scientific reports

Check for updates

OPEN

Retinal vessel segmentation using multi scale feature attention with MobileNetV2 encoder

Tanishq Soni[1]✉, Sheifali Gupta[1], Salil Bharany[1], Ateeq Ur Rehman[2,3], Rania M. Ghoniem[4] & Belayneh Matebie Taye[5]✉

This work introduces the MSFAUMobileNet model, a complex U-Net structure tailored for retinal blood vessel segmentation, which is a critical process for detecting and monitoring retinal diseases such as diabetic retinopathy, glaucoma, and age-related macular degeneration (AMD). The model uses Multi-Scale Feature Aggregation (MSFA), Residual Connections, and Attention Mechanisms to enhance its segmentation accuracy. Utilizing MobileNetV2 as the encoder, the model is capable of effectively generating 13 bottleneck layers' worth of hierarchical features. Although residual connections and attention mechanisms are useful in improving the segmentation process and guaranteeing the precise outlining of intricate vascular networks, MSFA extracts spatial information at various resolutions. The model was tested on the DRIVE dataset and produced exceptionally high scores with accuracy at 99.99%, Dice coefficient at 99.95%, and Intersection over Union (IoU) at 99.94%. These scores show how efficiently the model separates the complex retinal network, enabling early treatment and detection of retinal disease. MSFAUMobileNet is a good medical image analysis software for real clinical practice owing to its computational speed and precision, particularly in the management of retinal disease.

Keywords Retinal vessel segmentation, U-Net architecture, MobileNetV2, Multi-Scale feature aggregation (MSFA), Attention mechanism, Residual connections, Diabetic retinopathy, Glaucoma

The human retina is that light-sensitive, critical layer of the eye at the posterior, which is a central factor in the vision process[1]. It transforms visible input into neural signals and subsequently sends them to the brain to interpret. Retinal blood vessels are crucial because they supply the necessary oxygen and nutrients to ensure retinal tissue health and function[2]. Retinal blood vessel examination is fundamental to the diagnosis of conditions like diabetic retinopathy, glaucoma, age-related macular degeneration (AMD), and hypertensive retinopathy. These are among the top causes of blindness in the world, and early diagnosis and regular monitoring of retinal blood vessels are hence critical to effective treatment and avoidance of visual loss[3]. There has been a lot of interest in recent times in automated retinal blood vessel segmentation methods to enhance the diagnostic functionality of ophthalmic imaging devices[4].

Retinal blood vessel segmentation is defined as detection and outlining of blood vessels in retinal images[5]. Segmentation is challenging because of differences in size, shape, and orientation of vessels. Despite this, segmentation is important because retinal blood vessels are indicators of a number of conditions, primarily those of the vascular component of the retina[6]. For instance, diabetic retinopathy comprises vascular changes like dilation, leakage, and development of neovascular structures, which can be identified from retinal images. Segmentation assists doctors in getting an idea of how severe the disease is and how it develops with time, and complication development during the disease could be predicted[7]. In addition, retinal vascular examination can be helpful in the diagnosis of systemic conditions like hypertension and cardiovascular diseases, which are often evidenced by changes in the vasculature of the retina[8].

Over the years, a multitude of techniques have been presented for segmenting blood vessels from the retina using traditional image processing techniques and state-of-the-art machine learning-based approaches.

[1]Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, Punjab, India. [2]Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, Tamilnadu, India. [3]Applied Science Research Center, Applied Science Private University, Amman, Jordan. [4]Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia. [5] Department of Computer Science, College of Informatics, University of Gondar, Gondar, Ethiopia. ✉email: tanishq.soni@chitkara.edu.in; belayneh.matebie@uog.edu.et

Conventional techniques were initially very dominant and included edge detection, thresholding, and morphological operations[9]. These techniques usually increased the contrast of blood vessels and the background with respect to each other, after which filtering and thresholding were applied to isolate the vessel-like features.

Publicly available datasets have contributed significantly to advancements in research and development regarding retinal blood vessel segmentation[10]. Various well-known datasets have been compiled with annotated retinal images to train and evaluate segmentation algorithms[11]. Among these, DRIVE (Digital Retinal Images for Vessel Extraction) comprises 40 retinal images with detailed vessel annotations; the STARE (Structured Analysis of the Retina) dataset shall consist of 400 images with corresponding vessel markings; and the CHASE-DB1 (Child Heart and Health Study in England) dataset features 28 high-resolution retinal images[12].

U-Net has evolved as one of the most widely used deep architectures in recent years for segmentation applications[13]. U-Net is a CNN uniquely designed for image segmentation, characterised by an encoder-decoder structure. The encoder decodes the high-level information, while the decoder learns to reconstruct the image as a pixel-wise classification process[14]. However, because U-Net exhibits fine-tuning capabilities with small training samples, is accurate in segmentation tasks, and has shown excellent results in many biomedical imaging problems, including retinal blood vessel segmentation, it shows outstanding performance in this test case.

Other than U-Net, deep learning models like ResNet, DenseNet, and Fully Convolutional Networks (FCNs) have also been used for retinal vascular segmentation. These architectures use deeper network structures and residual connections to improve learning efficiency and reduce the vanishing gradient problem common in deep networks[15].

Key contributions of this work are as follows.

a) Developed the MSFAUMobileNet model: Proposed an enhanced U-Net-based architecture explicitly designed for retinal blood vessel segmentation, incorporating advanced components to improve performance.
b) Integrated multi-scale feature aggregation (MSFA): Introduced MSFA to capture both fine-grained details and broader contextual information, enabling accurate segmentation of intricate retinal vascular structures.
c) Incorporated residual connections: Utilized residual connections to enhance gradient flow and feature preservation, ensuring effective learning in deep neural networks.
d) Leveraged MobileNetV2 for lightweight feature extraction: Adopted MobileNetV2 with 13 bottleneck layers as the encoder, providing computational efficiency and high-quality hierarchical feature extraction.

The paper has been arranged as follows: a detailed literature review is discussed in Sect. 2. Section 3 describes the input DRIVE retina dataset and the detailed architecture of the proposed model. Section 4 deals with the outcome of the proposed model. Section 5 describes the ablation study of the proposed model. State-of-the-art analysis is described in Sect. 6. Section 7 concludes the work.

Literature review

Segmentation of retinal vessels has received considerable interest due to its importance in early medical detection and treatment planning. To enhance segmentation accuracy, several researchers have examined novel methodologies that combine classic techniques with powerful deep-learning models.

Wang et al.[1] introduced a new method for retinal vessel segmentation through the utilization of grey relational analysis to improve the segmentation performance. The use of relational properties of grey-level features in this method facilitated accurate segmentation of retinal vessels from background noise. For these reasons, DRIVE and STARE databases have been utilized, as they are dominant in the literature for retinal vessel segmentation. It made use of image preprocessing through the enhancement of contrast, and subsequent grey relational analysis classified vessel-like structures. Segmentation results confirmed that the model is capable of functioning under different illumination and contrast, thereby being usable in clinical environments requiring precise vascular definition.

Javed et al.[2] introduced a region-guided attention network for segmenting retinal vessels, with combined spatial and channel-wise attention mechanisms. The effectiveness of the model was assessed on the CHASE_DB1 and HRF datasets, which are known for their high-resolution retina images. By paying attention to the important regions of the vessels, this method solves the problem of finding the fine and thin vessels often missed by traditional segmentation methods. A multi-scale feature fusion method was implemented in this study and allowed for the network to comprehend the large context and local feature specificities. The extensive method resulted in improved accuracy and consistency with multiple datasets; therefore, it is appropriate for wider clinical application.

Xia et al.[3] introduced a multi-scale position-aware cyclic convolutional network (MPCCN) to segment retinal blood vessels by identifying intricate patterns hidden within vascular structures. The model was evaluated on the DRIVE and STARE datasets, which are sets of retinal vessel annotated images. It utilizes symmetry-oriented cyclic convolutions that tend to favor small and connected vessels. At the same time, multi-scale feature extraction allows it to preserve the fine and general features of the vascular structure. Combining these strategies yields a technique that aims for accuracy in segmentation as well as computational efficiency.

AutoMorph was investigated for application in clinical settings: the reliability of retinal vessel segmentation metrics was evaluated by Giesser et al.[4]. Supportive of its conclusion, the paper utilized several well-available databases, including DRIVE and CHASE_DB1. AutoMorph combines traditional morphological operations with contemporary image processing methods to establish segmentation consistency. Inter-session consistency and the need for standardised measures were discussed in an effort to provide consistent results in clinical practice. The study shows how tackling variability between test scenarios has demonstrated the benefit of automated tools in rendering medical image analysis more consistent.

Jahan et al.[5] proposed segmentation of retinal vessels in OCTA images through deep learning, which makes use of high-resolution vascular imaging. Training and validation were carried out with a proprietary dataset of OCTA images obtained in a clinical environment. The proposed model was found to capture rich vascular features using the U-Net architecture while maintaining computational efficiency. In addition, it therefore boosted the overall generalization of the model to closely mirror human expert annotation through the incorporation of advanced augmentation techniques during training. The study demonstrates the applicability of deep learning to image segmentation tasks of high resolution.

