Taylor & Francis
Taylor & Francis Group

Automatika

Journal for Control, Measurement, Electronics, Computing and Communications

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/taut20

A novel approach to macular edema detection: DeepLabv3+ segmentation and VGG with vision transformer classification

C. Kotteeswari, V. Chandrasekaran & S. Anitha

To cite this article: C. Kotteeswari, V. Chandrasekaran & S. Anitha (2024) A novel approach to macular edema detection: DeepLabv3+ segmentation and VGG with vision transformer classification, Automatika, 65:3, 1177-1190, DOI: 10.1080/00051144.2024.2352313

To link to this article: https://doi.org/10.1080/00051144.2024.2352313

Published online: 13 May 2024.

AUTOMATIKA
2024, VOL. 65, NO. 3, 1177–1190
https://doi.org/10.1080/00051144.2024.2352313

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

A novel approach to macular edema detection: DeepLabv3+ segmentation and VGG with vision transformer classification

C. Kotteeswari[a], V. Chandrasekaran[b] and S. Anitha[c]

[a]Department of Computer Science and Engineering, Velalar College of Engineering and Technology, Erode, India; [b]Department of Electronics and Communication Engineering, VelalarCollege of Engineering and Technology, Erode, India; [c]Department of Information Technology, Kongu Engineering College, Erode, India

ABSTRACT

The domain of deep learning has seen significant advancements, particularly in the context of detecting macular edema from images of the retina, in recent times. This study introduces an innovative model for identifying macular edema, employing two deep learning models: Deeplabv3 + and VGG with a vision transformer. The Deeplabv3 + model is used to segment the macula region in the retinal images. The segmented macula region is then fed into the VGG for feature extraction with a vision transformer model for detection. This approach leverages the strengths of both models in detecting accurately and efficiently. The Deeplabv3 + model can accurately segment the macula region, which is crucial for accurate detection. The VGG combined with a vision transformer model proves highly efficient in detecting even subtle changes in the macular region, signifying the existence of macular edema. The results of our experiments with the dataset show that the proposed method outperforms current cutting-edge techniques. With an outstanding precision rate of 99.53%, the suggested approach firmly solidifies its superiority. The results highlight the effectiveness of the proposed technique in precisely and effectively detecting pathological fluid accumulation in retina images. This ability can have a substantial influence on the early detection and management of eye disorders.

ARTICLE HISTORY
Received 1 November 2023
Accepted 2 May 2024

KEYWORDS
Macular edema; DeepLabv3+; deep learning; VGG; vision transformer; segmentation

1. Introduction

In developed nations, there has been a notable surge in the prevalence of diabetes-associated diseases, largely attributed to modern dietary habits. [1]. Diabetes mellitus, a medical condition characterized by either the impaired utilization of insulin or inadequate insulin production, results in elevated levels of blood sugar, instigating a gradual deterioration in various bodily regions. Among the areas significantly impacted, the delicate vascularity of the retinal and choroidal regions warrants particular attention. As outlined in research conducted by [2], common complications of diabetes include the onset of Diabetic Retinopathy and the condition known as Diabetic Macular Edema (DME). Diabetic Retinopathy is characterized by damage to the blood vessels in the retina caused by diabetes [3]. Conversely, DME involves the accumulation of fluid in the macula, the small central area of the retina. Vision impairment often results from these ailments in individuals affected by different retinal disorders, including diabetic retinopathy and age-related macular degeneration. Early detection and treatment of macular edema are crucial to preventing permanent vision loss.

Accurate identification and division of macular edema hold utmost importance in diagnosing the condition and devising effective treatment strategies. Advancements in AI technology have enabled autonomous analysis of a patient's condition, utilizing significant medical history and associated data, to identify the condition in mere seconds [4]. Lately, there has been an increasing application of computer vision methods to automate the recognition and differentiation of macular edema in retinal images. One such approach involves utilizing segmentation and feature extraction techniques to identify affected regions and differentiate them from healthy areas [5] The accurate detection and segmentation of macular edema remain critical for successful diagnosis and treatment planning. Recent studies have emphasized the use of computer vision techniques to facilitate automated detection and segmentation, with an approach involving the use of segmentation and feature extraction techniques to identify affected regions and distinguish them from healthy areas [6]. Advanced automated methods incorporating segmentation and feature extraction have been developed to detect macular edema. Deep learning techniques, notably using the DeepLabv3 + model, have enabled efficient and accurate categorization of retinal images into healthy and diseased groups. While models like VGG 16 and VGG 19 have demonstrated

CONTACT C. Kotteeswari koteeswaric213@gmail.com Department of Computer Science and Engineering, Velalar College of Engineering and Technology, Erode, India

high accuracy [7], their computational complexity has limited their accessibility. DeepLabv3 + addresses this challenge by employing techniques like atrous convolution to extract comprehensive context and information from images, crucial for tasks related to semantic segmentation.

The study is motivated by the crucial need for early detection and treatment of diabetic macular edema through routine fundus photo screening. To address this, the study intends to improve segmentation accuracy when diagnosing macular edema from fundus images. It aims to improve computer vision applications by optimizing neural network designs, namely VGG models. Combining advanced models like Deeplabv3 + and VGG with a vision transformer presents a viable strategy for achieving higher performance, thereby benefiting early diagnosis and treatment of eye problems.

The main contributions of this research are,

- Integration of EfficientNet with DeepLabv3 + for Parameter Reduction

One significant contribution of this research is the strategic integration of the DeepLabv3 + segmentation approach with the EfficientNet framework. The objective here is to minimize the parameters within the VGG architectures, which have traditionally been resource-intensive. This integration not only streamlines the model but also enhances its efficiency. By harnessing the inherent capabilities of the EfficientNet, the model is poised to achieve a more compact and resource-efficient design. This step is instrumental in addressing the challenges associated with parameter-heavy networks and paves the way for better real-world applications.

- Substitution of Dense Layers with Vision Transformer for Enhanced Performance

Another noteworthy facet of this research is the replacement of dense layers within the VGG network with a vision transformer. This transformation is crucial for two main reasons. Firstly, it seeks to decrease the count of model parameters during the training process, thus alleviating computational and memory requirements. Secondly, the employment of the vision transformer ushers in a novel approach to feature extraction and representation, which, while utilizing a limited number of parameters, brings about a significant improvement in the model's overall performance. This improvement reflects the potential of vision transformers in advancing the capabilities of deep learning models in computer vision applications.

- Comprehensive Evaluation Metrics

To gauge the efficacy of the proposed model, a battery of comprehensive evaluation metrics has been employed. These metrics include specificity, Jaccard index, dice score coefficient, loss, sensitivity, F1 score, positive predictive value, confusion matrix, number of parameters, mean squared error (MSE), ROC (Receiver Operating Characteristic) analysis, training and testing times, accuracy, kappa score, and mean absolute error (MAE). This multi-faceted evaluation approach ensures that the model's performance is assessed from various angles, providing a thorough understanding of its strengths and weaknesses.

Thus, this research aims to enhance deep neural networks such as Deeplabv3 + and VGG efficiency and performance through resource optimization and the integration of vision transformers. By combining these methodologies, the method improves the accuracy and robustness of macular edema detection from retinal pictures, addressing the crucial need for early diagnosis and treatment of eye problems.

The remaining part of the text is organized as follows: Section 2 explores different models used to identify pathological fluid accumulation in fundus images. Section 3 outlines the proposed methodologies. The assessment metrics and outcomes obtained from the selected dataset are discussed in section 4. Lastly, concluding remarks for this research are provided in section 5.

2. Related work

In [8,9]Despite convolutional neural networks (CNN) having segmentation efficiency that is inferior to these enhanced CNNs. The convolutional network's restricted depth, however, continues to present obstacles, including over-segmentation, fault segmentation, and the problem of multi-scale feature extraction. Notably, several studies, including [10–12], have effectively employed CNNs for tasks like subretinal fluid segmentation, pigment epithelium detachment segmentation, and retinal vasculature classification, respectively. To prevent loss of visual acuity, an integrated model was presented by [13]. In another study, [14] introduced a new technique for detecting diabetic macular edema (DME) which involved analyzing colour, wavelet decomposition, and automated lesion segmentation characteristics. In [15] mathematical morphology techniques to create a framework capable of identifying and assessing diabetic maculopathy. Their system relied on the identification and characterization of hard exudates within the macula region, in addition to evaluating the severity of maculopathy.

