smaller learning rate, outperforms the others, indicating the importance of fine-tuning hyperparameters. While batch size remains constant across all configurations, its influence on accuracy isn't directly evaluated here. In conclusion, these results underscore the importance of carefully tuning hyperparameters, particularly the learning rate, to achieve optimal model performance. Further exploration, potentially involving variations in batch size, could provide deeper insights into enhancing model training and accuracy.

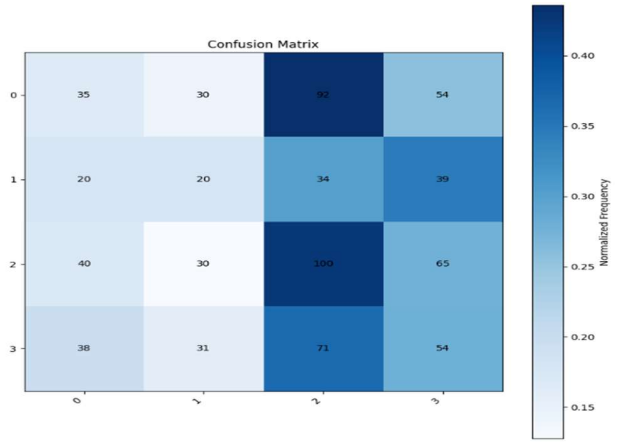


Figure 9: confusion matrix for VGG16

Figure 9 shows the analysis of the confusion matrix highlights the VGG16 notable challenges in accurately classifying specific categories, particularly with evident confusion patterns between Cataract (Class 0), Glaucoma (Class 2), and Normal (Class 3). Glaucoma (Class 2) exhibits the highest true positives at 100, yet simultaneously displays significant misclassifications, particularly stemming from Cataract (Class 0) and Normal (Class 3). Notably, Cataract (Class 0) frequently gets misclassified as Glaucoma (Class 2), while Normal (Class 3) showcases a comparable trend of misidentification. Diabetic (Class 1) demonstrates a more evenly distributed misclassification profile across other classes.

The overarching trend indicates a pronounced struggle for the model to effectively distinguish between these specific categories, thereby underscoring the imperative for enhanced feature engineering, a more balanced training dataset, or the adoption of more sophisticated classification algorithms to bolster its discriminatory prowess. This analysis underscores the critical need for targeted interventions to mitigate misclassification challenges and fortify the model's classification accuracy, especially for the intricate differentiation tasks inherent in the Cataract (Class 0), Glaucoma (Class 2), Normal (Class 3), and Diabetic (Class 1) categories.

For the two versions of YOLO, we have trained the models on different epoch numbers, as mentioned above. However, we have noticed that when training the two models on 100 epochs, it affects the model accuracy badly. Table 3 shows the training accuracy for the two versions of YOLO on different numbers of epochs.

Table 3: Epoch and accuracy for YOLO

Epoch	YOLOv5 accuracy	YOLOv8 accuracy
30	58%	86%
50	63%	92%
100	52%	81%

Figure 10 shows the analysis of the confusion matrix underscores the YOLOv5 model's pronounced classification challenges, particularly in discerning between Cataract, Glaucoma, and Normal classes. Cataract (Class 0) is recurrently misidentified as Glaucoma (Class 2), while Normal (Class 3) exhibits a comparable misclassification trend, indicating a significant overlap in the classification boundaries among these classes. Despite Glaucoma (Class 2) boasting the highest count of true positives (100), it contends with notable misclassifications, predominantly originating from Cataract (Class 0) and Normal (Class 3). In contrast, Diabetic (Class 1) showcases a more equitable distribution of misclassifications but still manifests confusion with other classes.