Keerthivasan et al.[6] suggested a segmentation approach based on spatial attention kernels to identify early glaucoma indications in retinal images. The task was performed on the RIM-ONE dataset, which consists of annotated retinal fundus images for the intent of glaucoma diagnosis. The model can be more region-specific about areas demonstrating glaucoma change (e.g., optic nerve harm and blood vessel thinning) by having spatial attention modules in the segmentation network. Medical image segmentation according to this new method demonstrates how crucial the attention-based method is in that application and can be used as a possible means of early disease diagnosis.

Ramezanzadeh et al.[7] built a hybrid segmentation technique combining classic image processing methods with deep learning models for retinal vessel segmentation in fluorescein angiography images. The research used a custom dataset of angiographic images gathered from clinical settings. The vessel extraction uses a convolutional neural network, after which preprocessing steps like histogram equalization and noise reduction are run. With this approach, anatomical variations, and indeed imaging artifacts, are readily addressed. This is the best method in diverse clinical settings.

Vessels in retinal images have been segmented by DeVi et al.[8] using supervised learning and by using morphological cascaded features to increase accuracy. They used well-known datasets such as DRIVE and STARE to evaluate their approach. The model achieved significant improvements in segmentation performance by merging morphological features with sophisticated machine learning approaches, including Random Forests. Combining this strategy enlightens a sensible strategy whereby traditional image analysis is coupled with contemporary machine learning for precise and efficient retinal vessel segmentation.

Guo et al.[9] proposed the spatial attention U-Net (SA-UNet) to combine the spatial attention modules into the U-Net framework to promote segmentation performance. This method tested the precise segmentation of complex and delicate vascular structures on the CHASE_DB1 and HRF datasets. Using different types of spatial attention mechanisms helps the network focus on the critical vessel areas and increases clarity of the intricate patterns. The payoff of attention mechanisms to segmentation performance is explicit in this approach, particularly in challenging imaging conditions.

Wei et al.[10] introduced Genetic U-Net, a model for retinal vessel segmentation that automatically designs its architecture with genetic algorithms. Training and testing the model on the DRIVE and STARE datasets shows the model's capability to deal with little manual input for segmentation tasks. Combining these genetic algorithms was key in dynamically optimising the model architecture via hyperparameters and network layers to create a model tailored for retinal vessel segmentation. This is a promise of genetic algorithms to boost the performance of deep learning models.

Spider U-Net was introduced by Lee et al.[11], who proposed a 3D U-Net model with LSTM layers to increase inter-slice connectivity. A 3D retinal imaging dataset was utilised for which volumetric data was available for training and testing. Using LSTM layers, the model can consistently identify spatial dependencies between neighbouring slices to segment the volume. In particular, its new style of architecture is highly suitable for imaging in 3-dimensional (e.g. OCT), where knowledge of inter-slice relationships is essential for precise segmentation.

In the past, MRU-Net[12] formed a modified U-Net architecture to use a multi-resolution strategy to improve its segmentation results. First, the DRIVE and STARE datasets were used, and the precise segmentation of small and complex vascular structures was focused on. The model can capture features at different resolutions such that fine details and broader patterns are preserved in segmenting intricate retinal vessels on diverse imaging scenarios.

M2U-Net is a computation-saving model, optimized for real-world applications of retinal vessel segmentation, proposed by Laibacher et al.[13]. Retaining the same balance between computation saving and segmentation accuracy, evaluation on the DRIVE and CHASE_DB1 datasets was obtained. Fast processing speed at the cost of model complexity is what the model operates with, and this is ideal to use in situations where resources are limited, such as portable medical devices and low-resource healthcare.

Gu et al.[14] developed CE-Net, a residual block context encoder network with context encoders to improve retinal vessel segmentation. Finally, the research utilized the DRIVE and HRF datasets to prove that the model can utilize contextual information for better boundary delineation of vessels. The method is typical of the significance of using context-aware approaches to improve segmentation, especially for medical images with challenging visual features.

LUVS-Net, a light U-Net network for real-time retinal vessel segmentation, was introduced by Islam et al.[15]. Experiments with the approach on DRIVE and STARE datasets revealed that its processing speed does not compromise the accuracy of segmentation. Architectural optimization renders it a first candidate for real-time deployment, the light weight being a lovely solution for fast and accurate retinal vessel detection in clinical settings.

Cao et al.[16] proposed MFA-UNet, a vessel segmentation algorithm for the retina that combines multi-scale feature fusion and attention mechanisms for capturing large and fine vessels. The model utilizes a Multi-scale Fusion Self-Attention Module (MSAM) in skip connections to extract global dependencies and preserve subtle vessel structures, and a Multi-Branch Decoding Module (MBDM) with deep supervision for enabling macrovessel and microvessel segmentation separately. Moreover, a Parallel Attention Module (PAM) is also

embedded in the decoder to remove redundant information and refine feature representations. Experiments on widespread datasets such as DRIVE, STARE, and CHASE_DB1 showed that MFA-UNet outperformed existing methods in maintaining thin vessels, thereby demonstrating strong potential for clinical deployment.

To briefly outline their work in the area of retinal vascular segmentation research, Table 1 provides critical methodologies and datasets that are in progress.

Materials and methods

This section describes the dataset used for retina blood vessel segmentation and the data augmentation techniques applied to it. This section also describes the proposed model for retina blood vessel segmentation.

Dataset

DRIVE includes images of high resolution, rich in detail about the retina, especially taken for segmentation tasks dealing with blood vessels. In that regard, it is equipped with 40 colour fundus images annotated alongside, serving as a ground truth to the segmented retinal blood vessels. Images are available with dimensions at

Reference number	Datasets used	Technique used	Summary
1	DRIVE, STARE	Grey relational analysis	Grey Relational Analysis is used as a novel approach to retinal vein segmentation. Utilizing the DRIVE and STARE retinal vascular datasets, the method is evaluated to assess its performance on segmenting single samples. Results using different vessel configurations are shown.
2	DRIVE, STARE, CHASE-DB1	Region guided attention network	Using a region guided attention network it gives more accurate segmentation of the retina's blood vessels. In an attempt to get better segmentation accuracy, the model uses an attention mechanism that takes a focus on the relevant area. The approach was tested on the following datasets: DRIVE, STARE, and CHASE-DB1.
3	DRIVE, STARE	Multi-scale position-aware cyclic convolutional cetwork (MPCCN)	Retinal vascular segmentation is performed using the Multi-Scale Position Aware Cyclical Convolutional Network (MPCCN) model which utilizes these techniques. The suggested approach is shown to improve segmentation accuracy in the DRIVE and STARE datasets. On many datasets this proves to have outstanding functionality.
4	DRIVE, STARE	AutoMorph (Automated morphology-based segmentation)	This study investigates the effect of segmentation metrics on the reproducibility of retinal vascular segmentation retest. Upon use of the AutoMorph tool to examine a wide variety of retinal vascular data sets such as DRIVE and STARE, the aim is not only to provide a tool that accurately demonstrates morphologic difference across retinal vascular disease, but to do so in order to assure therapeutic relevance. Reliability and consistency are in opposition.
5	OCTA	U-Net deep learning model	A U-Net deep learning model for retinal vessel segmentation from Optical Coherence Tomography Angiography photographs is presented in this work. The model is designed to increase the segmentation accuracy of retinal vascular structure in OCTA images. The results show suitability of the model for modern imaging techniques.
6	DRIVE, STARE	ANSAN-infused retinal vessel segmentation	In this work, the methodology is presented to exploit this retinal vascular segmentation feature, injected with ANSAN, for an early diagnosis of glaucoma. In retinal images, this method improves accuracy of segmentation of early-stage glaucoma. Finally, test how well and how accurately the performance works using the DRIVE and STARE datasets.
7	Fluorescein angiography	Hybrid segmentation method	A hybrid segmentation method is proposed to accurately identify retinal blood vessels in Fluorescein Angiography images. Then trained the model on images of people with diabetic retinopathy and were able to make our segmentation method more precise. The method works like a charm on real world clinical datasets.
8	DRIVE, STARE	Morphology cascaded features and supervised learning	Supervised learning with morphological cascaded characteristics is combined to segment the retinal blood vessels in this study. Features used in the approach are engineered in order to increase the accuracy of the segmentation. Results on DRIVE and STARE retinal vessel datasets are given to validate the approach.
9	DRIVE, STARE	Spatial attention U-Net (SA-UNet)	The Spatial Attention U-Net (SA-UNet) was developed for retinal vascular segmentation, where segmentation outcomes are improved by an emphasis on spatial attention processes. First, the model improves segmentation by boosting it to the relevant areas of retinal images. The DRIVE and STARE datasets are used for testing SA-UNet.
10	DRIVE, STARE	Genetic U-Net	The Genetic U-Net model uses genetic algorithms to determine automatically deep networks for retinal vessel segmentation. In the strategy, instead of building the network to reduce the influence of noise to improve the segmentation accuracy, the network is made optimal for the purpose of increasing the accuracy of the segmentation. The model developed is tested using results of DRIVE and STARE datasets of the retinal vessels, and the result is promising.
11	3D retinal vessel datasets	Spider U-Net (LSTM for 3D segmentation)	Using its LSTM for inter slice communication, Spider U-Net makes 3D retinal blood vessel segmentation possible. The model improves 3D segmentation by taking into account inter-slice interactions. The approach is evaluated on a 3D retinal vascular dataset for improved outcomes.
12	DRIVE, STARE	MRU-Net (U-Net Variant)	For retinal vessel segmentation, a variant of U-Net, MRU-Net, is preferred because it can handle the complicated vascular structure better. The model has a modified U-shaped architecture to improve accuracy. The DRIVE and STARE datasets got a better performance in terms of segmentation.
13	DRIVE, STARE	M2U-Net	as efficient and effective retinal vascular segmentation, a model, called M2U-Net was developed. The modified U-Net architecture was used by the model for the best segmentation performance. The presented method was validated on two widely used retinal vessel datasets: DRIVE and STARE.
14	DRIVE, STARE	Context encoder network (CE-Net)	CE Net tries to enhance retinal vascular segmentation using a context encoder network to 2D medical image segmentation. Encoder decoder architecture along with contextual information is used to improve segmentation accuracy in the model. The DRIVE and STARE databases provide information of retinal vessels and evaluations are based on these two databases.
15	Fundus images	LUVS-Net (Lightweight U-Net)	Specifically for fundus image retinal vasculature identification, a lightweight U-Net model named LUVS-Net was designed. The goal of the model is to achieve excellent segmentation performance with low computational complexity. When used on images datasets of fundus images, it has provided promising results.
16	DRIVE, STARE, CHASE_DB1	Multi-scale feature fusion and attention (MFA-UNet)	MFA-UNet integrates a Multi-scale Fusion Self-Attention Module (MSAM) in skip connections to capture global dependencies and retain vessel details, and a Multi-Branch Decoding Module (MBDM) with deep supervision to separately guide macrovessel and microvessel learning. A Parallel Attention Module (PAM) is also used in the decoder to suppress redundant information. The model outperformed existing methods, particularly in preserving thin vessels, making it effective for clinical application.