This framework [16] suggests using federated learning (FL) to forecast Parkinson's disease progression while addressing privacy concerns among health organizations. It focuses on highly interpretable models to enable human-understandable decisions, hence increasing AI trustworthiness. Experimental study shows that FL-based fuzzy rule-based systems are effective at achieving both data privacy and interpretability. The study [17] suggests developing a deep-learning model to predict center-involved diabetic macular edema (ci-DME) using fundus images. This model has a high sensitivity (85%) and a specificity of 80%, exceeding retinal specialists. It can also identify intraretinal and subretinal fluid, indicating promise for broader medical imaging applications.

In [18] a novel semantic categorization approach named Ens4B-UNet, which focuses primarily on medical images. The holistic model effortlessly combines four U-Net structures with previously trained foundational networks, facilitating the production of accurate segmentation results. In their study, [19] developed OCT-DeepLab, a revolutionary DL method specifically engineered for the precise categorization of pathological fluid accumulation in OCT images of the eye's macula. Expanding on the foundation of the DeepLab framework, the team integrated atrous spatial pyramid pooling (ASPP) into their model to facilitate the identification of macular edema across diverse characteristics. Additionally, they incorporated a fully connected CRF to refine the boundaries of the identified pathological fluid, thereby enhancing the accuracy of the segmentation results. The technology employs an approach in image processing, ML, and data analysis to find exudates and produce visual markers indicating the extent of Diabetic Macular Edema (DME). In a study conducted by [20], a comprehensive analysis of various imaging techniques was conducted to assess their merits and limitations in automating the detection and monitoring of DME.

In 2021, [21]. identified the macular region within the provided fundus images, subsequently extracting characteristics through the analysis of textural patterns, edges, and structural attributes. This method was utilized to distinguish between normal and abnormal macula. Lately, the utilization of DL has significantly risen in popularity for examining medical images [21]. CNNs have displayed extraordinary efficacy in a variety of tasks related to the analysis of medical images, such as segmentation, classification, and detection. The study outlined in reference [22] conducted a comprehensive assessment covering both the traditional non-DL approach and the DL approach utilized for the performance of Diabetic eye diseases (Retinopathy and Macular Edema). The review examined several aspects, including datasets, preprocessing, identification and selection of features, and classification techniques used in both non-deep learning and deep learning algorithms for grading DR and DME, as well as the evaluation metrics used to assess their performance.

A novel deep learning technique, as described in [23], combines the advantages of semi-supervised learning and transfer learning methodologies. By amalgamating these approaches, researchers developed a model with an impressive capability to accurately separate the optic nerve head in images of the back of the eye. This advancement highlights the potential of the model for automated segmentation and underscores its effectiveness in this critical task. Notably, this model offers the benefit of minimal storage space requirements and rapid training, setting it apart from other models that achieve comparable performance levels. This study [24] showed that an automated diagnostic system can identify retinal disorders, like diabetic macular edema (DME), at an early stage and can lead to more successful treatment outcomes. They developed a technique using OCT images to automatically detect cystoid areas, which are non-reflective gaps between the vitreoretinal layer and the inner-outer segment (IS-OS) layer, for the detection of cystoid ME (CME). A new technique is proposed in [25] for automatically screening Clinically Significant Macular Edema that addresses two primary difficulties encountered during such screenings – unbalanced data sets and exudate segmentation.

BrainSeg-Net [26] describes an encoder-decoder model for MR brain tumour segmentation that addresses issues such as location information loss and class imbalance. DLS [27] describes a Deep Learning System for Diabetic Macular Edema Detection Using OCT Data that outperforms human experts in specificity and sensitivity. An unsupervised fovea localization method based on the BVV model [28] improves resilience across public datasets. TransDeepLab [29] combines the hierarchical Swin-Transformer and DeepLabv3 to increase medical image segmentation accuracy. HMLC [30] presents a hybrid multilayered classification method for retinal disorders, which achieves good accuracy using CNN-VGG19 models. An end-to-end design [31] integrates ResNet50 and SENet for diabetic macular edema grading, improving accuracy without lesion segmentation. Finally, a deep learning strategy for oral cancer detection [32] uses sensory capabilities, transfer learning, and the Inception-V3 algorithm to obtain high accuracy.

This segment elaborates on the suggested method for identifying macular edema. The initial phase involves an elucidation of the experimental data set employed in the classification of macular edema. Subsequently, we delve into the proposed design based on deep learning and its process for accurately recognizing and categorizing fundus images as either macular or normal.

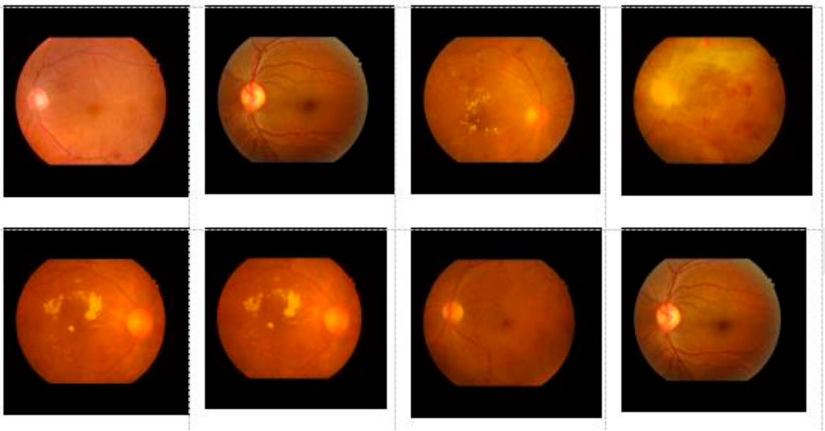


Figure 1. Sample fundus images.

3. Materials and methods

3.1. Dataset

The Indian Diabetic Retinopathy image dataset containing 516 images has been utilized for this experiment and from the total 516 images, the images were split into two sets as 413 images for training the data set, and 103 images are used to test a model. The collection comprises images of diabetic retinopathy (DR) and/or diabetic macular edema (DME), as well as normal retinal structures. Each image includes ground truths about the presence of DR, DME, and normal retinal structures, making supervised learning and model validation easier. To make the dataset bigger, augmentation was done. Figure 1 shows some of the sample back-of-the-eye images from the dataset. The dataset contains two classes namely normal and macular edema.

3.2. Method outline

An outline of the suggested model is provided in this section. This study aims to reduce the parameters in the system for classifying fundus images. Figure 2 shows the outline of the suggested model for macular edema classification.

This study employs advanced machine learning methods to streamline the structure by minimizing the parameters at every stage. Figure 2 depicts the framework of the new combination model designed for categorizing macular edema. The suggested model is divided into four phases. The procedure begins with gathering retinal images, which are later prepared and enhanced with additional data. The data collection has been parted into three subsets: the training, the testing, and the validation. The pre-processed images are segmented using DeepLab v3 + which uses the EfficientNet model as a backbone architecture. Moreover, this DeepLab model incorporates an array of methodologies like dilated convolutions, atrous spatial pyramid pooling (ASPP), and bypass connections to enhance its performance while concurrently minimizing the parameter count. By reducing the parameters in the DeepLab model, this enhancement boosts its performance when tackling image segmentation tasks. Then, segmented output is used as input to VGG models (VGG-16 and VGG-19) for feature extraction. To further decrease the no. of model parameters of the VGG model, dense layers in the model are replaced by Vision Transformer for classification. This model categorizes the output into two classes including normal and macular. Finally, the classifier results are evaluated using various metrics.

