EMERGING ADVANCES IN INTEGRATED TECHNOLOGY
e-ISSN: 2773-5540

EmAIT

Vol. 6 No. 1 (2025) 28-39
https://publisher.uthm.edu.my/ojs/index.php/emait

Comparative Evaluation of Diabetic Retinopathy Detection Using VGG-16 and ResNet-50 Models: Insights from Matthew's Correlation Coefficient and Cohen's Kappa Metrics

Leon Ewe[1]*, Aisha Siddiqa[1]

[1] *School of Computing*
Universiti Utara Malaysia (UUM), Sintok, Kedah, MALAYSIA

*Corresponding Author: ewe_leon2@ahsgs.uum.edu.my
DOI: https://doi.org/10.30880/emait.2025.06.01.004

Article Info

Received: 2 December 2024
Accepted: 17 January 2025
Available online: 10 June 2025

Keywords

Comparative evaluation, convolutional neural networks, diabetic retinopathy, ResNet-50, VGG-16

Abstract

The diagnosis of diabetic retinopathy utilizing artificial intelligence is a subject of debate, particularly involvement of convolutional neural networks (CNN). This article employs VGG-16 and ResNet-50 models for the diagnosis of degrees of diabetic retinopathy by analyzing the common retinal features of diabetic retinopathy, which are microaneurysms, hemorrhages, macula oedema, and exudates. The datasets from the APTOS 2019 Blindness Detection fundus images, which were resized to 224×224 pixels, are used for model training to prevent overfitting and processed via yCbCr color filter to emphasis on the image brightness for ease of diagnosis and evaluation. The evaluation results were determined by the metrics of Matthew's Correlation Coefficient (MCC) and Cohen's Kappa, both capable of handling imbalanced data, ensure reliable agreement with true labels, and provide balanced insights into the model's predictive accuracy. The findings revealed that the VGG-16 performed better than ResNet-50 via measurement of Matthew's Correlation Coefficient in classifying images in a balanced manner, while ResNet-50 performs better than VGG-16 according to values of Cohen's Kappa thanks to the deep layers of the model.

1. Introduction

Diabetic Retinopathy (DR) is a degenerative eye condition that contributes to major vision impairment among adults, specifically among the elderly population. Often occur as a symptom of diabetes mellitus, DR causes damage to the retinal vessels which causes leakages, hemorrhages and catastrophic vision loss if proper treatment is neglected (National Eye Institute, 2024). Proliferative Diabetic Retinopathy (PDR), a stage in which aberrant, fragile vessels grow on the retina, while Non-Proliferative Diabetic Retinopathy (NPDR) is defined by modest bulges in retinal blood vessels and minor leakages. A study conducted by Lima et al. (2016) estimated that 35.0% across 93 million people worldwide suffer from various stages of diabetic retinopathy (DR), making it a critical public health concern especially among the elderly and at-risk populations.

Diabetic retinopathies were typically categorized according to the number of microaneurysms, hemorrhages, macula oedema, and exudates present in the retina as shown in Fig.1, obtained from one of the sample proliferate diabetic retinopathy images from the 2019 APTOS Blindness Images. Microaneurysms, as one of the visible signs of diabetic retinopathy, occurs in the form of tiny bulges in the vessels of the retina that poses a risk of blood or fluid leakage. Retinal hemorrhage, which is correlated with the increasing number of such features with an

increasing severity of diabetic retinopathy, occurs when blood vessels rupture and bleed into retinal tissues. Macula oedema, which its presence poses a higher severity risk of diabetic retinopathy, occurs by the thickening of the macula. When a leakage in retinal blood vessels happen, usually accompanied with microaneurysms, into the macula, it poses a risk of central vision loss. Exudates, a common retinal feature that occurs among severe and proliferate cases of diabetic retinopathy, are yellowish deposits of lipids accumulating in the retina caused by leakage of retinal blood vessels which forms in clusters around damaged vessels and are a sign of blood-retina barrier breakdown.

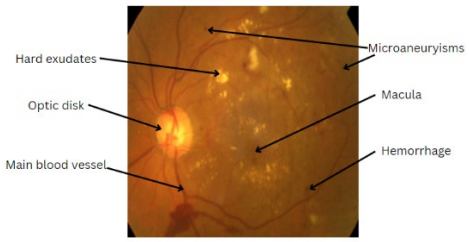


Fig. 1 *Sample of processed images with labelled DR stages*

Various studies have been conducted for the diagnosis of diabetic retinopathy using artificial intelligence, specifically in the field of CNN. The related studies include the utilization of VGG-19, ResNet-101, MobilenetV2, and Inception-ResNet apart from VGG-16 and ResNet-50. These models were then evaluated and analyzed for determining the efficient models to be utilized in this article, and both VGG-16 and ResNet-50 were determined considering the computational costs and efficiency of the models evaluated, and the utilization of semantic segmentation architectures used such as DeepLabV3+ and PSP-Net+. In summary, the studies conducted underscore the comparative evaluation of various classification models like VGG-16 and ResNet-50, showcasing a robust pipeline for effective diabetic retinopathy diagnosis.