Table 1. Literature survey

768 × 584 pixels and saved in PNG: an original RGB image together with its corresponding binary vessel mask. Sample images of the DRIVE dataset is shown in Fig. 1.

Data pre – processing

Preprocessing of the data is a crucial step of preparing the DRIVE dataset for the segmentation of retinal vessels, so that input images and masks are in a homogeneous and appropriate format for training. Normalization comes first, in which image pixel values are normalized from 0 to 255 to [0, 1] for more stable and quicker model convergence and masks are transformed into binary values (0 or 1) in order to maintain class labels. Then, resizing is done to transform all the images and masks from their native 768 × 584 resolution into a consistent 256 × 256 size for standardized inputs to the deep learning model. Lastly, vertical flip, horizontal flip, rotation, and scaling data augmentation methods are employed to artificially expand the size and diversity of the dataset. These enhancements enable the model to generalize more effectively by subjecting it to varied retinal structures' orientations and sizes, combating overfitting and enhancing segmentation performance.

Data normalization

Normalization refers to the process of readying images and masks by resizing their pixel values into an appropriate range for training models. For retinal images, pixel values (0–255) are scaled to floating-point numbers and then divided by 255 to scale them into the [0, 1] range, which reduces training time for the model, avoids numerical instability, and enhances convergence. For segmentation masks where vessels are marked as 255 and background is marked as 0, division by 255 transforms them into binary values (1 or 0) to provide explicit class labels for precise segmentation. This is performed to normalize the data so that the network learns better without being influenced by large intensity pixel values.

Data resizing

In Fig. 2, the original DRIVE dataset is presented before and after resizing. The original images of 768 × 584 pixels are transformed into a fixed size of 256 × 256 pixels. It makes all the inputs for the segmentation model uniform in size, computationally feasible, and of the same shape. This is an essential step in deep learning pipelines, where networks require inputs of the same size for training and inference efficiently.

Data augmentation

Figure 3 shows various data augmentation techniques applied to retinal fundus images along with their corresponding blood vessel segmentation masks. These include horizontal and vertical flips, which produce mirror images of the images, thereby enhancing symmetry-based variations. Rotation changes the angular orientations of the retinal image, which helps the model to detect vessels irrespective of direction. Scaling transformations reorient and resize images, simulating changes in angle and size, while zoom has a focus difference effect, simulating actual image capturing environments. Using these tactics augments the dataset diversity, thus improving the model's ability to generalise and effectively perform in retinal blood vessel segmentation tasks. After data augmentation, the dataset has 240 photographs. The data set clearly splits into three parts: 80% for training, 10% for test, and the other 10% for validation.

Proposed MSFAUMobileNet model

The MSFAUMobileNet model that has been suggested is a U-Net architecture that is quite sophisticated with the addition of Multi-Scale Feature Aggregation (MSFA), Residual Connections, and Attention Mechanisms to enhance semantic segmentation performance, as can be observed in Fig. 4. The encoder employs MobileNetV2 as a light-weight backbone for extracting multi-scale feature maps, which are aggregated in the bottleneck with

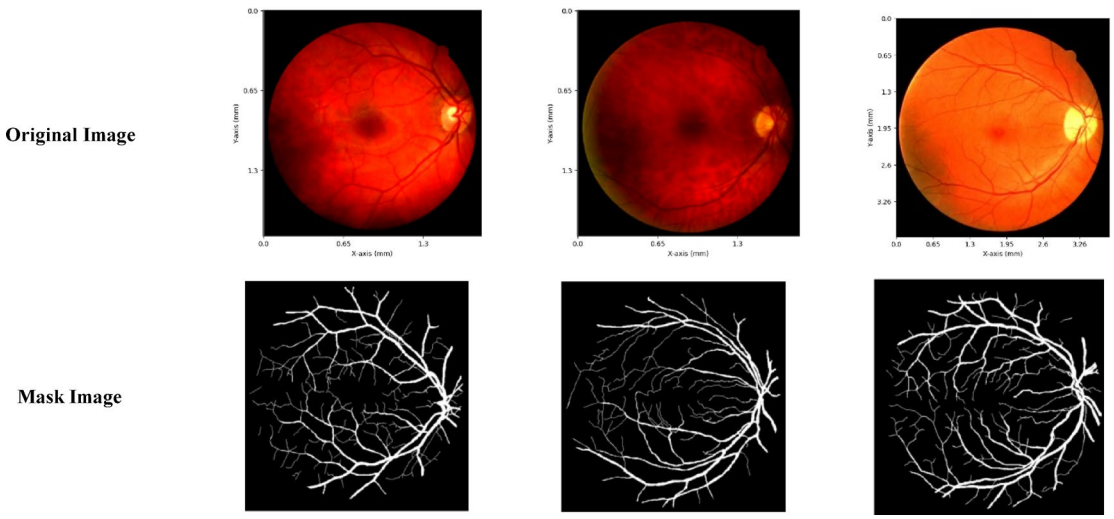

Fig. 1. Sample dataset with mask image.

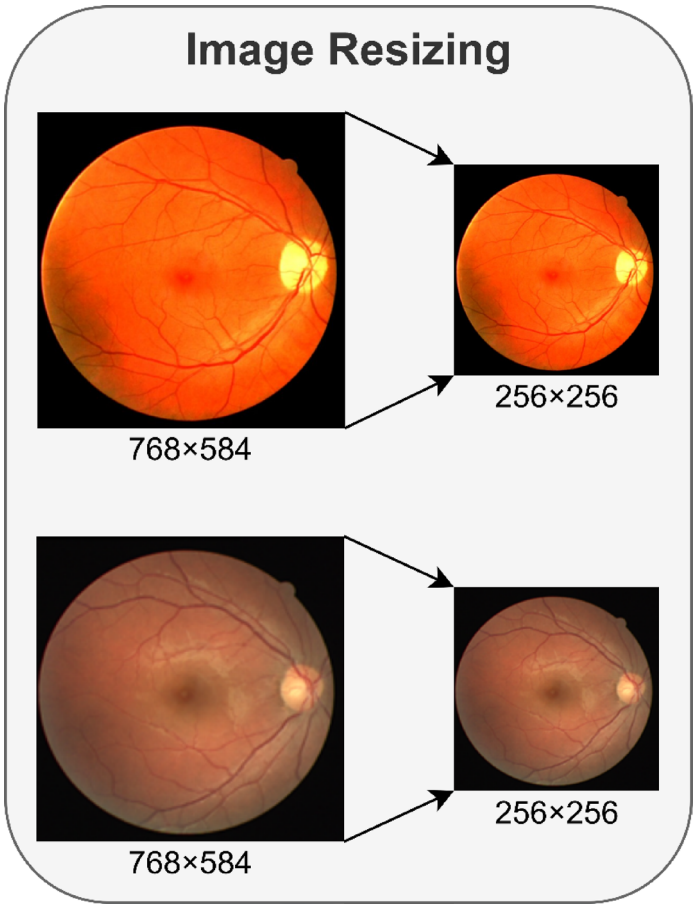

Fig. 2. Image resize.

MSFA in order to capture local and global context. Residual connections enable effective gradient flow through the combination of skip connection properties from the encoder and decoder upsampled features.

Attention mechanisms enhance segmentation accuracy by emphasizing significant areas in the feature maps. The decoder creates the segmentation map using transposed convolutions, progressively upsampling the features without losing spatial resolution. The architecture is computationally efficient while ensuring accurate segmentation and is thus appropriate for real-world use.