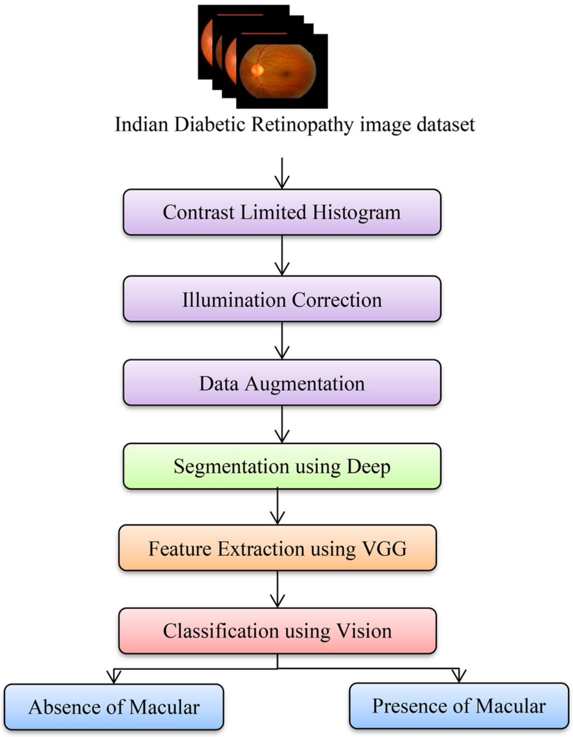

Indian Diabetic Retinopathy image dataset
Contrast Limited Histogram
Illumination Correction
Data Augmentation
Segmentation using Deep
Feature Extraction using VGG
Classification using Vision
Absence of Macular
Presence of Macular

Figure 2. Overview of suggested work.

3.3. Data preprocessing

The primary objective of the preprocessing stage is to eliminate any noise and irregularities present in the retinal fundus image, ultimately enhancing its quality and improving the contrast. Preprocessing plays a crucial role in normalizing the image and correcting non-uniform intensities, in addition to its role in contrast improvement and noise reduction. By eliminating artifacts and enhancing accuracy in subsequent processing stages, preprocessing contributes significantly to reducing errors caused by low-quality images. Consequently, preprocessing is an essential operation for enhancing overall image quality. The outputs obtained from preprocessing serve as the initial input for data training. Incorporating additional preprocessing steps is essential for augmenting the information accessible to the disease diagnosis system.

3.3.1. Contrast enhancement

To develop the visual standard and informative value of the original images before processing, we employed image enhancement techniques, including contrast enhancement and illumination correction. We utilized the Contrast Limited Adaptive Histogram Equalization (CLAHE) technique, developed by [33] to enhance image visibility. CLAHE is a modified version of AHE that applies the enhancement function to all neighbouring pixels and derives a transformation function. It differs from AHE in that it limits contrast. For grayscale retinal images, we employed CLAHE and adjusted the “clip limit” parameter to minimize image noise. By generating a gray-level mapping and histogram clipping, we evenly distributed pixel numbers across gray levels within the contextual area, ensuring a balanced average pixel count.

$$n_{avg} = \frac{n_{r-a_{pix}} * n_{r-b_{pix}}}{n_g} \tag{1}$$

Here, n_{avg}denotes the mean number of pixels, $n_g$represents the count of gray levels within the contextual region, $n_{r-a_{pix}}$ signifies the number of pixels in the directional extent of the contextual region, and $n_{r-b_{pix}}$ indicates the count of pixels in the b direction of the contextual region. The precise clip limit is subsequently computed by,

$$n_{ACL} = n_C * n_{avg} \tag{2}$$

3.3.2. Illumination correction

The objective of this preprocessing technique is to alleviate the impact of uneven illumination in retinal images, which is commonly referred to as the scenario effect, as explained by [34]. To accomplish this, the intensity of each pixel is computed using the following equation:

$$pix' = pix + \mu_{da} - \mu_{la} \tag{3}$$

Where μ_{da} is the desired average intensity, μ_{la} is the local average intensity, and pix, pix” represents the initial and latest pixel size values, respectively.

3.3.3. Data augmentation

Overfitting avoidance is critical for the employed DL models [34,35]. To enhance the dataset, we incorporated various data transformations, including cropping, rotation, and flipping. The implementation of this method led to a quintupling of the image count, as well as resized and improved versions of the original images when compared to the initial dataset. Cropping was employed to eliminate noise and extraneous outliers while focusing on the retinal area. The objective was accomplished by isolating the central patch of the image, encompassing the most crucial section of the retina while disregarding the black contour and irrelevant regions. Rotation involves rotating the cropped images from various angles. Specifically, we rotated the images by 90, 120, 180, and 270 degrees. This rotational augmentation further enriched the dataset and introduced variations in the orientation of the retinal structures. Flipping, on the other hand, entailed horizontally or vertically flipping the images. This transformation was applied to introduce mirror images of the original dataset, providing additional diversity and enabling the model to capture variations in the orientation of retinal features. By employing these data transformations, we were able to expand the dataset and improve its diversity, thereby enhancing the performance and robustness of our model for retinal image analysis tasks.

```
Pseudocode for Data pre-processing
Input: image- fundus image
Output: Augmented image
function CLAHE(image, clip limit) // Contrast Enhancement
    grayscale_image = convert_to_grayscale(image)
    clahe = create_CLAHE(clip Limit = cliplimit)
    enhanced_image = clahe.apply(grayscale_image)
return enhanced_image
function illumination_correction(image, desired_intensity)
// Illumination Correction
    local_average_intensity = calculate_local_average_intensity
    (image)
    correction_factor = desired_intensity / local_average_intensity
    corrected_image = image * correction_factor
return corrected_image
set: augmented_images = []
For image in images do // Cropping
    cropped_image = crop_image(image)
    augmented_images.append(cropped_image)
    For angle in [90, 120, 180, 270] do // Rotation
        rotated_image = rotate_image(image, angle)
        augmented_images.append(rotated_image)
    End For
        horizontally_flipped_image = flip_image_horizontally
        (image)
        vertically_flipped_image = flip_image_vertically(image)
        augmented_images.extend([horizontally_flipped_image,
        vertically_flipped_image])
return augmented_images
```

3.4. Segmentation using DeepLabv3+

Segmentation is critical since it enables the accurate identification and separation of certain regions within