In this article, we propose a methodology that is to conduct a comparative evaluation of both VGG-16 and ResNet-50 CNN models by measuring their respective Matthew's Correlation Coefficient (MCC) and the Cohen's Kappa, which were measured and calculated based on the obtained values of true positives, true negatives, false positives and false negatives, obtained prior to the preprocessing of the images that were pre-resized into 224×224 pixels and applying yCbCr color filter. The rationale for using MCC and Cohen's Kappa for CNN model evaluation is to analyze and evaluate their ability to handle imbalanced data, ensure reliable agreement with true labels, and provide balanced insights into the predictive accuracy of the involved models. This is especially vital in clinical diagnostics, where models need to be both statistically robust and clinically interpretable to support effective patient care. This field of study represents the analysis on the performance comparison between VGG-16 model and ResNet-50 model, which is one of the subfields of the CNN model that adapts the use of artificial intelligence in the medical field. The evaluation and review of the models used in this article provides an insight on the efficiency and reliability of the CNN models in evaluating medical diagnosis symptoms.

The remaining parts of this article consists of four (4) sections: Sections 1, 2, 3, and 4. Section 1 explains the Related Work which consists of the utilization of VGG-16 and ResNet-50 models in other fields of study, utilization of other CNN models in various studies, and the applied databases in the detection of diabetic retinopathy. Section 2 describes the Methodology which explains the hardware, models and databases used for this article, apart from the. Section 3 describes the Results and Findings obtained from the epoch training of both VGG-16 and ResNet-50 models. Section 4 consists of the Conclusion which summarizes the article.

2. Related Works

This section describes the literature review and related works that compromises three sub-sections: the utilization of VGG-16 and ResNet-50 models, utilization of CNN models and the datasets used for diabetic retinopathy detection.

2.1 Utilization of VGG-16 and ResNet-50 models

Convolutional neural networks (CNN) have been employed in the diagnosis of diabetic retinopathy mainly by commonly used CNN models, which are VGG-16 and ResNet-50. VGG-16 model is primarily utilized for several clinical diagnosis part from diabetic retinopathy, which Melinda et al. utilized ResNet-50 was used for the facial classification of autism spectrum disorder using DeepLabV3+, a semantic segmentation architecture developed from DeepLabV2+[1]. By leveraging ResNet-50 as the backbone, the study detects facial feature patterns associated with autism spectrum disorder (ASD), which are often subtle and require high-resolution feature extraction as ASD diagnosis especially facial recognition, poses a risk of misdiagnosis due to the high misdiagnosis

nature of patients with ASD. DeepLabV3+, which is particularly effective for semantic segmentation tasks, was combined with ResNet-50 in this study to enhance the precision of classification by segmenting facial regions and focusing on distinguishing characteristics, which was proven with the post-segmentation accuracy of 85.9%, an increasing trend compared to pre-segmentation output.

Besides, Sharma and Guleria conducted an analysis of pneumonia detection utilizing VGG-16 model on the datasets of x-ray images of the patients' chests. The success detecting of the subtle radiographic features indicative of pneumonia utilizing VGG-16 in pneumonia detection, as demonstrated in this study, reinforces the model's versatility and utility in various diagnostic applications, including diabetic retinopathy, pneumonia, and other imaging-based diagnostics in healthcare [2]. PSP-Net+ is also employed by a study conducted by Ye et al. for utilization of VGG-16 model in diagnosis of prostate tumour, which employs the usage of PSP-Net+VGG-16 deep learning model. The study indicates a precise classification accuracy and recognition rates on prostate MRI images which produces a superior accuracy in processing prostate tumour diagnosis due to the [3]. The comparisons between the VGG-16 and ResNet-50 models are displayed in Table 1 below.

Table 1 *Comparison between VGG-16 and ResNet-50 models*

Comparison	VGG-16	ResNet-50
Features	Sequential, simple architecture, uses 3x3 convolutions	Residual learning, deep network, skip connections
Strengths	High feature extraction for small datasets, simple architecture	Residual learning prevents vanishing gradients, efficient for deep networks
Weakness	Computationally heavy, prone to overfitting, lacks flexibility	Requires more memory, not optimal for mobile applications

2.2 Utilization of CNN models

Several CNN models were employed in various field that includes VGG-19, ResNet-101, ResNet-152, MobilenetV2, and Inception ResNet V2. Tyagi et al. compares the efficiency between VGG-16 and VGG-19 models and found out that the VGG-19, with its deeper architecture and additional layers, achieves marginally higher accuracy than VGG-16 in image classification tasks [4]. However, VGG-16 outperforms well in computational efficiency compared to VGG-19, making the former a preferred time-saving options for CNN models. This finding also suggests that deeper CNN models such as VGG-19 may capture more detailed feature representations but at the cost of increased computational costs, a major drawback in resource-constrained environments. Zakaria et al. in their study demonstrates that ResNet-101 and ResNet-152 both achieve higher classification accuracies compared to ResNet-50, owing to the increased depth and ability to capture complex features at the cost of expensive computational costs, which hinders the effectiveness in real-time applications or on devices with limited processing power [5]. Despite this, Suo et al. conducted a study regarding the efficiency of the CS-ResNet-101 and highlighted the model's well performance in medical image analysis tasks, such as diabetic retinopathy detection and tumor classifications, both require subtle pattern recognition [6].