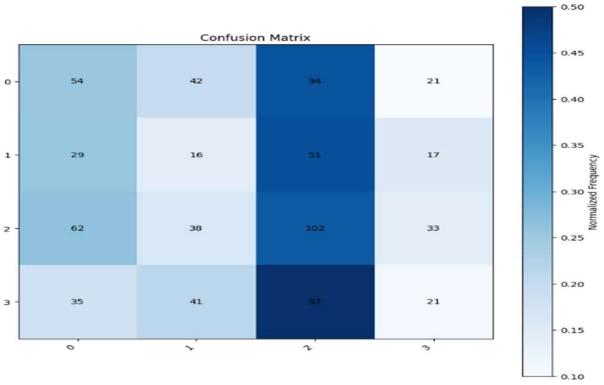


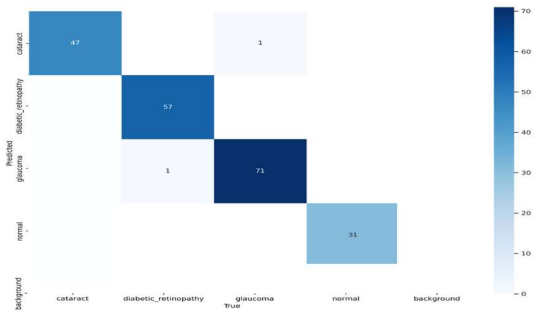
Figure 10: Confusion matrix for YOLOv5

The model's evident struggle to effectively differentiate between these classes highlights the imperative for refined feature engineering, a more diverse and balanced training dataset, or the adoption of advanced classification algorithms to bolster its discriminative acuity and curtail misclassification errors. This analysis accentuates the critical necessity for targeted enhancements to ameliorate the model's classification accuracy, particularly concerning the intricate distinctions within the Cataract (Class 0), Glaucoma (Class 2), and Normal (Class 3) categories, thereby fortifying its overall performance in real-world classification tasks.

Lastly, the analysis of the confusion matrix for YOLOv8 in Figure 11 underscores a commendable level of accuracy, with a predominant alignment of predictions along the diagonal. Notably, the model achieved accurate predictions in 47 cases of Cataract, 57 instances of Diabetic Retinopathy, 71 occurrences of Glaucoma, and 31 occurrences of Normal. Despite this impressive

performance, minor misclassifications were identified, notably a singular case of Cataract and Glaucoma wrongly predicted as Diabetic Retinopathy. These isolated misclassifications hint at areas where further refinement is warranted to enhance the model's capacity to effectively differentiate between closely related conditions.

This evaluation affirms YOLOv8's overall efficacy in classification tasks, as evidenced by the majority of accurate predictions aligning with the respective classes. However, the identification of sporadic misclassifications underscores the continuous need for iterative improvements to fortify the model's precision and mitigate misclassification risks, particularly concerning



nuanced distinctions between analogous conditions.

Figure 11: Confusion matrix for YOLOv8

In the final assessment of the three various deep learning models, namely, YOLOv5, YOLOv8, and VGG16, it is evident that YOLOv8 stands out as the most notable performer, showcasing an impressive accuracy of 92.5%, precision of 88%, and recall of 87%. Particularly in the classification of glaucoma, with a striking accuracy of 93%, YOLOv8 excels in identifying this crucial condition. The model also demonstrates robust performance in detecting diabetic retinopathy and cataracts, with accuracies of 84% in each case. This outstanding performance can be attributed to YOLOv8's advanced architecture, enabling accurate detection across multiple classes.

Comparatively, YOLOv5 presents a competitive option with a decent accuracy of 63%, precision of 72%, and recall of 70%, although it falls short when benchmarked against YOLOv8. Similarly, VGG16, with a respectable accuracy of 78%, precision of 81%, and recall of 77%, trails behind the YOLO models, suggesting potential limitations in its architecture for medical image classification. Overall, the superior accuracy, precision, and recall values, combined with robust performance across various medical conditions, position YOLOv8 as the preferred choice among the evaluated models.