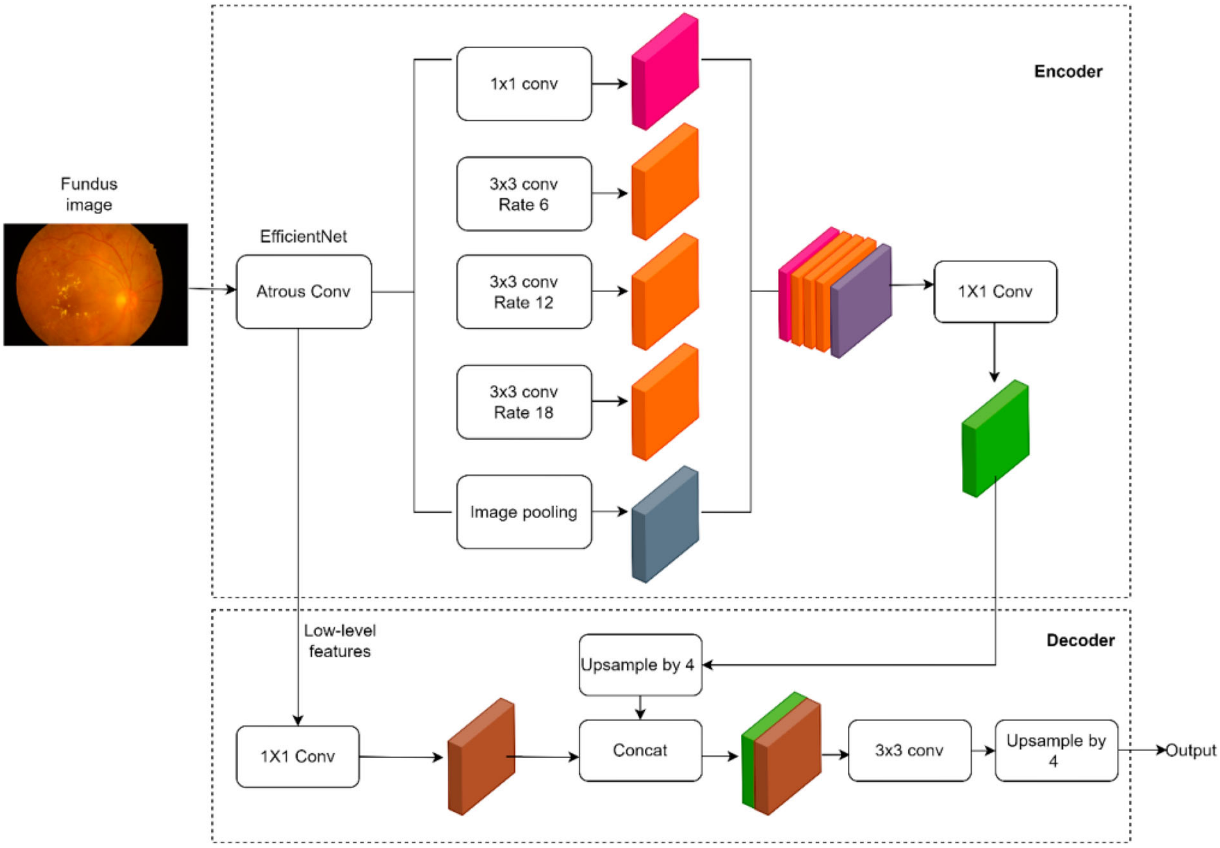


Figure 3. Segmentation using Deeplabv3+.

complicated images such as retinal scans, which is useful in diagnosing macular edema. This distinction allows for targeted investigation and action, resulting in more accurate diagnosis and individualized therapy planning. Without segmentation, collecting useful information from images would be difficult, impeding effective healthcare decision-making and possibly leading to misdiagnosis or inadequate treatment.

DeepLabV3 + utilizes an advanced deep neural network architecture that includes an encoder-decoder framework and a spatial pyramid pooling module, as shown in Figure 3. By incorporating depthwise separable convolution into the ASPP and decoder modules and exploring the backbone feature extraction network, the network enhances both its speed and effectiveness for semantic segmentation tasks. An encoder integrates a feature extraction network with an ASPP module, and the decoder improves both the network's efficiency and semantic segmentation issues, delivering precise outcomes for semantic segmentation.

The encoder component of the system employs atrous convolution to capture background data at various balances, while the decoder component efficiently enhances object boundaries and segmentation outcomes. By manipulating filters, atrous convolution permits the network to control feature resolution with CNN and obtain a range of multi-scale information. When applied to a two-dimensional input feature map s, atrous convolution computes an output feature map z using convolution filter f, as shown below,

$$z[j] = \sum_{t} s[j + r.t]f[t] \tag{4}$$

The equation above represents the output feature map z computed from the input feature map s using atrous convolution with a rate of r, which determines the stride for sampling the input image. The readers can refer to [8] for more details on the equation. Standard convolution operates with a fixed rate of r equal to 1. However, to decrease computational complexity, depthwise separable convolution is employed. This method employs a combination of spatial convolution and point-wise convolution to efficiently process individual input channels. In the DeepLab model, an energy function, originally introduced by [8] is utilized.

$$E = \sum_{m} \theta_m(l_m) + \sum_{n} \theta_{mn}(l_m, l_n) \tag{5}$$

θm(lm) In the equation mentioned above, the label assignment for each pixel is denoted by the variables l, m & n where m varies from 1 to N. The unary function θm(lm) is used to represent the value associated with the label assignment for pixel l in label m.

$$\theta_m(l_m) = -\log P(l_m) \tag{6}$$

The assignment of label probability at pixel m, denoted as P(lm), is calculated using a specific formula. To compute the probabilities for all connecting pairs of image pixels, m, and n, the following expression is utilized:

$$\theta_{mn}(l_m, l_n) = \mu(l_m, l_n)[g_1 exp \times \left(-\frac{||P_m - P_n||^2}{2\sigma_\alpha^2} - \frac{||I_m - I_n||^2}{2\sigma_\beta^2}\right) + g_2 exp\left(-\frac{||P_m - P_n||^2}{2\sigma_\gamma^2}\right) \quad (7)$$

In the equation above, the value of $\mu(l_m, l_n)$ is equal to 1 when $l_m = l_n$ and 0 otherwise.

3.5. VGG-16 and VGG-19 for feature extraction

In image classification, the extraction of features greatly impacts the accuracy of classification tasks. Features are categorized as local or global, depending on their characteristics such as colour, shape, or texture. Deep convolutional neural network models, like VGGNet, have become popular for this purpose. VGGNet, developed jointly by the visual geometry group at Oxford University and researchers from Google DeepMind [9], is known for its straightforward architecture and strong performance. VGG16 and VGG19 are constructed using 3 × 3 convolutional kernels and 2 × 2 maximum pooling layers. The utilization of pre-trained DNNs involves extracting profound image features. VGG16, for example, is pre-trained on extensive datasets like ImageNet, which serves to decrease the time and computational resources required for training. The framework consists of 13 layers for convolution and 3 layers that are fully connected. It is structured into five segments, each of which incorporates multiple convolutional layers and one pooling layer. The number of kernels used for feature extraction in the first block's two convolutional layers is 16, and the subsequent pooling layer reduces the image size. The remaining blocks follow a similar architecture, with the exception that blocks 1 and 2 have two convolutional layers, while blocks 3–5 have three convolutional layers with varying kernel numbers in each layer to increase network depth and improve accuracy. The structure of VGG16 is illustrated in Figure 4.

The VGG19 model, a popular method for image classification, utilizes multiple 3 × 3 filters in each of its 16 convolutional layers for feature extraction. Its architecture, as shown in Figure 4, includes 5 groups of these layers, each by a max-pooling layer. After these layers, the model employs a classifier for classification tasks. When given an image, the model outputs the corresponding label for the depicted object. In this research, we utilize a pre-trained VGG19 model for attribute selection and implement a deep-learning approach for classification. However, due to the high number of parameters computed by the CNN model, we perform dimensionality reduction by applying a linear embedding layer and subsequently a classification method.

3.5.1. Convolutional layer

The convolution operation plays an important part in extracting image features. The process entails performing convolutions on the attribute maps from the previous layers using the resulting feature maps and simultaneously refining the convolutional kernels. This process, known as training, can be mathematically expressed as follows:

$$Q_v^{(n)} = \sum_u F_u^{(n-1)}(r, s) * k_{uv}^{(n)}(r, s) + b_v^{(n)} \quad (8)$$

$$Q_v^{(n)} = \sum_u \sum_{p,q=0}^{C} F_v^{(n-1)}(p, q) k_{uv}^{(n)}(r - p, s - q) + b_v^{(n)} \quad (9)$$

Let (r, s) be a pixel coordinate. $F_u^{(n-1)}$ represents the u-th feature map of the (n-1)-th layer, while $k_{uv}^{(n)}$denotes the convolutional kernel that links the u-th input feature map to the v-th output feature map on the n-th layer. C represents the size of the convolutional kernel, and $b_v^{(n)}$denotes the v-th bias of the n-th layer. The symbol $*$ indicates the 2-D convolutional operation. To enhance the nonlinear characteristics of the network and strengthen its ability to express classifications, a nonlinear activation function is linked to each convolutional layer. This can be expressed as follows:

$$F_v^{(n)}(r, s) = \sigma(z_v^{(n)}) \quad (10)$$

The symbolσrepresents the non-linear activation function known as ReLU. Within the VGG neural network, pooling is utilized to decrease the count of training parameters. Normally, a 2 × 2 pooled window size is used, merging the values of four pixels into one result. Maximum pooling selects the highest value from the four pixels, whereas average pooling computes the mean value. In the VGG neural network, the employed pooling technique is maximum pooling.

$$F_v^{(n)}(r, s) = \underset{p,q=0...W-1}{max} F_v^{(n)}(r + p, s + q) \quad (11)$$

Here W denotes the dimensions of the pooling window.