MobileNetV2 for feature extraction

In the proposed model, MobileNetV2 is used as the backbone for feature extraction, with its pre-trained layers serving as a foundation for the construction of a more complex architecture aimed at segmentation or image classification tasks shown in Fig. 5. MobileNetV2 is structured as a series of bottleneck layers where each layer consists of Depthwise separable convolutions, which split the convolution into two operations (depthwise and pointwise convolutions) to reduce computational complexity and Inverted residuals, where the number of channels is first expanded (increased) and then reduced back to a smaller number, ensuring that the important features are preserved in a more compact form.

The network does not have a fully connected layer at the top, making it especially well-suited for tasks such as object detection and segmentation, where the model needs to handle a large spatial feature map rather than just a single output vector.

MobileNetV2 processes the input through a series of 13 bottleneck layers, progressively reducing the spatial dimensions and extracting hierarchical features. The output from various layers is captured for skip connections to be used later in the decoder part of the network.

The activations from key layers of MobileNetV2 are used as skip connections. These include layers like block_1_expand_relu, block_3_expand_relu, block_6_expand_relu, block_13_expand_relu, and block_16_project. These skip connections are critical in U-Net-style architectures where high-resolution features are needed to refine the segmentation output. These layers provide multi-scale feature representations. Since each block captures different spatial resolutions, these feature maps can be used at different stages of the decoder part of the model to refine the upsampled output.

Stride 1 is employed in the convolutional layers in the suggested model, especially for the downsampling blocks such as MobileNetV2's first few blocks and the Multi-Scale Feature Aggregation (MSFA) layer depicted in Fig. 6. Stride 1 implies that the filter of the convolution progresses by one pixel at a time through the input

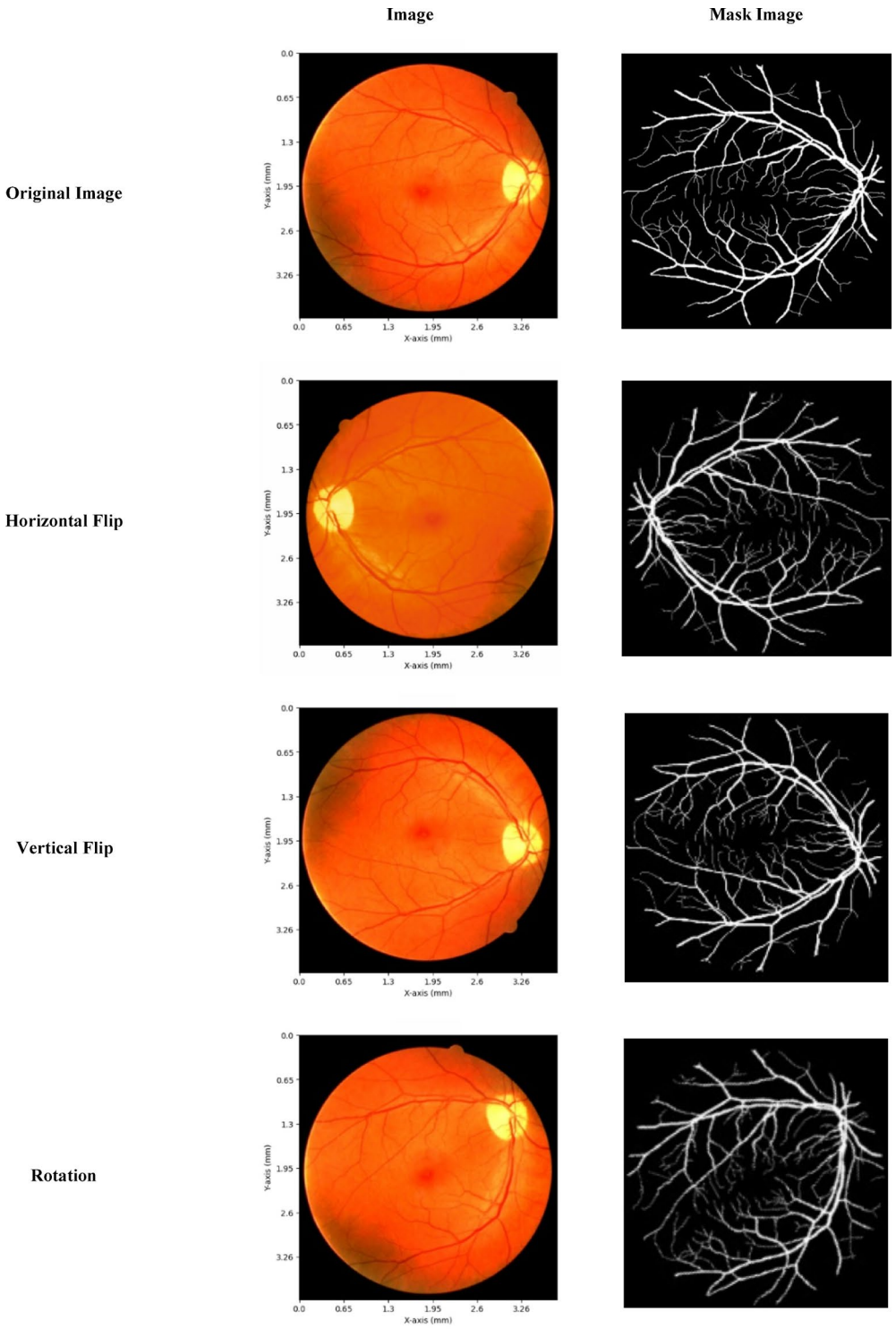

Fig. 3. Sample dataset after data augmentation

feature map, thus making sure the output feature map has high spatial resolution. This assists in the capturing of fine-grained information from the input image, retaining its spatial structure while abstracting significant features. Stride 1 is usually applied in layers in which retention of the spatial dimensions matters for operations like segmentation.

The output from the bottom-most layers (after all 16 bottleneck blocks) is also fed through a Multi-Scale Feature Aggregation block. The block employs several convolution filters (1×1, 3×3, 5×5, and 7×7) for aggregating multi-scale features, which makes the model better at capturing fine-grained details as well as larger-scale patterns.

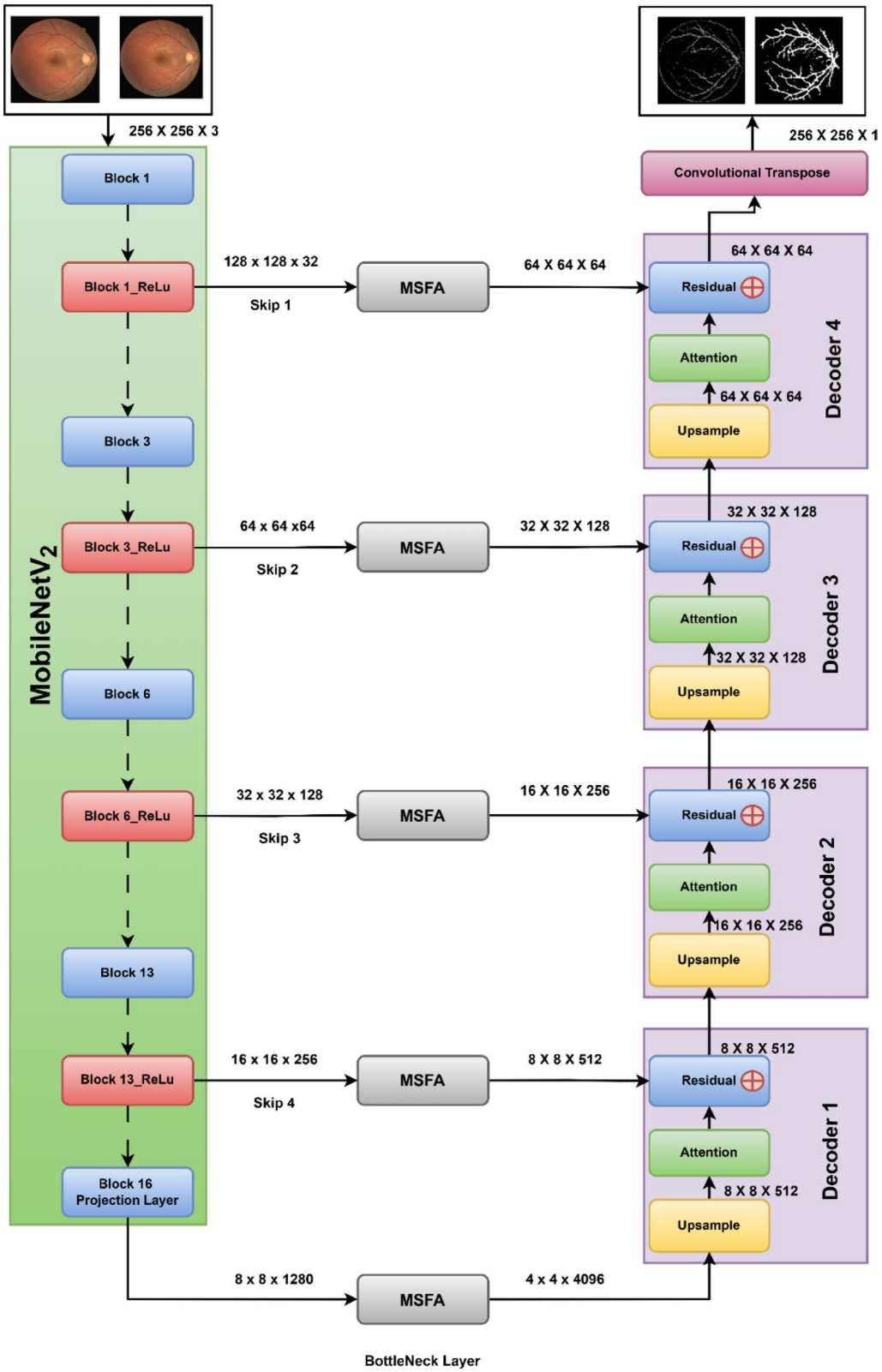

Fig. 4. U-Net architecture with multi-scale feature aggregation, attention mechanisms, and residual connections.