The study suggests that the increased depth of ResNet-101 compared to ResNet-50 enhances the former's ability to detect intricate features, which affects accuracy of diagnosing conditions from medical imagery. A study on MobileNetV2 separately conducted by Wiratama et al. and Bhatta indicated that the detection of diabetic retinopathy utilizing the model provides a CNN evaluation with compact environment compatibility but at the cost of lower model accuracy that affects the medical diagnosis of diabetic retinopathy [7][8]. Varghese and Pandian in their study regarding Inception-ResNet-V2 demonstrated that the merger with the Inception architecture with residual connections, enabling the model to capture both fine and complex patterns in images while reducing issues like vanishing gradients in deep networks but at the cost of computationally expensive and complexity upon implementation, which was proven by a sharp increase of the model accuracy, precision, and the F1 score of 94.74%, 93.99% and 94.11% respectively [9]. Setiawan's study on the VGG-16 model, used Matthew's Correlation Coefficient (MCC), which can effectively capture the predictive quality of these models via image segmentation metrics in skin lesion. It highlights the strength and capabilities of the model in differentiating classes more evenly instead of just capturing majority class predictions [10]. By achieving a balance between sensitivity and specificity across imbalanced classes, MCC ensures that the model is not biased toward any

categories predetermined. The equation for the MCC, which requires the acquisition of true positives, true negatives, false positives, and false negatives is stated in Equation 1.

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (1)$$

Kesuma and Rudiansyah on their study of lung CT-scan image for detection of Covid-19 utilizing ResNet-50 architecture, used Cohen's kappa (κ) as one of the measurements for the accuracy of the model [11]. Cohen's Kappa can help assess the consistency and reliability of these models by adjusting for chance. This is particularly relevant in imbalanced datasets, as Cohen's Kappa reduces the impact of majority class dominance on evaluation scores, apart from providing a measure of agreement strength, supporting confidence that CNN's classifications are consistent with expert human diagnoses. The equation of Cohen's Kappa, which requires the value of the observed (p_o) and expected accuracy (p_e) is as in Equation 2 below.

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (2)$$

The features, strengths and weaknesses between each CNN model are then elaborated in Table 2.

Table 2 *Comparison between CNN models*

Comparison	VGG-19	ResNet-101	ResNet-152	Mobilenet-V2	Inception-ResNet-V2
Features	Extension of VGG-16, but with 19 layers instead of 16	More layers and deeper network with residual connections compared to ResNet-50	Even deeper than both ResNet-50 and ResNet-101 with 152 layers, residual connections	Lightweight, optimized for mobile applications	Hybrid of Inception and ResNet, strong feature extraction
Strengths	Identical to VGG-16 but deeper	Enhanced depth for better accuracy on complex tasks	Very deep network with high accuracy for complex tasks	Lightweight, faster, lower computational cost	Combines strengths of Inception and ResNet for improved accuracy and efficiency
Weakness	Identical to VGG-16, but with higher computational costs	Higher computational and memory requirements than ResNet-50	Requires more memory than ResNet-50 and 101, not optimal for mobile applications	Lower accuracy than deeper models, limited to simpler tasks	Computationally intensive, requires high memory and processing power

2.3 Datasets Used in Diabetic Retinopathy Models

The APTOS 2019 Dataset was obtained and conducted which all images have been rated by clinicians from the scale of 0 to 4, which are scaled according to the five categories of diabetic retinopathy: No DR, Mild, Moderate, Severe, and Proliferate DR respectively. This dataset was used during the APTOS 2019 Blindness Detection, a featured code competition to detect and diagnose diabetic retinopathy, which the solutions were used in the 4th Asia Pacific Tele-Ophthalmology Society (APTOS) Symposium compiled by Dane et al. [12]. Apart from that, another dataset regarding diabetic retinopathy is obtained and compiled by Castillo Benítez et al. from the Department of Ophthalmology of the Facultad de Ciencias Médicas (FCM), Universidad Nacional de Asunción (UNA), Paraguay. The fundus images acquired in the dataset consists of three main categories, which are Non-Proliferative Diabetic Retinopathy (NPDR), Proliferative Diabetic Retinopathy (PDR) and Advanced PDR that were divided into seven subcategories that denotes the severity of diabetic retinopathy, which are No DR signs, Mild (or early) NPDR, Moderate NPDR, Severe NPDR, Very Severe NPDR, PDR and Advanced PDR [13]. Another dataset that is worth investigating is the Indian Diabetic Retinopathy Image Dataset (IDRiD) conducted by Porwal et al.,

which were obtained by capturing fundus images from citizens of rural India and being utilized for medical image classification, fovea detection and optic disc detection [14].