3.5.2. Loss function

Upon completion of forward propagation in a neural network, updating network parameters involves following specific rules determined by loss functions like cross-entropy loss or MSE. The training aim is to lower the loss value, leading to network optimization. Cross-entropy loss is particularly notable for its capacity to gauge the similarity between training samples and the

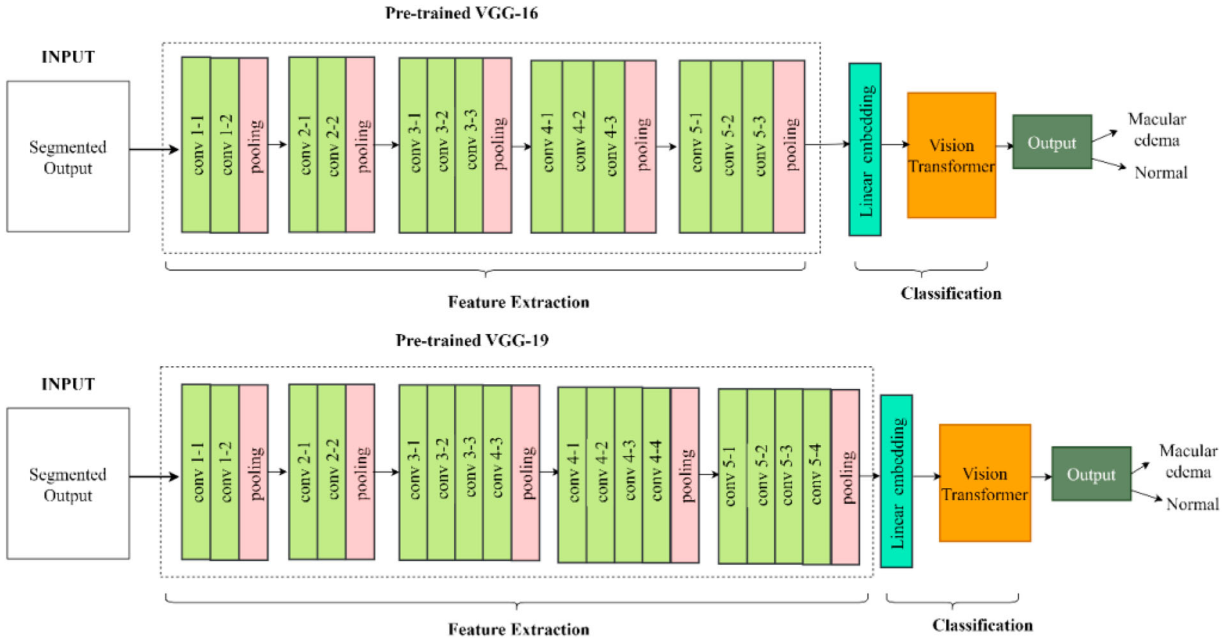


Figure 4. Architecture of VGG-16 and VGG-19 for feature extraction.

model distribution, offering a more accurate measure of the model's fit to the data. This notion can be articulated as follows:

$$L(w, b) = -\sum_{g=1}^{E} y^{(g)} \log p(y_g|q^{(L)}; w, b) \qquad (12)$$

In a neural network, the weight and bias sets for each layer are typically labelled as "w" and "b" respectively. Additionally, the actual label for the g-th class is denoted by "$y^{(g)}$".

3.6. Classification using vision transformer

3.6.1. Linear embedding layer

Initially, the patches undergo linear projection using the embedding matrix E, resulting in a model dimension vector d. This vector then undergoes further processing in the encoder. The embedded representations, along with a learnable classification token v class, play a vital role in the classification task. Despite the Transformer model treating the embedded image patches as an unordered set, positional encodings are introduced to preserve spatial layout information. These positional encodings (Ip) are added to the input, maintaining spatial relationships between patches. This process generates the embedded sequence of patches represented as Equation (13), with each patch accompanied by the token 0.

$$y_0 = [u_{class}; a_1 I; a_2 I; \ldots; a_n I] + I_q,$$
$$I \in T^{(q^2 e)\times f}, I_p \in T^{(n+1)\times f} \qquad (13)$$

3.6.2. Vision transformer encoder

Before being inputted into the Transformer encoder, the sequence of patches is embedded. The Transformer encoder consists of N layers, each comprising two key components: (1) a multihead self-attention block based on Equation (14), and (2) a fully connected feedforward dense block as per Equation (15). The MLP block comprises two dense layers, separated by a GeLU activation function. Both parts of the encoder utilize residual skip connections and are preceded by a layer of normalization. The input for this process is the embedded patches sequence, designated asy_0.

$$y'_n = MSA(NL(y_{n-1})) + y_{n-1},\ n = 1 \ldots ..N \qquad (14)$$

$$y_n = MLP(NL(y'_n)) + y'_n,\ n = 1 \ldots ..N \qquad (15)$$

The initial element y_N^0 from the sequence is captured within the last encoding layer and utilized as input for an external head classifier. This classifier is responsible for predicting the corresponding class label.

$$z = NL(y_N^0) \qquad (16)$$

In the Transformer framework, the MSA block assumes a vital role within the Transformer's encoder by assessing the relative importance of each patch embedding in the sequence. Comprising four layers – linear, concatenation, self-attention, and a final merging layer – the MSA block captures the relationships between patches, facilitating effective information processing and representation within the model. Its significance within the Transformer architecture is underscored by its pivotal role.

The self-attention (SA) mechanism within the Transformer employs attention weights to assess the significance of different elements in a sequence. These weights are determined by calculating the dot-product of the query (q) and key (k) vectors, adjusted by a scaling factor that is influenced by the dimension of the key vector (DK). By applying a softmax function to the scaled dot-product, the attention weights are generated and subsequently used to calculate a weighted sum of the value (v) vectors across the sequence. The SA block generates q, k, and v vectors by multiplying the input sequence with learned matrices,U_{qkv}. Each element in the sequence receives a Q vector, enabling the evaluation of relevance between elements. Finally, the softmax output is multiplied by the v vectors to produce the final attention-weighted representation of the input sequence. The equations for the SA block are given as follows:

$$[q, k, v] = zU_{qkv}, U_{qkv} \in R^{d \times 3D_k} \tag{17}$$

$$A = softmax\left(\frac{qk^T}{\sqrt{D_k}}\right), \; A \in R^{n \times n} \tag{18}$$

$$SA(x) = A.v \tag{19}$$

The MSA block employs h attention heads to compute scaled dot-product attention in parallel, with each head using distinct learned Query, Key, and Value weight matrices. The resulting attention outputs from all heads are linked and then linearly projected to the desired dimension through a feed-forward layer parameterized by a learnable weight matrix W. The output of the MSA block is thus the concatenation of the h attention outputs transformed by the feed-forward layer. This operation is given by the equation:

$$MSA(z) = Concat(SA_1(z); SA_2(z); \ldots SA_h(z))W, \quad W \in R^{h.D_k \times D} \tag{20}$$

4. Result and discussion

4.1. Experimental design

The suggested models are coded in the Python programming language utilizing the Keras module, which is based on machine learning. Python is compatible with TensorFlow and is ideal for developing a neural network. This is beneficial for both CPU and GPU operations. To adjust the model's hyperparameters precisely, a grid search is used to identify the parameters that produce the best performance on the given test data. To ensure efficient model training, it is vital to choose the most appropriate hyperparameters. In this instance, the number of training epochs has been fixed at 50, and the learning rate is established at 0.00001.

4.2. Performance evaluation

To assess the efficiency of the suggested neural network classifier contrasted to the current classifiers, performance metrics were utilized. These metrics include the dice score coefficient (DSC) as described by [27], the Jaccard index (J) as described by [26], sensitivity, accuracy, specificity, and F1-score measurements. Furthermore, to assess the classification outcomes against randomly assigned values, [25] introduced the Kappa coefficient. A greater Kappa coefficient signifies a higher level of precision in the classification process.

Table 1. Results of classification evaluation using the confusion matrix parameters and three distinct learning rates.