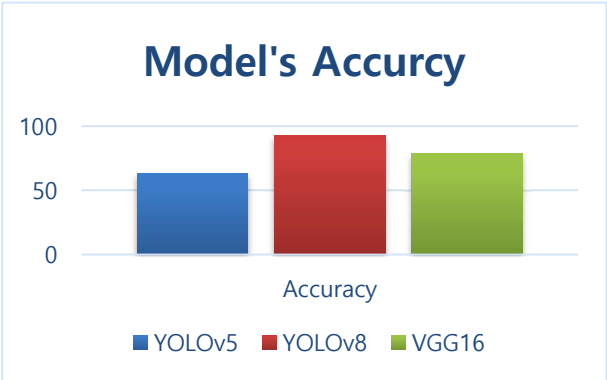


Figure 12: Mode's accuracy comparison

Furthermore, the graphical representation in Figure [12] underscores the accuracy comparison of YOLOv5, YOLOv8, and VGG16, with YOLOv8 leading with an accuracy of around 90%, outperforming the other models. YOLOv5 lags with roughly 70% accuracy, while VGG16 shows moderate performance at about 80%. This reaffirms YOLOv8's effectiveness in medical image classification.

Table 4: Model's accuracy table

Model	Recall	Precision	Accuracy
YOLOv8	87%	88%	92%
YOLOv5	70%	72%	63%
VGG16	77%	81%	78%

Table 4 presents a concise summary of the accuracy metrics for each model, emphasizing YOLOv8's strong performance with 92% accuracy, 88% precision, and 87% recall. In contrast, YOLOv5 and VGG16 exhibit lower accuracy levels, further highlighting the superiority of YOLOv8 in this context.

6. Conclusion

The aim of this study was to assess the effectiveness of deep learning models in classifying ocular diseases based on retinal images. We conducted a comparative analysis of the performance of the YOLOv8,

YOLOv5, and VGG16 models, with a specific focus on varying hyperparameter settings for VGG16.

Our examination of the VGG16 training configurations underscored the importance of hyperparameter tuning, particularly the impact of the learning rate on achieving optimal accuracy. While further research into the effects of batch size is recommended, these findings provide valuable insights for future implementations of VGG16 in ocular disease classification tasks.

Furthermore, the study highlights the clear superiority of YOLOv8 in this particular domain. It achieved an impressive accuracy of 92.5%, surpassing both YOLOv5 (63%) and VGG16 (78.80%) by a significant margin. Notably, YOLOv8 demonstrated exceptional performance in identifying critical conditions such as glaucoma (93% accuracy) and exhibited strong detection capabilities for diabetic retinopathy and cataracts (84% accuracy each). These findings suggest that YOLOv8's advanced architecture is well-suited for the complexities of medical image classification in ocular disease diagnosis.

In summary, this study not only highlights the effectiveness of various deep-learning models in ocular disease classification but also sets the stage for future advancements in this field. YOLOv8's superior accuracy and robust performance across different ocular conditions position it as a promising tool for the development of a more comprehensive and accurate automated diagnosis system, ultimately leading to improved patient care. Subsequent research can explore the potential benefits of integrating these models and further refine hyperparameter optimization techniques to achieve even greater diagnostic accuracy.

Acknowledgment

We extend our sincere gratitude to the faculty and staff of Umm Al-Qura University for their support and guidance throughout the execution of the "Classification of Ocular Diseases: A Comparative Study of YOLOv8, YOLOv5, and VGG16" research project. Special thanks are due to Amal AlShahrani for their invaluable mentorship and expertise, which significantly contributed to the success of this endeavor.

References

[1] Resnikoff, S., Pascolini, D., Etya'ale, D., Kocur, I., Pararajasegaram, R., Pokharel, G. P., & Mariotti, S. P. (2004). Global data on visual impairment in the year 2002. Bulletin of the World Health Organization, 82(11), 844-851.