2.4 Summary

Medical diagnostics have been greatly improved by recent developments in convolutional neural networks (CNNs), especially in imaging-based applications including Covid-19 lung imaging, pneumonia, ASD, and diabetic retinopathy (DR). In spite of this, both VGG-16 and ResNet-50 are among the commonly employed model architectures, each offering their own unique advantages. VGG-16, with its simple sequential architecture, excels in small dataset feature extraction at the cost of overfitting issues and high computational costs. The VGG-16 model has been effectively utilized in pneumonia detection and prostate tumor diagnosis with PSPNet+. Meanwhile, ResNet-50, featuring deeper networks and residual connections, addresses vanishing gradients and is well-suited for complex tasks and its application via DeepLabV3+ for precise ASD facial classification. Beyond these, deeper models like VGG-19, ResNet-101, and ResNet-152 offer superior accuracy for complex tasks but at higher computational costs. Lightweight models such as MobileNetV2 provide efficiency for mobile applications but at reduced accuracy, while Inception-ResNet V2 balances feature extraction and computational efficiency Supporting these advancements, datasets like APTOS 2019 Blindness Detection Dataset, UNA Diabetic Retinopathy Dataset, and IDRiD enable robust model training and evaluation for various DR stages, facilitating progress in accurate disease detection and classification. The list of related works is shown in Table 3.

Table 3 *List of literature review matrix*

Author(s)	Aim	Field of study	Model/Techniques/Datasets used	Result
Melinda et al., 2024	To propose a facial feature detection for diagnosis of ASD	Diagnosis of autism spectrum disorder (ASD)	ResNet-50 and DeepLabV3+	Pre-segmentation accuracy of 83.7%, followed by post-segmentation accuracy of 85.9%
Sharma and Guleria, 2023	To conduct a chest x-rays examination using VGG-16 and various Neural Networks	Pneumonia diagnosis	VGG-16, CXR image datasets	Accuracy value of first and second datasets are 92.15% and 95.4%, VGG-16 with neural networks performs well
Ye et al, 2023	To facilitate the segmentation of prostate tumor from MRI images using deep learning models	Diagnosis of prostate tumor	PSP-Net+VGG-16	Prostate MRI classification accuracy and identification rates based on VGG16 are high, particularly when PSP-Net+VGG16 is combined.
Tyagi et al., 2024	To compare the efficiency between VGG-16 and VGG-19	Identification of medicinal plant leaves and disease detection	VGG-16, VGG-19	VGG-16 excels in computational efficiency, while VGG captures detailed representations
Zakaria et al., 2021	To determine the best architecture based on deep residual network (ResNet)	Diagnosis of pulmonary diseases	ResNet-50, ResNet-101 and ResNet-152	The accuracies of ResNet-50, ResNet-101, and ResNet-152 were 90.37%, 89.79%, and 67.57%, respectively.
Suo et al., 2024	To develop an automated DR	Diagnosis and classification of	FCSAM, CS-ResNet-101, APTOS 2019 dataset	Accuracy, specificity, sensitivity and F1

	classification using FCSAM and ResNet-101	diabetic retinopathy		score of the CS-ResNet-101 model are 98.1%, 99.6%, 98.1% and 98.1%,
Wiratama et al., 2023	To propose a deep learning method using MobilenetV2 architecture	Diagnosis and classification of diabetic retinopathy	MobileNetV2, APTOS 2019 dataset	Macro precision, recall, and f1-score were 92.8%, 92.6%, and 92.4%, respectively, with testing accuracy of 92.6%.
Bhatta, 2023	To propose a Mobilenet-Driven Learning for early detection of DR	Diagnosis and classification of diabetic retinopathy	Densenet, Mobilenet, ResNet-50	Training accuracy of 0.92, F1 score of 0.78
Varghese and Pandian, 2023	To present an effective eye disease classification model based on Inception Resnet V2 model with fine-tuning mechanism	Diagnosis of eye disease classification	Inception-ResNet-V2	Model accuracy increases from 81.00% to 94.74%, F1 score increases from 81.49% to 94.11%
Setiawan, 2020	To find the predictive quality of these models via image segmentation metrics in skin lesion	Skin lesion detection	Otsu's thresholding, Coye, and Grabcut image segmentation methods	MCC proven useful in assessment upon obtaining the best image segmentation performance via Grabcut algorithm
Kesuma and Rudiansyah, 2023	To facilitate the classification of Covid-19 diseases using CT lung scans	Classification of Covid-19 Diseases	ResNet-50	Cohen's Kappa value of with training accuracy of

3. Methodology

In this section, the methodology of this article is explained in detail for the article, which compromises the dataset utilized, CNN models used which are VGG-16 and ResNet-50, Hardware and software used.

3.1 Datasets

Retinal fundus images which were used for the evaluation of diabetic retinopathy detection using CNNs were obtained from the APTOS 2019 image dataset which were resized into 224×224 pixels for ease of evaluation. The rationale for the utilization of Rath's dataset is that the 224×224 pixels used were the standard size used for model pretraining, apart from ability of the model to leverage pretrained weights, providing a beneficial starting point for transfer learning, which can lead to faster convergence and improved performance due to reduction of computational costs and elimination of overfitting risks. In this dataset, a total of 3,662 fundus images were pre-defined with degrees of severity on the scale from 0 to 4 with increased accordingly, which are:

- 0 = No_DR
- 1 = Mild
- 2 = Moderate
- 3 = Severe
- 4 = Proliferate_DR

Emerging Advances in Integrated Technology Vol. 6 No. 1 (2025) p. 28-39

The database images were assigned randomly into two batches: 80% of the images were utilized for training, while the remaining 20% were used for testing. The model was compiled using the Adam optimizer and evaluated with sparse categorical cross-entropy and accuracy metrics. The training process employed a batch size of 32 and was executed over 30 epochs. To balance training time and model performance and reduce the possibility of overfitting, 30 epochs with a batch size of 32 were chosen. If there is no improvement after more than 15 checks, the period training is terminated. Early stopping is set at a patience of 15. Fig. 2 shows the distribution of images categorized in accordance with the severity of diabetic retinopathy as tabulated in a chart.