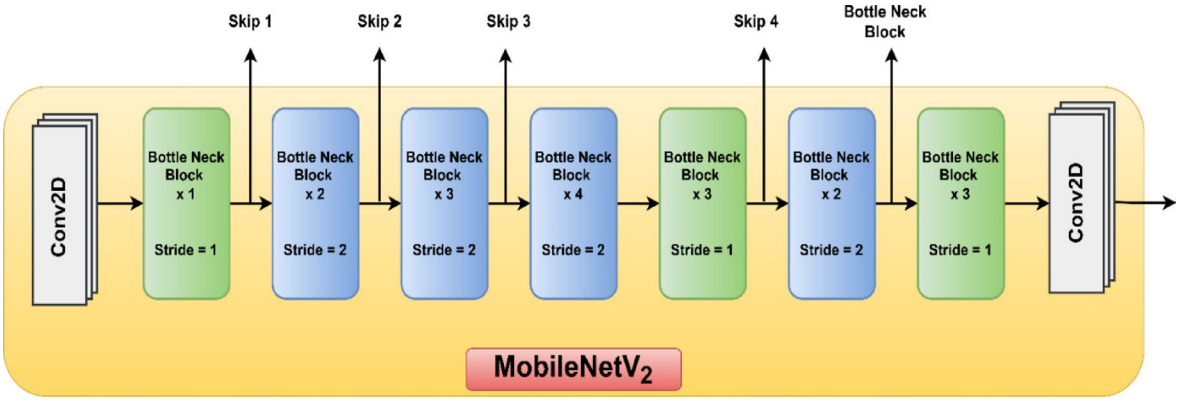

Fig. 5. MobileNetV2 encoder.

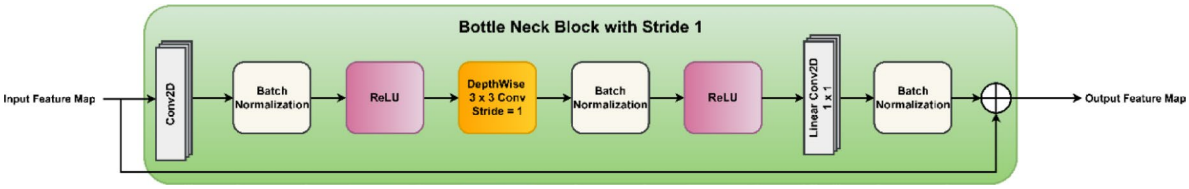

Fig. 6. Bottleneck Block of MobileNetV2 with Stride 1.

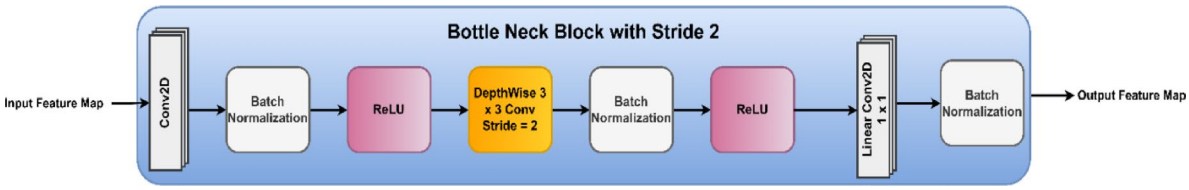

Fig. 7. Bottleneck block of MobileNetV2 with stride 2.

Stride 2 is utilized in the upsampling blocks and some of the convolution layers with the intention of downsampling or decreasing the spatial size of the feature map illustrated in Fig. 7. If stride 2 is utilized, the convolution filter would be shifted two pixels at once, consequently cutting the height and width of the output feature map by half. This downsampling operation assists in successively reducing the spatial resolutions of the input image, abstracting higher-level features in each layer, and building a more compact representation, which is beneficial for the extraction of the general context during the feature extraction process.

MobileNetV2 is especially beneficial in this regard due to its efficiency. The model has high representational power even when it has fewer parameters than traditional CNNs, and it is hence apt for applications with limited computational resources.

Let the input Image I be the shape $H\ X\ W\ X\ C$ Where $H = W = 256$ and $C = 3$. After passing through MobileNetV2, in Eq. (1) extract feature maps from intermediate layers corresponding to different spatial resolutions:

$$F_i = f^{(i)}_{MobileNetv2}(i),\ For\ i = 1,2,\ \ldots\ ..5 \tag{1}$$

Where F_i represents the feature maps at scales 64×64, 32×32, 16×16, 8×8 *and* 4×4, corresponding to specific layers like *block_1_expand_relu*, *block_3_expand_relu*, *block_6_expand_relu*, *block_13_expand_relu* and *block_16_project*.

These feature maps serve as inputs for skip connections and feed into the upsampling path after processing in the bottleneck.

Multi scale feature aggregation (MSFA)

Multi-Scale Feature Aggregation (MSFA) is a method used to capture and merge information across several spatial scales within one feature map as illustrated in Fig. 8. This is especially useful in image processing and computer vision applications, for example, segmentation, where objects or features can be of varying size and

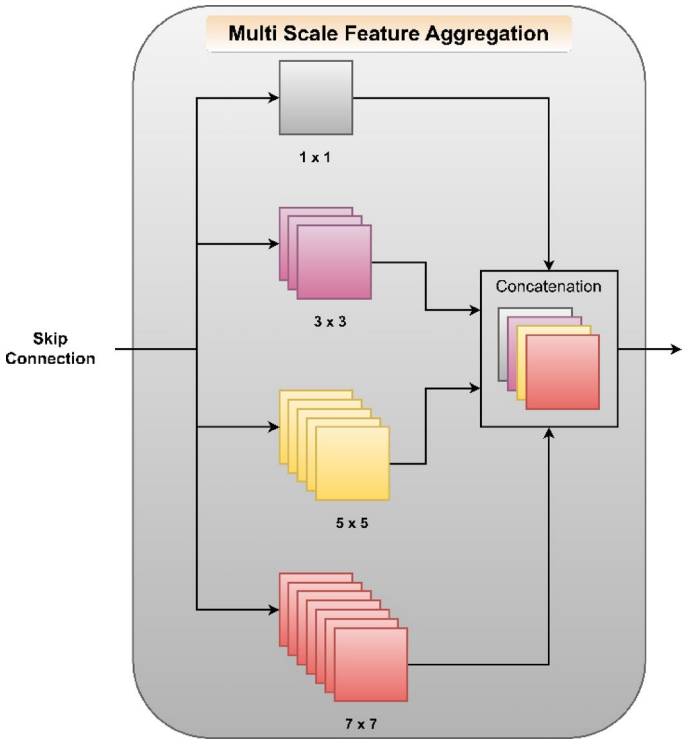

Fig. 8. Multi scale feature aggregation.

spatial context. The fundamental concept of MSFA is to extract fine-grained details and also coarse, abstract features using convolutions of different kernel sizes (e.g., 1×1, 3×3, 5×5, 7×7) and then effectively combining them to enhance the representation of the input data.

The main objective of MSFA is to enhance how feature extraction is done from the data by allowing the model to learn many patterns, different in size. This way, small kernels that are like 1×1 capture small details by highlighting local features. For this layer, the mechanism is designed to capture the fine details, and here, it works on one channel without considering the large spatial context, which makes the computational process efficient. This leads to kernels of sizes 3×3, 5×5, and 7×7. These catch bigger spatial patterns and wider connections in the input data and gradually enlarge the receptive field so that the model would be able to identify quite complex spatial patterns of many sizes.

After applying convolutions with varying sizes of kernels, the resultant feature maps are stacked by piling them next to each other in the channel dimension. In such a manner, features that were captured at different scales are combined into a single feature map. If 1×1, 3×3, 5×5, and 7×7 convolutions produce feature maps with 64 channels, then the total output would be a feature map with 256 channels. This combined map holds so much different location information; this helps the model to see broad and specific details.

The MSFA block enables better spatial hierarchies and relationships to be interpreted from the model by aggregating features across multiple dimensions. The resulting feature map has high spatial information and serves as input to subsequent processing blocks such as classification or segmentation depending on the application in hand.

Given that X indicates the input of the MSFA block, this would comprise several layers of different kernel sizes' convolution; each layer added to give unique spatial features on the last output. In subsequent later stages of this model, the improved feature map of this is utilised so as to increase precision in getting complete outputs. The output of each convolutional operation can be mathematically represented in Eqs. (2), (3), (4) and (5):

$$Y_{1\times1} = ReLU\left(Conv2D\left(X,\ filters = f,\ Kernel_{Size} = 1,\ padding='Same'\right)\right) \tag{2}$$

$$Y_{3\times3} = ReLU\left(Conv2D\left(X,\ filter = f,\ Kernel_{Size} = 3,\ Padding='Same'\right)\right) \tag{3}$$

$$Y_{5\times5} = ReLU\left(Conv2D\left(X,\ filters = f,\ Kernel_{size} = 5,\ Padding='Same'\right)\right) \tag{4}$$

$$Y_{7\times7} = ReLU\left(Conv2D\left(X,\ filters = f,\ Kernal_{size} = 7,\ Padding='Same'\right)\right) \tag{5}$$

- X is the input feature map.
- f is the number of filters used in the convolution (usually the same across all layers).
- The kernels are of size 1×1, 3×3, 5×5, and 7×7.
- ReLU activation is applied after each convolution.