	Learning rate	0.00001	0.0001	0.001
VGG-19	Accuracy	99.67	99.53	99.00
	sensitivity	99.65	99.25	98.38
	specificity	99.87	99.67	99.54
	positive predictive value	99.35	99.27	98.67
	F1 score	99.52	99.32	98.07
VGG-16	Accuracy	98.53	98.47	98.17
	sensitivity	98.29	99.15	98.38
	specificity	98.47	99.53	99.37
	positive predictive value	98.25	99.23	98.18
	F1 score	98.47	99.29	98.00

Cohen's kappa (K) has been calculated using the Equation (21).

$$K = \frac{accuracy_{(predicted)} - accuracy_{(expected)}}{accuracy_{(expected)}} \tag{21}$$

Evaluation of image segmentation performance can be assessed through various criteria. Nevertheless, the Mean Intersection-Over-Union (MeanIoU) emerges as the predominant and precise evaluation metric overall. This metric indicates the point of convergence between the forecasted values of the approach and the actual numbers of the sample labels. The union ratio is computed by summing the average of the intersections for each class. It can be expressed mathematically as follows:

$$MeanIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ij} - p_{ii}} \tag{22}$$

The value of positive prediction pertains to the likelihood that a particular set of pixels has been accurately recognized, as denoted by the true positive (TP) value.

$$Positivepredictivevalue = \frac{TP}{TP + FP} \tag{23}$$

4.3. Classification performance using learning rate

To explore the suitable training parameters, we proceeded with the fine-tuning of the initial learning rate (LR) value and subsequently assessed the resultant performance. Table 1 presents the performance of each LR value.

The obtained results showed better performance for the VGG-19 model at different learning rates (0.00001, 0.0001, 0.001) indicating high accuracy rates ranging from 99.00% to 99.67%. The sensitivity values

Table 2. Performance comparison between non-segmented and segmented images.

	Parameter	Segmented imaged	Non-segmentation images
VGG-19	Accuracy (%)	99.53	99.47
	Sensitivity (%)	99.37	99.28
	Specificity (%)	99.68	99.62
	positive predictive value (%)	99.27	99.12
	F1 score (%)	99.26	99.18
	Training time (second)	2.68 × 105	2.75 × 105
	Testing time (second)	28.24	33.14
VGG-16	Accuracy (%)	98.34	98.23
	Sensitivity (%)	98.78	98.65
	Specificity (%)	98.63	99.58
	positive predictive value (%)	98.45	98.32
	F1 score (%)	98.40	98.27
	Training time (second)	2.72 × 105	2.85 × 105
	Testing time (second)	30.47	37.48

also remain consistently high, ranging from 98.38% to 99.65%, which demonstrates the approach's capability to correctly find positive instances. Similarly, the specificity values show excellent performance, ranging from 99.54% to 99.87%, reflecting the approach's proficiency in accurately finding negative instances. The positive predictive values range from 98.67% to 99.35%, indicating the model's reliability in correctly predicting positive instances. The F1 scores range from 98.07% to 99.52%, representing the balance between precision and recall.

4.4. Comparison of time for training and testing between segmented and non-segmented images

This research examines the effectiveness of classifying macular edema disease and the time taken for analysis using non-segmented and segmented images. Findings in Table 2 reveal that employing DeepLabv3 + to study two sets of fundus disease images, one being non-segmented and the other segmented, results in superior performance with the segmented approach. Specifically, the classification accuracy for VGG-19 and VGG-16 using segmented fundus images stands at 99.53% and 98.34%, respectively. Notably, the training and testing times for the macular edema disease classifier are significantly faster when using segmented images. For example, the VGG-19 network requires 2.68 × 105 s for training and 28.24 s for testing, compared to the non-segmented images with training times of 2.75 × 105 s and testing times of 33.14 s for the same network. Moreover, the adoption of the EfficientNet backbone in place of the DeepLabv3 + encoder enhances the model's ability to efficiently extract high-level features from input images, leading to quicker training and inference times without compromising accuracy. The deeper architecture of VGG-19 with 19 layers allows it to effectively capture intricate patterns, thereby contributing to its superior performance in feature extraction compared to the 16-layer VGG-16.

Table 3. Comparison of number of Parameters of VGG-16 and VGG-19.

Model	Parameters (M)
VGG-16	14.71
VGG-19	20.02

4.5. Comparison of the number of parameters of VGG-16 and VGG-19

The parameter count in VGG denotes the total learnable parameters within the neural network architecture. A high parameter count can lead to overfitting and hinder the model's generalization capability. Additionally, it can be resource-intensive in terms of computational power and memory during both training and usage. VGG-16 and VGG-19 are known for their substantial parameter counts, with VGG-16 having 138 million and VGG-19 having 143 million parameters. The fully connected layers contribute significantly to this count. Notably, the fully connected layers in VGG-16 consist of three layers with 4096 neurons each, totalling 37,752,832 parameters, whereas in VGG-19, these layers have 62,378,344 parameters. The removal of these layers from each model effectively reduces the overall parameter count, as illustrated in Table 3.

4.6. Error rate

The primary utilization of the MSE and MAE lies in evaluating prediction error rates and model performance. MAE gauges the discrepancy between the actual and predicted values through the computation of the medium absolute difference across the dataset. The Mean Squared Error, which measures the disparity between the actual and forecasted values, is calculated by taking the average disparity from the dataset and then squaring it.

$$MAE = \sum_{i=1}^{n} \frac{|y_i - \widehat{y_i}|}{n} \tag{30}$$

$$MSE = \sum_{i=1}^{n} \frac{(y_i - \widehat{y_i})^2}{n} \tag{31}$$

Where y_i and $\widehat{y_i}$ represents observed value and predicted value respectively.

The MSE and MAE curves of VGG-16 and VGG-19 are depicted in Figures 5 and 6.

In general, a lower Mean Absolute Error (MAE) or MSE score signifies improved model performance. VGG-19, known for its ability to extract high-level image features, contributes to enhanced accuracy and reduced error rates. More data is required for training and validation compared to other models, which aids in boosting performance. Through this extensive data, the model can grasp intricate data patterns, resulting in precise predictions and reduced errors. The model has

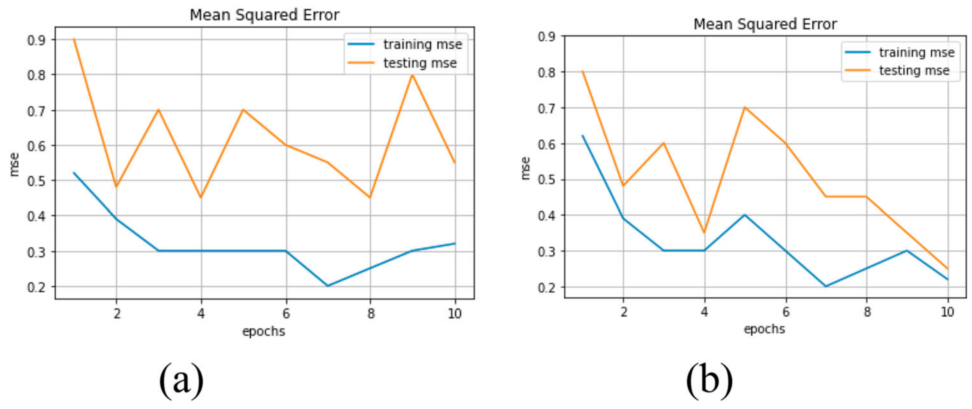


Figure 5. MSE curve of (a) VGG-19 and (b) VGG-16.