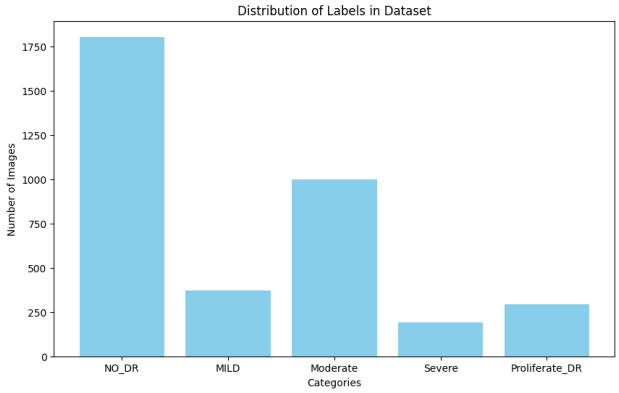


Fig. 2 *Distribution of images by DR stages*

The database images were then pre-processed applying yCbCr color filter, which separates brightness information (Y) from the chroma information (Cb and Cr) due to the reliance of brightness information of computer vision tasks, especially convolutional neural networks. The processed images were then saved into an independent image folder as shown in Fig. 3.

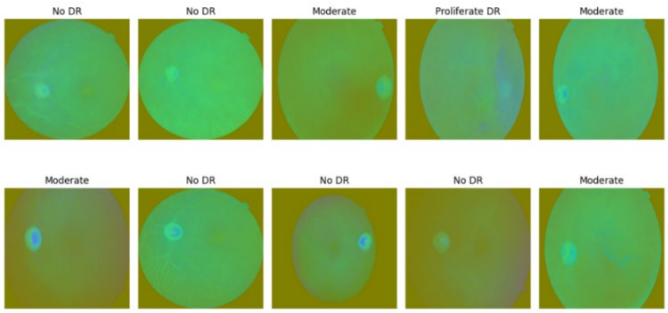


Fig. 3 *Sample of processed images with labelled DR stages*

Correlation among the filename, label and levels in a database is also a key indicator when CNN-related evaluations were conducted, as CNN models learns to associate specific features in the images with their corresponding labels, apart from elimination of bias in the model's training and evaluation as the labels and levels were evaluated based on the categories of degrees of diabetic retinopathy. The correlation among the filepath, label and levels of the dataset are tabulated as shown in Table 4.

Table 4 *Comparison between DeepLabV3+ and PSP-Net+ architectures*

Column	Filepath	Label	Level
Filepath	1.000000	0.028856	-0.029236
Label	0.028856	1.000000	0.177177
Level	-0.029236	0.177177	1.000000

3.2 CNN Models

VGG-16 and ResNet-50, the CNN models used in this study, were selected as the primary focus. One popular convolutional neural network architecture for image classification tasks is the VGG-16 model, which is based on the VGG-Net. Each of the 13 convolutional layers that make up the VGG-16 model's architectural network has two to three convolution blocks, four activation blocks, and five dense blocks. To solve the vanishing gradient issue, the ResNet-50 model's architecture comprises 50 layers with a sequence of residual blocks. The model structure includes 48 convolutional layers, one max-pooling layer, and one average-pooling layer, making it deeper than VGG-16 but also more efficient in training due to the residual learning framework. Fig. 4 shows the model architectures of VGG-16 and ResNet-50 models, based on the output data obtained from the Jupyter Notebook that displays the layers of the model.

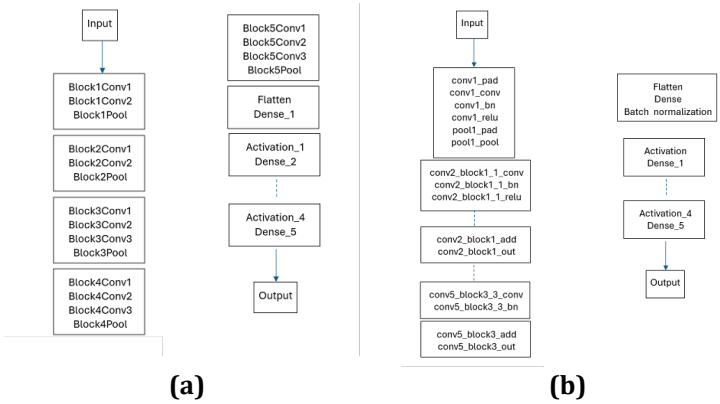


Fig. 4 *Model architectures (a) VGG-16; (b) ResNet-50*

3.3 Hardware and Software used

The next step is the use of the hardware and software for the comparative evaluation. For the hardware used, the laptop is used to train and evaluate the CNN models with the the specifications as follows:

- Intel® Core™ i5-8250U CPU @ 1.60GHz - 1.80 GHz processor
- 12.0 GB RAM
- 512 GB SSD and 1 TB HDD storage
- Intel® UHD Graphics 620 Graphics

The software used is Jupyter Notebook, as the model evaluation mainly utilizes Python programming language that provides a large array of libraries crucial for CNN evaluation tasks, especially TensorFlow, and Matplotlib. The software is used as it provides an interactive environment for an iterative workflow, ideal for analyzing correlations, inspecting Dataframe structures, and testing data transformation.