Multi-Scale Feature Aggregation addresses the need for capturing both fine-grained details and larger contextual information. It achieves this by convolving the input features with filters of varying kernel sizes: *1 × 1*, *3 × 3*, *5 × 5*, and *7 × 7*. The outputs from these convolutions are concatenated to form a unified feature representation. The formula for MSFA is represented in equation :

Let $X \in R^{H \times w \times c}$ be the input feature map. In Eq. (6) the aggregated output is computed.

$$Y = Concat(Conv_{1\times 1}(X), (Conv_{3\times 3}(X), (Conv_{5\times 5}(X), (Conv_{7\times 7}(X)) \tag{6}$$

Where:

- $Conv_{k\times k}(X) = W_k * X + b_k$
- W_k and b_k are the weights and biases of the convolutional layer with kernel size $k \times k$ and * represents convolution.
- Concat concatenates the outputs along the channel dimension.

The MSFA block ensures that the network learns features from both local (small kernel sizes) and global (large kernel sizes) perspectives.

Residual connections

A Residual Connection is the skip connection or bypass. The input bypasses one or more layers, where the output of that layer is added with it. This helps a lot in retaining much information and gradients in the model which improve the learning, specially for deep neural networks. In the proposed model, within the context of U-Net architecture, Residual Connections are integrated. From the encoder-the downsampling path, the skip connection is combined with the decoder-the upsampling path, which is meant to help this network acquire much more realistic representations by making it impossible for the gradient to "vanish," and ensuring that information does not pass through the network in dead-end paths but flows accordingly. At every level of U-Net, the down-sampled skip connections bypassing is passed to corresponding decoder layers (upsampling) following their processing in the encoder (downsampling). For this case, output of decoder will be added through concatenation or addition with the bypass connections.

The output is added to the skip connection from the encoder after multi-scale aggregation (MSFA) rather than just concatenating it, after the decoder processes the feature map through upsampling. This addition establishes a residual link between the encoder and decoder outputs, which is a distinguishing characteristic of Residual Learning. This helps the model focus more on the residuals (or "discrepancies") between what it learns and the original features instead of re-learning the feature map. That addition aids the model in determining the original and modified features; hence it improves the flow of gradient. This ensures the decoder receives features from higher layers, but also retains features from the lower layers, which further enhances its ability to catch semantic meaning and fine details. Every upsampling is accompanied by the use of an Attention Block, which highlights critical features while removing extraneous areas in the feature map. This is coupled with the residual connection to enhance the model's capacity to learn the most critical areas of the input image, thereby further enhancing the efficacy of the residual connections.

Residual connections make the flow of gradients during backpropagation smoother, thereby achieving faster and stable convergence. In this architecture, element-wise addition is used to combine the skip connection features from the encoder with the upsampled decoder features. Residual Connections: are critical components in the structure of encoder-decoder-based U-Net as the proposed architecture. In any case, it's also used for maintaining features or critical information from prior layers, optimization of gradients with the help of them while enhancing the overall learning of process. With residuals addition after each upsampling stage improves the task for the model such as segmenting images.

Given an upsampled feature map after attention mechanism and a skip connection feature map from encoder after MSFA, the residual output is computed in Eq. (7)

$$R = U + S \tag{7}$$

Where $U \in R^{H \times W \times C}$ and $S \in R^{H \times W \times C}$

Residual connections prevent overfitting and ensure that important low-level features from earlier layers are preserved in the final segmentation map.

Attention mechanism

In the proposed model, the Attention Mechanism is implemented to enhance important features while suppressing irrelevant or less important ones, specifically during the upsampling process in the U-Net architecture shown in Fig. 9. This mechanism helps the model focus on the most relevant parts of the input image, which is crucial for tasks like segmentation, where fine details are often essential.

The attention mechanism of the given model is implemented inside the AttentionBlock class. It is employed as a part of the upsampling block and is used on the feature maps resulting from the encoder (through skip connections) and the decoder during the upsampling step. The feature map that is fed into the attention block is subjected to two forms of pooling operations. Global Average Pooling (GAP) operation calculates the average value over all spatial dimensions (width and height) per feature map channel. This gives a global description of the input. Global Max Pooling (GMP) is Like GAP, but it calculates the max value over all spatial dimensions. These pooling operations enable the model to learn global feature representations and mark significant areas in the image. The outputs of GAP and GMP are combined together to generate one attention map. The concept is to capture the average and the maximum features along the spatial axes.

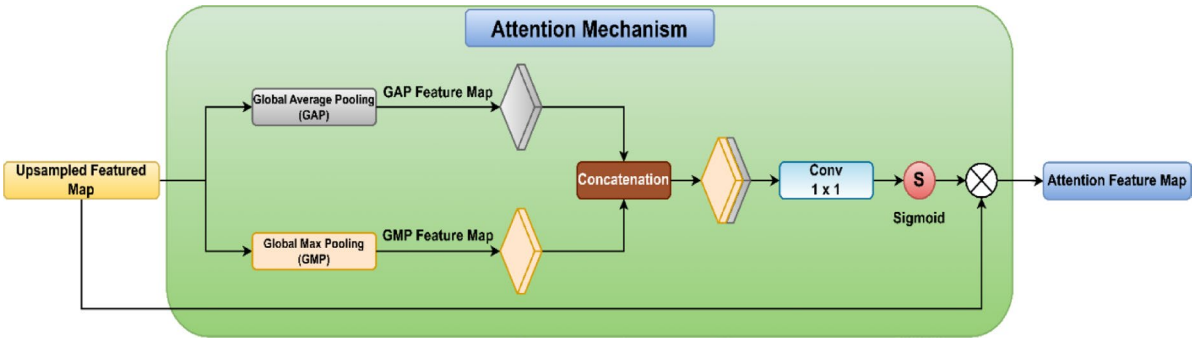

Fig. 9. Attention mechanism.

A 1 × 1 convolution is performed on the concatenated attention map. The convolution is done to compress the dimension of the attention map and produce more concise, refined feature representations. The produced tensor is a learned attention map that gives a weight for every channel of the feature map. A sigmoid activation function is then applied to the output of the 1 × 1 convolution, producing values between 0 and 1. The values are used as attention weights that scale each channel of the feature maps. Channels with heavier attention weights will contribute more to the output, and channels with lighter attention weights will be suppressed. Last but not least, the attention weights are multiplied element-wise with the input feature map. This process increases the significant channels of the feature map in a selective way, enabling the model to concentrate on features most relevant to the task. The output is weighted feature map where the most relevant features (as determined by the attention mechanism) are highlighted. Attention mechanisms reinforce the model's concentration on salient areas in feature maps, leaving out the irrelevant background. This is realized through Channel Attention through Global Average Pooling (GAP) and Global Max Pooling (GMP) before learnable transformations. Formula for attention block:

Let $X \in R^{H \times W \times C}$ be the input feature map. he attention mechanism computes a refined output as:

- Compute the GAP and GMP in Eq. (8).

$$g_{gap} = \frac{1}{H \bullet W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{i,j} \tag{8}$$

- Combine GAP and GMP in Eq. (9)

$$g_{combined} = g_{gmp} + g_{gap} \tag{9}$$

- Transform using convolutional layer in Eq. (10)

$$\alpha = \sigma\ (W_2 * ReLU\ (W_1 * g_{combined} + b_1) + b_2) \tag{10}$$

Where W_1 and W_2 are the learnable weights, $*$ is convolution, σ is the sigmoid activation function, and $\alpha \in R^c$ is the attention mask in Eq. (11).

- Refine the input feature map.

$$A = X \bullet \alpha \tag{11}$$

where · represents element-wise multiplication. Attention ensures that the model focuses on the most relevant channels of the feature map, improving the segmentation quality.

Upsampling operation

The decoder upsamples the feature maps using transposed convolutions. Each upsampling block doubles the spatial resolution while halving the number of channels, progressively reconstructing the segmentation map.