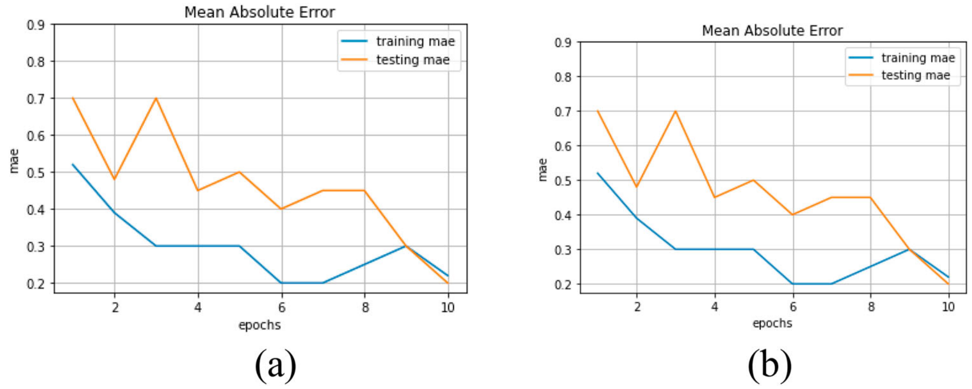


Figure 6. MAE curve of (a) VGG-19 and (b) VGG-16.

been trained using regularization techniques such as dropout and weight decay to counter overfitting, leading to better generalization. Fine-tuning the model's hyperparameters can further diminish MAE and MSE errors. Model performance is influenced by factors like the training data's quality and size, specific hyperparameter selection during training, and the model's architecture.

4.7. *Training loss and validation loss*

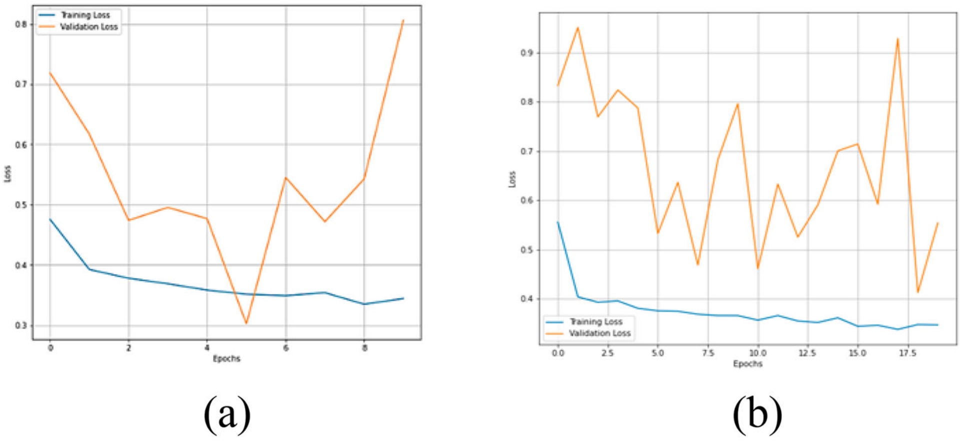
A crucial element within neural networks involves the loss function, which is accountable for evaluating the accuracy of the method's predictions. During training, the model is taught using training data. Additionally, the assessment of the performance of a DL approach on the evaluation set for classifying brain images involves utilizing the validation loss.

When the dense layers are excluded from both VGG models, the model is effectively shortened at the final convolutional layer, reducing the count of trainable parameters. This reduction can lead to a decrease in the model's overall intricacy, potentially making it easier to train and resulting in a lower loss. Figure 7 displays the approach loss on the training and validation datasets for the suggested VGG-16 and VGG-19, respectively. Additionally, the convolutional layers in the VGG model are designed to learn hierarchical representations of features in the input images, and these features may already be sufficiently informative for the classification task without the need for the fully connected layers.

Table 4. Performance of the segmentation algorithm.

Segmentation algorithms	DSC	J	Kappa	MeanIoU
DeepLabv3+	98.54	97.46	96.36	95.93
ResNet-50	94.97	95.37	94.93	94.28
MobileNetV2	96.66	93.78	95.48	95.37
Inception v3	93.57	92.39	93.34	94.48
ResNext50	95.54	94.57	94.76	94.75

In Table 4, the segmentation results of five algorithms (DeepLabv3+, ResNet-50, MobileNetV2, Inception v3, and ResNext50) are compared. DeepLabv3 + stands out with the highest DSC, J, and Kappa scores, indicating superior segmentation accuracy. When combined with EfficientNet as its backbone, DeepLabv3 + benefits from advanced feature extraction capabilities, leading to improved performance. EfficientNet's high resolution and depth enable it to capture intricate image details, essential for precise object localization. Despite DeepLabv3+'s dominance, ResNet-50 and MobileNetV2 also exhibit strong performance. In contrast, Inception v3 shows relatively lower performance, but optimization could enhance its effectiveness. DeepLabv3 + and MobileNetV2 emerge as the most effective algorithms for segmentation tasks, with other options like ResNet-50, Inception v3, and ResNext50 also showing promise, albeit with slightly lower performance. The notable performance



(a) (b)

Figure 7. The loss during the training and validation stages for VGG-16 and VGG-19.

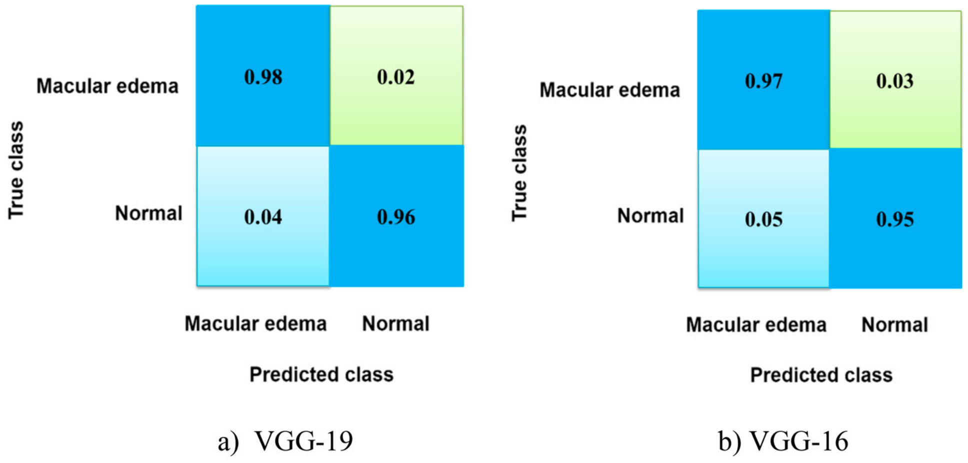


Figure 8. Confusion matrix. a) VGG-19; b) VGG-16.

of DeepLabv3 + suggests its potential applicability in domains like medical imaging or autonomous driving where high accuracy is critical.

4.8. *Confusion matrix*

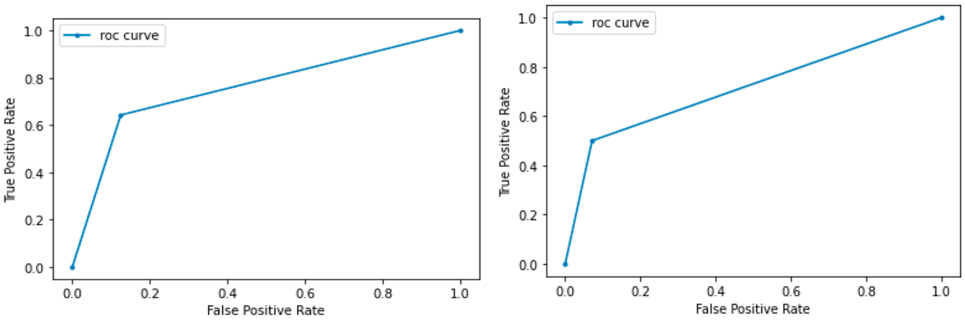
The test information is used for constructing a confusion matrix, which assesses the effectiveness of the suggested technique. The rows of this matrix contain the actual class information, while the columns contain the predicted class information. This matrix produces four possible scenarios: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Figure 8 demonstrates the confusion matrix for the suggested model, illustrating the findings of an analysis on a dataset involving the classes “normal” and “macular edema.” The model demonstrates a high specificity of 98% by correctly predicting the data.

4.9. *Receiver operator characteristic(ROC)*

The ROC curve represents a significant measurement in tasks related to classification and identification. It is created by plotting the rate of true positive (TP) results in comparison to the rate of false positive (FP) results. In the ninth figure, it was observed that the ROC curve was near the upper left corner, implying the precise classification of macular edema and normal categories from fundus images through the suggested VGG-19 approach. Notably, in Figure 9(a), the VGG-19 model surpassed the VGG-16 model in Figure 9(b) as it was positioned nearer to the left corner, signifying a higher true positive rate and enhanced accuracy of the approach. With its increased layers, VGG-19 demonstrates adeptness in capturing intricate features within the input images, thereby leading to improved classification performance.