4. Results

The results for the comparison between VGG-16 and ResNet-50 models are obtained after 30 epochs were processed for each model, which all training accuracy, validation accuracy and test accuracy were done after considered the model loss of both models, which were evaluated using training loss and validation loss that were recorded. During the epoch training conducted, the training data of the VGG-16 and ResNet-50 network conducted is achieved on the training dataset as shown in Table 5.

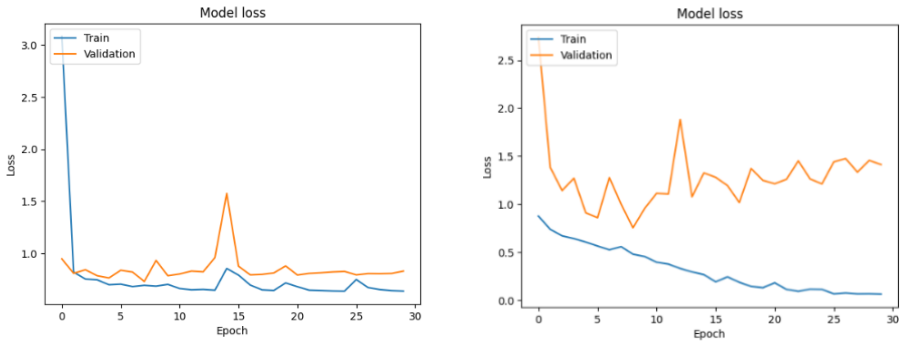
Table 5 *Comparison between findings of VGG-16 and ResNet-50 models*

Metrics	VGG-16 (%)	ResNet-50 (%)
Training Accuracy	76.41	98.67
Validation Accuracy	72.70	76.07
Test Accuracy	72.79	75.45

This indicates that the VGG-16 model correctly classified 76.41% of the training data. The findings in the ResNet-50 show a much higher training accuracy of 98.67%, which suggests that the model in the existing findings was able to fit the training data much better than in the own findings. The VGG-16 model's validation accuracy indicates that 65.57% of the unseen validation data was classified correctly. This relatively low value suggests that

Emerging Advances in Integrated Technology Vol. 6 No. 1 (2025) p. 28-39

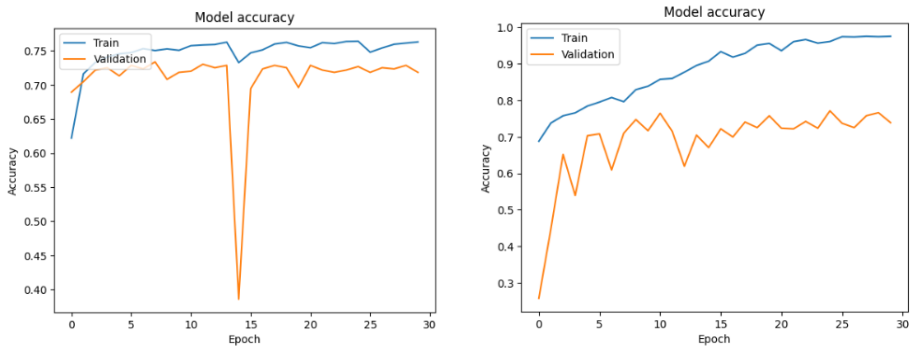
the model may not generalize well to new data, possibly due to underfitting compared to the ResNet-50 model's validation accuracy of 76.07%. In the evaluation of the training, validation, and test accuracies, the model loss is also taken into consideration, which summarizes the. The model loss per epoch for VGG-16 and ResNet-50 model, which records the training and validation losses were shown in a line chart as shown in Fig. 5.



(a) **(b)**

Fig. 5 *Model loss graph (a) VGG-16; (b) ResNet-50*

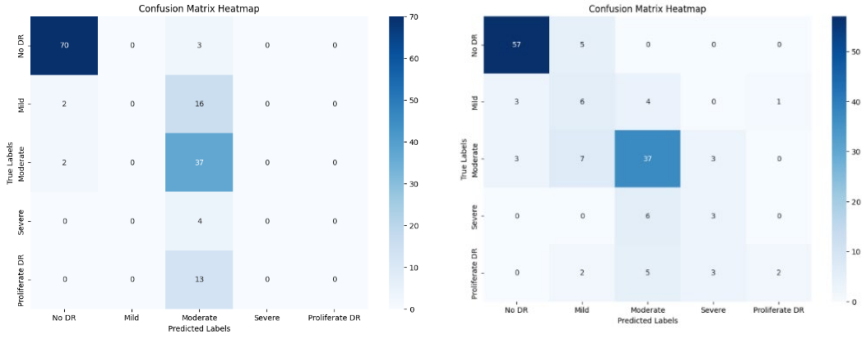
Based on the charts above, the blue line represents the training loss, which the training loss of VGG-16 model experiences decreases sharply in the first epoch and continues to decline, indicating that the VGG-16 model is effectively learning from the training data, while the training loss of ResNet-50 model decreases uniformly throughout the training process, which is an indicator that the ResNet-50 model is learning and fitting the training data well compared to the VGG-16 model. The orange line represents the validation loss, which suggests that the VGG-16 model generalizes well in the early stages based on the initial decrease of the validation loss, while the ResNet-50 model experiences initial decreases on the validation loss but starts to fluctuate and remains relatively high after the first few epochs. The model accuracy per epoch for VGG-16 and ResNet-50 model, which records the training and validation accuracies were shown in a line chart as shown in Fig. 6.