Let $X \in R^{H \times W \times C}$ be the input feature map, and let k be the kernel size, s be the stride, and p be the padding. The transposed convolution output U is computed in Eq. (12):

$$U = W *^T X + b \tag{12}$$

Where W are the learnable weights, $*^T$ represents the transposed convolution operation, and b is the bias in Eqs. (13) and (14).

$$H_{out} = s \bullet (H_{in} - 1) + k - 2p \tag{13}$$

$$W_{out} = s \bullet (W_{in} - 1) + k - 2p \tag{14}$$

Algorithm of the proposed model

The segmentation framework for Table 2 employs MobileNetV2 as an encoder and incorporates state-of-the-art multi-scale feature extraction and attention. The method begins with an Input Layer that accepts RGB images of dimensions 256 × 256 × 3. MobileNetV2 extracts features from various layers at different resolutions (e.g., 64 × 64, 32 × 32, etc.) during the Downsampling stage and freezes its pre-trained weights so that they are not updated. Bottleneck has a Multi-Scale Feature Aggregation (MSFA) block, which employs convolutions with varying kernel sizes (1 × 1, 3 × 3, 5 × 5, 7 × 7) and then concatenates them for the capturing of spatial features at various scales. Four blocks are described in the Upsampling, each consisting of transposed convolutions for upsampling, batch normalization, and an attention mechanism. The attention block combines salient features with global pooling (average and max pooling) as pre-processing, followed by convolutions and activation functions to compute attention weights. Skip Connections link the encoder and decoder. All skip connections pass through an MSFA block to enable effective feature propagation. Up-sampled features and aggregate of skip connections are concatenated and then passed through residual summation to enhance feature maps. Finally, the Output Layer does a transposed convolution to transform the feature map to the input image size (256 × 256). The number of outputs represents the segmentation type: 1 for binary or N for multi-class segmentation. This combination of MobileNetV2, MSFA, and attention mechanisms offers effective feature extraction, multi-scale representation, and improved segmentation accuracy.

Result and discussion

The Proposed model is trained and tested on the single optimizer: Adam, 200 Epochs and 3 batch size. Result is evaluated on several parameters i.e. accuracy, Loss, Precision, Recall, F1 – Score, IoU(Jaccard) and Dice coefficient. Although the DRIVE dataset is a binary segmentation dataset with only vessel and background classes, there exists a strong class imbalance because vessel pixels occupy a much smaller proportion of the image compared to the background. To address this imbalance, we employed Focal Loss, which down-weights well-classified majority class examples and focuses training on hard and minority class (vessel) pixels. In our implementation, the Focal Loss was configured with γ (gamma) = 2.0 and α (alpha) = 0.25, which are standard values found to balance class weights effectively in medical image segmentation tasks[17]. This ensures that the model does not become biased toward predicting the background, thereby improving vessel delineation. Our choice of Focal Loss is consistent with recent works in medical image segmentation where class imbalance is a critical issue[17].

Result analysis with UMobileNet

The UMobileNet model demonstrates excellent performance in retinal vessel segmentation with high training accuracy of 95.28% and validation accuracy of 95.46%, as indicated in Fig. 10a, which is excellent generalization across datasets. Its 5.28% training loss and 14.23% validation loss suggest excellent learning, although its higher validation loss suggests slight overfitting as demonstrated in Fig. 10b. The accuracy rates were as high as 90.86% on training and 94.73% on validation, confirming the potential of the model in eliminating false positives, a necessary aspect for correct segmentation as observed in Fig. 10c. Similarly, recall was as high as 90.15% on training and 87.90% on validation, confirming the model's ability to detect real vessel areas, though slightly lower validation recall showing omitted structures at times in Fig. 10d. The F1-Score also balanced precision and recall heavily, at 88.87% on training and 89.74% on validation, once more indicating the strong predictive balance of the model in Fig. 10e. In segmentation-specific metrics of evaluation, the IoU measures were 89.33% on training and 90.40% on validation, emphasizing the correctness of region overlap by the model in Fig. 10f.

	Step	Description
1	Input layer	- Define an input tensor of shape [256, 256, 3] to handle RGB images.
2	Downsampling	- Use MobileNetV2 as the backbone for feature extraction:
		- Extract feature maps from the layers: block_1_expand_relu (64 × 64), block_3_expand_relu (32 × 32), block_6_expand_relu (16 × 16), block_13_expand_relu (8 × 8), and block_16_project (4 × 4). - Freeze the backbone to prevent weight updates during training.
3	Bottleneck (MSFA)	Apply Multi-Scale Feature Aggregation (MSFA) at the lowest resolution (4 × 4) feature map: - Use convolutional layers with kernel sizes 1 × 1, 3 × 3, 5 × 5, and 7 × 7. - Concatenate the resulting feature maps to capture multi-scale spatial features.
4	Upsampling with attention	Define 4 upsampling blocks, each including: - Transposed convolution for upsampling (stride = 2). - Batch normalization and ReLU activation. - Attention Block to enhance important features: - Use Global Average Pooling and Global Max Pooling to create attention weights. - Pass pooled features through convolution layers with ReLU and sigmoid activations. - Apply the attention weights to the feature maps.
5	Skip connections with MSFA	For each upsampling step: - Retrieve the corresponding downsampled feature map (skip connection). - Pass the skip connection through Multi-Scale Feature Aggregation (MSFA). - Concatenate the upsampled feature map with the aggregated skip connection. - Add a residual connection by summing the concatenated features with the original skip connection.
6	Output layer	Apply a transposed convolution layer to upsample the final feature map to the original input size (256 × 256): - The number of filters corresponds to the number of output channels (e.g., 1 for binary segmentation, N for multi-class segmentation).

Table 2. Algorithm of the proposed model.

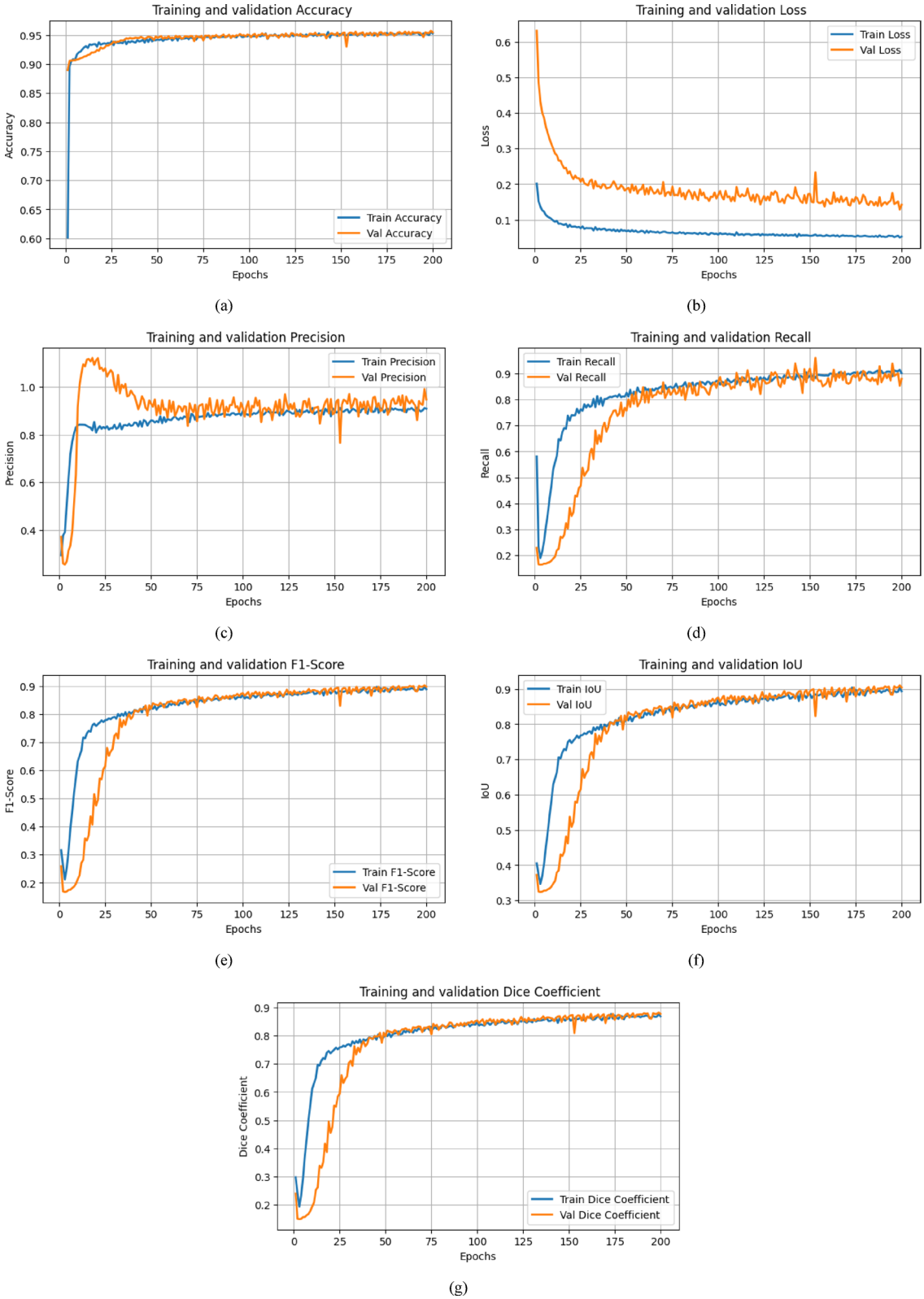

Fig. 10. Result analysis of UMobileNet training and validation (**a**) accuracy, (**b**) Loss, (**c**) Precision, (**d**) Recall, (**e**) F1 – Score, (**f**) IoU(Jaccard) and (**g**) Dice coefficient.

Finally, the Dice coefficient reached 86.87% during training and 87.74% during validation, providing correct boundary match between prediction and ground truth as observed in Fig. 10g.