5. Conclusion

The suggested method involving the utilization of Deeplabv3+-based segmentation and VGG combined with a vision transformer for identifying macular edema has demonstrated encouraging outcomes. This amalgamation of deep learning models effectively tackles the challenges of precisely and swiftly detecting macular edema in retinal images. The Deeplabv3 + model’s segmentation of the macula region is crucial for pinpointing the area of interest in macular edema detection. Simultaneously, the VGG integrated with the vision transformer model is adept at identifying subtle



(a) (b)

Figure 9. (a) ROC Curve of VGG-19 and (b) ROC curve of VGG-16.

changes within the macula region that could potentially signify the presence of macular edema. The integration of these models results in a heightened accuracy and sensitivity in identifying macular edema. Implementing this proposed method in clinical practice has the potential to enable early identification and treatment of macular edema, consequently preventing irreversible harm to the retina and preserving vision. Furthermore, it can ease the burden on clinicians, streamlining the process of macular edema detection and enhancing the accuracy of diagnosis. Future advancements may include the development of a user-friendly interface tailored for clinical usage and the validation of this approach on a more expansive and diverse dataset, ensuring its resilience and applicability across various scenarios.

Authorship contributions

All authors are contributed equally to this work

Disclosure statement

No potential conflict of interest was reported by the author(s).

Human and animal rights

No violation of Human and Animal Rights is involved.

Data availability statement

Data sharing is applicable to this article as a publicly available dataset analyzed during the current study

References

[1] Lovic D, Piperidou A, Zografou I, et al. The growing epidemic of diabetes mellitus. Curr Vasc Pharmacol 2020;18(2):104–109.

[2] Lahmiri S. Hybrid deep learning convolutional neural networks and optimal nonlinear support vector machine to detect presence of hemorrhage in retina. Biomed Signal Process Control. 2020;60(101978):101978.

[3] Foo VHX, Gupta P, Nguyen QD, et al. Decrease in choroidal vascularity index of Haller's layer in diabetic eyes precedes retinopathy. BMJ Open Diabetes Res Care. 2020;8(1):e001295.

[4] Wang Z, Keane PA, Chiang M, et al. Artificial intelligence and deep learning in ophthalmology. In: *Artificial intelligence in medicine*. Cham: Springer International Publishing; 2021. p. 1–34.

[5] Bhardwaj C, Jain S, Sood M. Deep learning-based diabetic retinopathy severity grading system employing quadrant ensemble model. J Digit Imaging. 2021;34(2): 440–457.

[6] Li F, Wang Y, Xu T, et al. Deep learning-based automated detection for diabetic retinopathy and diabetic macular oedema in retinal fundus photographs. Eye. 2022;36(7):1433–1441.

[7] Sitaula C, Hossain MB. Attention-based VGG-16 model for COVID-19 chest X-ray image classification. Appl Intell. 2021;51(5):2850–2863.

[8] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv [cs.CV]. 2014.

[9] Chen L-C, Papandreou G, Kokkinos I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell 2018;40(4):834–848.

[10] Hu J, Chen Y, Yi Z. Automated segmentation of macular edema in OCT using deep neural networks. Med Image Anal 2019;55:216–227.

[11] Lu D, Heisler M, Lee S, et al. Deep-learning based multiclass retinal fluid segmentation and detection in optical coherence tomography images using a fully convolutional neural network. Med Image Anal 2019;54:100–110.

[12] Soomro TA, Afifi AJ, Zheng L, et al. Deep learning models for retinal blood vessels segmentation: A review. IEEE Access. 2019;7:71696–71717.

[13] Thulkar D, Daruwala R, Sardar N. An integrated system for detection exudates and severity quantification for diabetic macular edema. J Med Biol Eng 2020;40(6):798–820.

[14] Maurya PK, Gupta V, Singh M, et al. Automated detection of diabetic macular edema involving cystoids and serous retinal detachment. Opt Laser Technol 2020;127(106157):106157.

[15] Rajput GG, Reshmi BM, Rajesh IS. Automatic detection and grading of diabetic maculopathy using fundus images. Procedia Comput Sci. 2020;167:57–66.

[16] Shahid AH, Singh MP. A deep learning approach for prediction of Parkinson's disease progression. Biomed Eng Lett 2020;10(2):227–239.

[17] Varadarajan AV, Bavishi P, Ruamviboonsuk P, et al. Predicting optical coherence tomography-derived diabetic macular edema grades from fundus photographs using deep learning. Nat Commun. 2020;11(1):130.

[18] Abedalla A, Abdullah M, Al-Ayyoub M, et al. Chest X-ray pneumothorax segmentation using U-Net with EfficientNet and ResNet architectures. PeerJ Comput Sci. 2021;7(e607):e607.

[19] Wang Z, Zhong Y, Yao M, et al. Automated segmentation of macular edema for the diagnosis of ocular disease using deep learning method. Sci Rep 2021;11(1):13392.

[20] Ajaz A, Kumar H, Kumar D. A review of methods for automatic detection of macular edema. Biomed Signal Process Control. 2021;69(102858):102858.

[21] Khalid S, Akram MU, Shehryar T, et al. Automated diagnosis system for age-related macular degeneration using hybrid features set from fundus images. Int J Imaging Syst Technol 2021;31(1):236–252.

[22] Mathews MR, Anzar SM. A comprehensive review on automated systems for severity grading of diabetic retinopathy and macular edema. Int J Imaging Syst Technol 2021;31(4):2093–2122.

[23] Bengani S, Jothi AA, Vadivel S. Automatic segmentation of optic disc in retinal fundus images using semi-supervised deep learning. Multimed Tools Appl. 2021;80(3):3443–3468.

[24] Tao Z, Zhang W, Yao M, et al. A joint model for macular edema analysis in optical coherence tomography images based on image enhancement and segmentation. BioMed Res Int 2021;2021:6679556.

[25] Chalakkal R, Hafiz F, Abdulla W, et al. An efficient framework for automated screening of Clinically Significant Macular Edema. Comput Biol Med 2021;130(104128):104128.

[26] Rehman MU, Cho S, Kim J, et al. BrainSeg-Net: brain tumor MR image segmentation via enhanced encoder-decoder network. Diagnostics (Basel). 2021;11(2):169.

[27] Liu X, Ali TK, Singh P, et al. Deep learning to detect OCT-derived diabetic macular edema from color retinal photographs: a multicenter validation study. Ophthalmology Retina. 2022;6(5):398–410.

[28] Fu Y, Zhang G, Li J, et al. Fovea localization by blood vessel vector in abnormal fundus images,". Pattern Recognit. 2022;129:108711.

[29] Azad R, Heidari M, Shariatnia M, et al. TransDeepLab: convolution-free transformer-based DeepLab v3 + for medical image segmentation. arXiv [eess.IV]. 2022.

[30] Udayaraju P, Jeyanthi P, Sekhar BV. A hybrid multilayered classification model with VGG-19 net for retinal diseases using optical coherence tomography images. Soft comput. Sep 2023;27(17):12559–12570.

[31] Fu Y, Lu X, Zhang G, et al. Automatic grading of diabetic macular edema based on end-to-end network. Expert Syst Appl. 2023;213(1):118835.

[32] Babu PA, Rai AK, Ramesh JV, et al. An explainable deep learning approach for oral cancer detection. J Electr Eng Technol. 2023: 1–2.

[33] Zuiderveld KJ. Contrast limited adaptive histogram equalization. In: Graphics gems. 1st ed. San Diego, CA: Academic Press Professional, Inc.; 1994. p. 474–485.

[34] Gao W, Peng M, Wang H, et al. Incorporating word embeddings into topic modeling of short text. Knowl Inf Syst 2019;61(2):1123–1145.

[35] AbdelMaksoud E, Barakat S, Elmogy M. A computer-aided diagnosis system for detecting various diabetic retinopathy grades based on a hybrid deep learning technique. Med Biol Eng Comput 2022;60(7): 2015–2038.