(a) **(b)**

Fig. 6 *Model accuracy graph (a) VGG-16; (b) ResNet-50*

The training accuracy of the VGG-16 model increased from the first epoch which records at 0.62 and remained consistent in between 0.70 until 0.78. The validation accuracy remains consistent with an exceptional outlier on the validation accuracy of the model on the 15th epoch that recorded at 0.3857. This is due to a sharp spike on the validation loss that also occurred simultaneously in the 15th epoch. However, the overall trend is positive, with the best validation accuracy achieved at 30th epoch. Meanwhile, the model training and validation accuracies for the ResNet-50 model shows a balanced yet lower accuracy, but with an outlier on the 14th epoch which shows a sharp decline of the validation accuracy. Meanwhile, the model training and validation accuracies for the ResNet-50 model shows an overall increasing trend per epoch, which gradually raises to a higher accuracy level with the best validation accuracy achieved at 30th epoch. This could imply that the ResNet-50 model either had more complexity, better tuning, or overfitting. The accuracy and loss charts plotted mirrored the actual predicted labels, which are concluded in the confusion matrix heatmap. Fig. 7 shows the confusion matrix heatmap for the VGG-16 model.



(a) **(b)**

Fig. 7 *Confusion matrix (a) VGG-16; (b) ResNet-50*

The confusion matrix of both models indicates that the VGG-16 model correctly classifying images with the "No DR" and "Moderate" label which most predictions are falling on the diagonal, but in most images with "Mild", "Severe" and "Proliferate DR" labels, they were mostly misclassified under the "Moderate" label, further reinforcing the VGG-16 model's bias towards "Moderate" label, apart from a few of the images labelled as "Mild" and "Moderate", both with two images, were misclassified under the "No DR" label. Meanwhile the ResNet-50 shows reasonable performance, with most predictions falling on the diagonal, indicating correct classifications. However, there are some misclassifications, particularly between "Mild", "Moderate", and "Severe" categories. Based on the confusion matrix heatmaps above, both number of true positives, true negatives, false positives and false negatives are compared in Table 6.

Table 6 *Comparison between true positives, false positives, true negatives and false negatives of VGG-16 and ResNet-50 models*

Metrics (%)	VGG-16	ResNet-50
True positives	[70, 0, 37, 0, 0]	[57, 6, 37, 3, 2]
True negatives	[70, 145, 72, 143, 134]	[79, 119, 82, 132, 134]
False positives	[4, 0, 36, 0, 0]	[6, 14, 15, 6, 1]
False negatives	[3, 2, 2, 4, 13]	[5, 8, 13, 6, 10]

The comparison table above highlighted several strengths and weaknesses on both VGG-16 and ResNet-50 models. Compared to VGG-16, ResNet-50 demonstrates broader effectiveness. VGG-16 generally records fewer errors across most classes, except for "Moderate", where ResNet-50 performs better. Both model architectures maintain similar true negative rates, though ResNet-50 slightly excels in some cases, such as "No DR" class. However, ResNet-50 suffers from more false negatives in specific classes, such as "Moderate" and "Proliferate DR" classes, indicating it misses more true instances. Overall, VGG-16 appears more conservative, with fewer false detections, while ResNet-50 offers more balanced detection across classes but at the cost of higher false negatives in some scenarios. Table 7 shows comparison between MCC and Cohen's Kappa of both VGG-16 and ResNet-50 models.

Table 7 *Comparison between the values of MCC and Cohen's Kappa of VGG-16 and ResNet-50 models*

Metrics	VGG-16	ResNet-50
Matthew's Correlation Coefficient (MCC)	0.6028	0.5826
Cohen's Kappa	0.5599	0.5801

After computing the true positives, false positives, true negatives and false negatives of each result, the Matthew's Correlation Coefficient for the VGG-16 model and the ResNet-50 model are 0.6028 and 0.5826 respectively while for the Cohen's Kappa, the values of VGG-16 model and the ResNet-50 model are 0.5599 and 0.5801 respectively. The evaluation presents that the VGG-16 model performs better than ResNet-50 model according to the measurements of Matthew's Correlation Coefficient, but ResNet-50 model outperforms VGG-16

model in accordance with Cohen's Kappa. The varying performance results stemmed from ResNet-50, being a CNN model better suited for complex feature extraction and can capture fine-grained details and patterns more effectively than VGG-16. This depth may contribute to higher overall agreement with true labels, which Cohen's Kappa rewards. VGG-16, on the other hand, has a simpler and straightforward architecture than ResNet-50. While it may not capture intricate features as effectively, it might be better at achieving a balance in correct and incorrect classifications across classes. This balance may lead to a better correlation between predictions and actual classes, which MCC measures well compared to Cohen's Kappa.