[2] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[3] Abbas, Q. (2017) 'Glaucoma-deep: Detection of glaucoma eye disease on retinal fundus images using Deep Learning', *International Journal of Advanced Computer Science and Applications*, 8(6). doi:10.14569/ijacsa.2017.080606.

[4] Grewal, Parampal S., et al. "Deep Learning in Ophthalmology: A Review." Canadian Journal of Ophthalmology, vol. 53, no. 4, Aug. 2018, pp. 309–313, https://doi.org/10.1016/j.jcjo.2018.04.019.

[5] Elloumi, Y., Akil, M. and Boudegga, H. (2019) 'Ocular diseases diagnosis in fundus images using a deep learning: Approaches, tools and performance evaluation', Real-Time Image Processing and Deep Learning 2019 [Preprint]. doi:10.1117/12.2519098.

[6] Ting, Daniel S.W., et al. "Deep Learning in Ophthalmology: The Technical and Clinical Considerations." Progress in Retinal and Eye Research, vol. 72, Sept. 2019, p. 100759, https://doi.org/10.1016/j.preteyeres.2019.04.003.

[7] H. Ahmed and S. Hameed, "Eye Diseases Classification Using Hierarchical MultiLabel Artificial Neural Network," IEEE, Jul. 2020, doi: https://doi.org/10.1109/it-ela50150.2020.9253120.

[8] Li, N. et al. (2021) 'A benchmark of ocular disease intelligent recognition: One shot for multi-disease detection', Benchmarking, Measuring, and Optimizing, pp. 177–193. doi:10.1007/978-3-030-71058-3_11.

[9] D ipu, Nadim Mahmud, et al. "Ocular Disease Detection Using Advanced Neural Network Based Classification Algorithms." ASIAN JOURNAL of CONVERGENCE in TECHNOLOGY, vol. 7, no. 2, https://doi.org/10.33130/ajct.2021v07i02.019.

[10] O. Bernabe, E. Acevedo, A. Acevedo, R. Carreno, and S. Gomez, "Classification of Eye Diseases in Fundus Images," IEEE Access, vol. 9, pp. 101267–101276, 2021, doi: https://doi.org/10.1109/access.2021.3094649.

[11] Dai, Ling & Wu, Liang & Li, Huating & Cai, Chun & Wu, Qiang & Kong, Hongyu & Liu, Ruhan & Wang, Xiangning & Hou, Xuhong & Liu, Yuexing & Long, Xiaoxue & Wen, Yang & Lu, Lina & Shen, Yaxin & Chen, Yan & Yang, Xiaokang & Zou, Haidong & Sheng, Bin & Jia, Weiping. (2021). A deep learning system for detecting diabetic retinopathy across the disease spectrum. Nature Communications. 12. 10.1038/s41467-021-23458-5.

[12] Shamsan, Ahlam, et al. "Automatic Classification of Colour Fundus Images for Prediction Eye Disease Types Based on Hybrid Features." Diagnostics, vol. 13, no. 10, 11 May 2023, p. 1706, https://doi.org/10.3390/diagnostics13101706.

[13] Arif, Z., Fu'adah, R. Y. N., Rizal, S., & Ilhamdi, D. (2023). Classification of eye diseases in fundus images using convolutional neural network (CNN) method with efficientnet architecture. Classification of eye diseases in fundus images using Convolutional Neural Network (CNN) method with EfficientNet architecture | Arif | JRTI (Jurnal Riset Tindakan Indonesia) (iicet.org)

[14] Babaqi, Tareq & Jaradat, Manar & Yildirim, Ayse & Al-Nimer, Saif & Won, Daehan. (2023). Eye Disease Classification Using Deep Learning Techniques. 10.48550/arXiv.2307.10501.

[15] J eny, Afsana & Junayed, Masum Shah & Islam, Md Baharul. (2023). Deep Neural Network-Based Ensemble Model for Eye Diseases Detection and Classification. Image Analysis & Stereology. 42. 77-91. 10.5566/ias.2857.