5. Conclusion

In conclusion, by conducting a comparative performance of VGG-16 and ResNet-50 based on the calculated metrics of MCC and Cohen's Kappa. By preprocessing retinal images to 224×224 pixels, applying yCbCr color filter and categorizing them by DR severity, this study aims to identify the most efficient model for DR detection. The outcome of this study which VGG-16 has a higher Matthew's Correlation Coefficient (MCC) at 0.6028 and ResNet-50 scores better on Cohen's Kappa at 0.5801 indicates the balance classification accuracy of VGG-16, while the depth of ResNet-50 enhances label agreement with detailed pattern recognition.

Acknowledgement

We sincerely express our acknowledgement and gratitude to our lecturer for the A241 STIN5014 (Artificial Intelligence) course, for their invaluable guidance, unwavering support and insightful feedback. We appreciate the cooperative spirit and helpful contributions of our Universiti Utara Malaysia (UUM) colleagues and fellow researchers. Finally, we would like to express our sincere gratitude to our family and friends for their steadfast support. We are appreciative of your crucial contributions to our academic journey, and this work is the result of everyone's combined efforts.

Conflict of Interest

Authors declare that there is no conflict of interests regarding the publication of the paper.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design:*** *Leon Ewe;* ***data collection:*** *Leon Ewe;* ***analysis and interpretation of results:*** *Leon Ewe Aisha Siddiqa;* ***draft manuscript preparation:*** *Leon Ewe, Aisha Siddiqa. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] Melinda, M., Aqif, H., Junidar, J., Oktiana, M., Basir, N. B., Afdhal, A., & Zainal, Z. (2024). Image Segmentation Performance using Deeplabv3+ with Resnet-50 on Autism Facial Classification. *JURNAL INFOTEL, 16(2),* 441-456.

[2] Sharma, S., & Guleria, K. (2023). A deep learning based model for the detection of pneumonia from chest X-ray images using VGG-16 and neural networks. *Procedia Computer Science, 218,* 357-366.

[3] Ye, L. Y., Miao, X. Y., Cai, W. S., & Xu, W. J. (2022). Medical image diagnosis of prostate tumor based on PSP-Net+ VGG16 deep learning network. *Computer Methods and Programs in Biomedicine, 221.*

[4] Tyagi, K., Vats, S., & Vashisht, V. (2024). Implementing Inception v3, VGG-16 and VGG-19 Architectures of CNN for Medicinal Plant leaves Identification and Disease Detection. *Journal of Electrical Systems*, *20*(7s), 2380-2388.

[5] Zakaria, N., Mohamed, F., Abdelghani, R., & Sundaraj, K. (2021). Three resnet deep learning architectures applied in pulmonary pathologies classification. *2021 International Conference on Artificial Intelligence for Cyber Security Systems and Privacy (AI-CSP)* (pp. 1-8). IEEE.

[6] coal and gangue based on improved PSPNet. *Measurement*, *201*, 111646.

[7] Wiratama, A. B., Fu'adah, Y., Saidah, S., Magdalena, R., Ubaidah, I. D. W. S., & Simanjuntak, R. B. J. (2023, April). Diabetic Retinopathy Classification Based on Fundus Image Using Convolutional Neural Network (CNN) with MobilenetV2. *Proceeding of the 3rd International Conference on Electronics, Biomedical Engineering, and Health Informatics: ICEBEHI 2022, 5–6 October, Surabaya, Indonesia* (pp. 89-102). Singapore: Springer Nature Singapore.

[8] Bhatta, S. (2023). Empowering Rural Healthcare: MobileNet-Driven Deep Learning for Early Diabetic Retinopathy Detection in Nepal. *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, *5*(4), 290-302.

[9] Varghese, R. E., & Pandian, I. A. (2023). Inception-Resnet V2 Based Eye Disease Classification Using Retinal Images. *2023 3rd International Conference on Mobile Networks and Wireless Communications (ICMNWC)* (pp. 1-5). IEEE.
[10] Setiawan, A. W. (2020). Image segmentation metrics in skin lesion: accuracy, sensitivity, specificity, dice coefficient, Jaccard index, and Matthews correlation coefficient. In *2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM)* (pp. 97-102). IEEE.
[11] Kesuma, L. I., & Rudiansyah, R. (2023). Classification of Covid-19 Diseases Through Lung CT-Scan Image Using the ResNet-50 Architecture. *Computer Engineering & Applications Journal*, *12*(1).
[12] Dane, S., Kathrik, & Maggie (2019). *APTOS 2019 Blindness Detection, Version 1.* Asia Pacific Tele-Ophthamology Society. Retrieved 20 October 2024 from https://www.kaggle.com/c/aptos2019-blindness-detection/overview
[13] Castillo Benítez, V. E., Castro Matto, I., Mello Román, J. C., Vázquez Noguera, J. L., García-Torres, M., Ayala, J., Pinto-Roa, D. P., Gardel-Sotomayor, P. E., Facon, J., & Grillo, S. A. (2021). Dataset from fundus images for the study of diabetic retinopathy. Data in Brief, 36, 107068. https://doi.org/10.1016/j.dib.2021.107068
[14] Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., & Meriaudeau, F. (2018). Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research. *Data*, *3*(3), 25.