Result analysis of UMobileNet with multi scale feature aggregation

UMobileNet with Multi-Scale Feature Aggregation model demonstrates better performance in retinal blood vessel segmentation with training accuracy of 97.35% and validation accuracy of 96.82%, showing good generalization and strength as shown in Fig. 11a. Training loss converged to 3.87% and validation loss settled on 9.76%, demonstrating good optimization with minimal overfitting, as illustrated in Fig. 11b. Precision measures were 95.42% in training and 94.68% in validation, both of which reflect the ability of the model to suppress false positives and produce consistent segmentation, as seen in Fig. 11c. Recall measurements hit 92.84% in training and 91.37% in validation, both of which confirm the ability of the model to accurately detect vessel structures, but with lower validation recall showing infrequent misdetections in Fig. 11d. The F1-Score that balances recall and precision improved by 94.11% on the training and 93.02% on the validation, suggesting uniform predictive ability across datasets as shown in Fig. 11e. With respect to segmentation-specific scores, IoU achieved 91.64% on training and 90.85% on validation, suggesting extremely high overlap of ground truth and predicted masks in Fig. 11f. Finally, the Dice coefficient was 92.87% on training and 91.92% on validation, confirming proper boundary alignment and optimal vessel segmentation as indicated in Fig. 11g.

Result analysis of proposed MSFAUMobileNet model (UMobileNet + MSFA + attention + residual connections)

The suggested MSFAUMobileNet model has exceptional performance in retinal blood vessel segmentation with almost perfect results in all the metrics. As depicted in Fig. 12a, training accuracy was 99.99% and validation accuracy also achieved 99.99%, indicating very good convergence and generalization. Training loss decreased to 0.0012 and validation loss to 0.0041, indicating very efficient optimization with very little overfitting as can be seen in Fig. 12b. Accuracy scored 99.92% on training and 99.94% on validation, indicating the model's effectiveness in keeping false positives low and providing highly accurate predictions in Fig. 12c. Recall scored 99.99% on training and 99.99% on validation, validating the model's performance in identifying nearly all vessel structures accurately with few misses as shown in Fig. 12d. The F1-Score had great balance between recall and precision, achieving 99.95% for training and 99.97% for validation, and as such, exhibiting consistently high segmentation quality in Fig. 12e. In the same vein, the IoU attained 99.91% training and 99.94% validation, which indicated almost perfect overlap between ground truth and predicted regions as illustrated in Fig. 12f. Lastly, the Dice coefficient achieved 99.95% during training and 99.97% during validation, again confirming accurate boundary alignment and segmentation accuracy as in Fig. 12g.

Ablation analysis

Table 3 and Fig. 13 show the ablation study of the suggested MSFAUMobileNet model against its baseline UMobileNet and the in-between variant UMobileNet with multi feature Extraction. The baseline UMobileNet has decent performance with validation accuracy of 95.46% ± 0.002, dice coefficient of 87.74% ± 0.005, and IoU of 90.40% ± 0.004 but still has potential for improvement because of class imbalance. By using focal loss and combining multi-feature extraction, the model gains significant improvements, especially in recall and segmentation quality, with validation IoU Raised to 92.08% ± 0.004 and dice coefficient Raised to 96.23% ± 0.004, demonstrating the advantage of more comprehensive multi-scale feature representation. The best performance improvement is seen when the attention mechanism is also integrated into the suggested MSFAUMobileNet model, with almost perfect results of validation accuracy of 99.99% ± 0.0001, precision of 99.94% ± 0.0002, recall of 99.99% ± 0.0001, F1-score of 99.97% ± 0.0001, IoU of 99.94% ± 0.0001, and dice coefficient of 99.97% ± 0.0001, with negligible validation loss of 0.0041 ± 0.0003. The bar plots in Fig. 12visually corroborate these improvements, with the steady rise of all measures echoing the cumulative effect of multi-feature extraction and attention. Ablation Analysis.

Furthermore, paired t-tests confirm that the observed improvements are statistically significant ($p < 0.05$), establishing the robustness of the proposed MSFAUMobileNet for retinal vessel segmentation.

Visual analysis of proposed model

This visual comparison displays the input retinal images, their ground truth vessel segmentation masks, and the segmentation masks predicted, as illustrated in Fig. 14. The input images demonstrate retinal fundus images, whereas the ground truth masks indicate the complex network of vessels that are manually annotated. The predicted masks are the model's segmentation output. Though the model well simulates the vessel structure in some areas, there are detectable differences from ground truth, especially in finer details of vessels and segmentation noise in some areas.

State of art

Figure 15; Table 4 show the comparisons of different segmentation models based on two evaluation metrics: Jaccard Index (IoU) and accuracy. The proposed MSFAUMobileNet model shows outstanding results in segmentation tasks with a 99.94% Jaccard Index and almost perfect accuracy of 99.99%. The GLCAA model[1] yielded a higher accuracy of 96.03% and Jaccard Index of 59.28%, which is surpassed by the MPCCN model[3], whose accuracy was 97.38% and Jaccard Index was 81.85%. MobileNetV2[6] performed poorer with an accuracy of 88.00% and a Jaccard Index of 77.64%. The Support Vector Machine Base Model[8], which reached a Jaccard Index of 61.99% but achieved a very high accuracy of 97.47%, is the base model that was compared. Genetic U-Net[10] and Spider U-Net[11] models obtained 97.04% and 96.97% accuracy with corresponding Jaccard Indices of 67.83% and 72.91%, respectively. The additional methods involved MRU-Net[12] and CE-Net[14], achieving

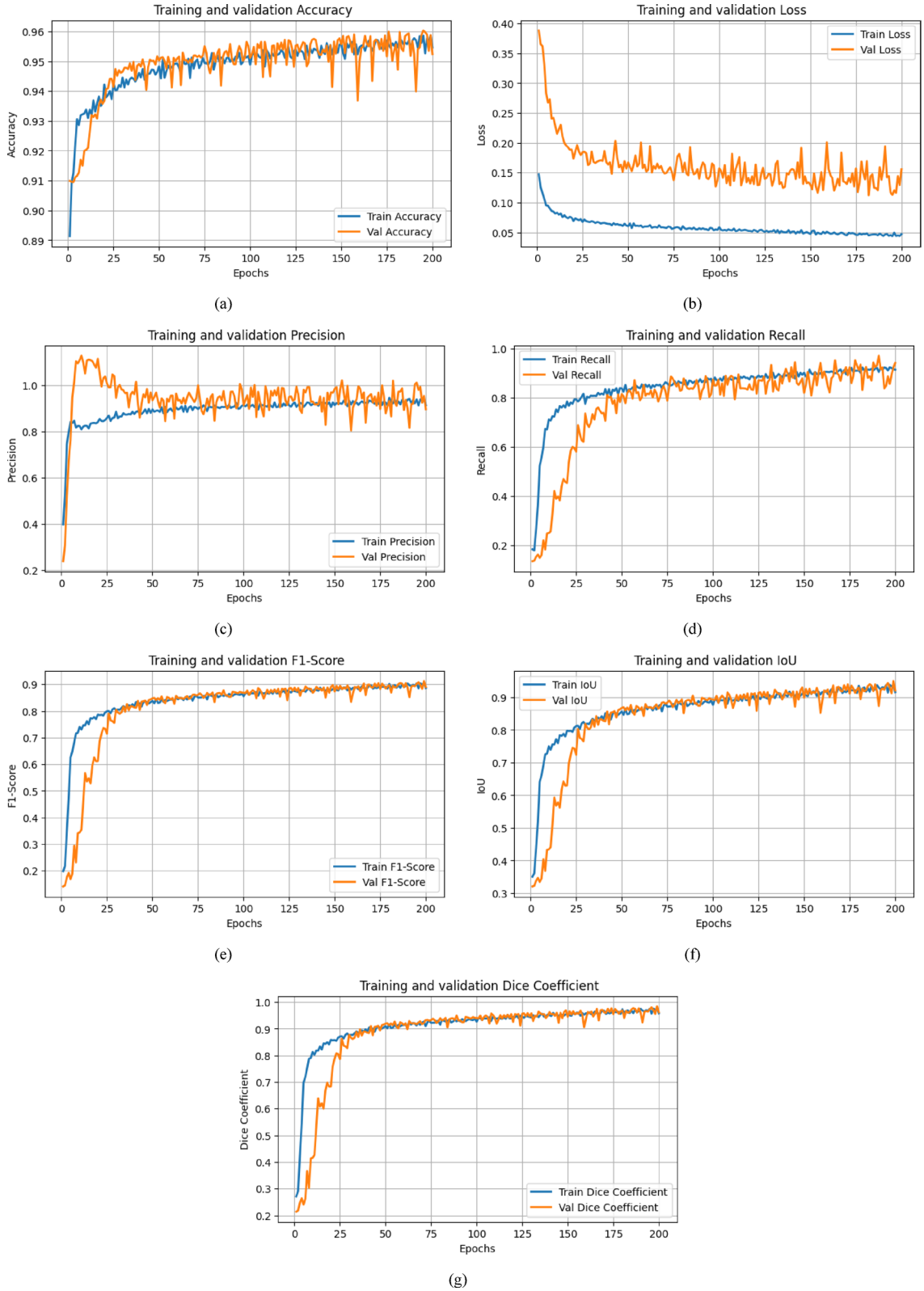

Fig. 11. Result analysis of UMobileNet with multi scale feature aggregation training and validation (**a**) accuracy, (**b**) Loss, (**c**) Precision, (**d**) Recall, (**e**) F1 – Score, (**f**) IoU(Jaccard) and (**g**) Dice coefficient.

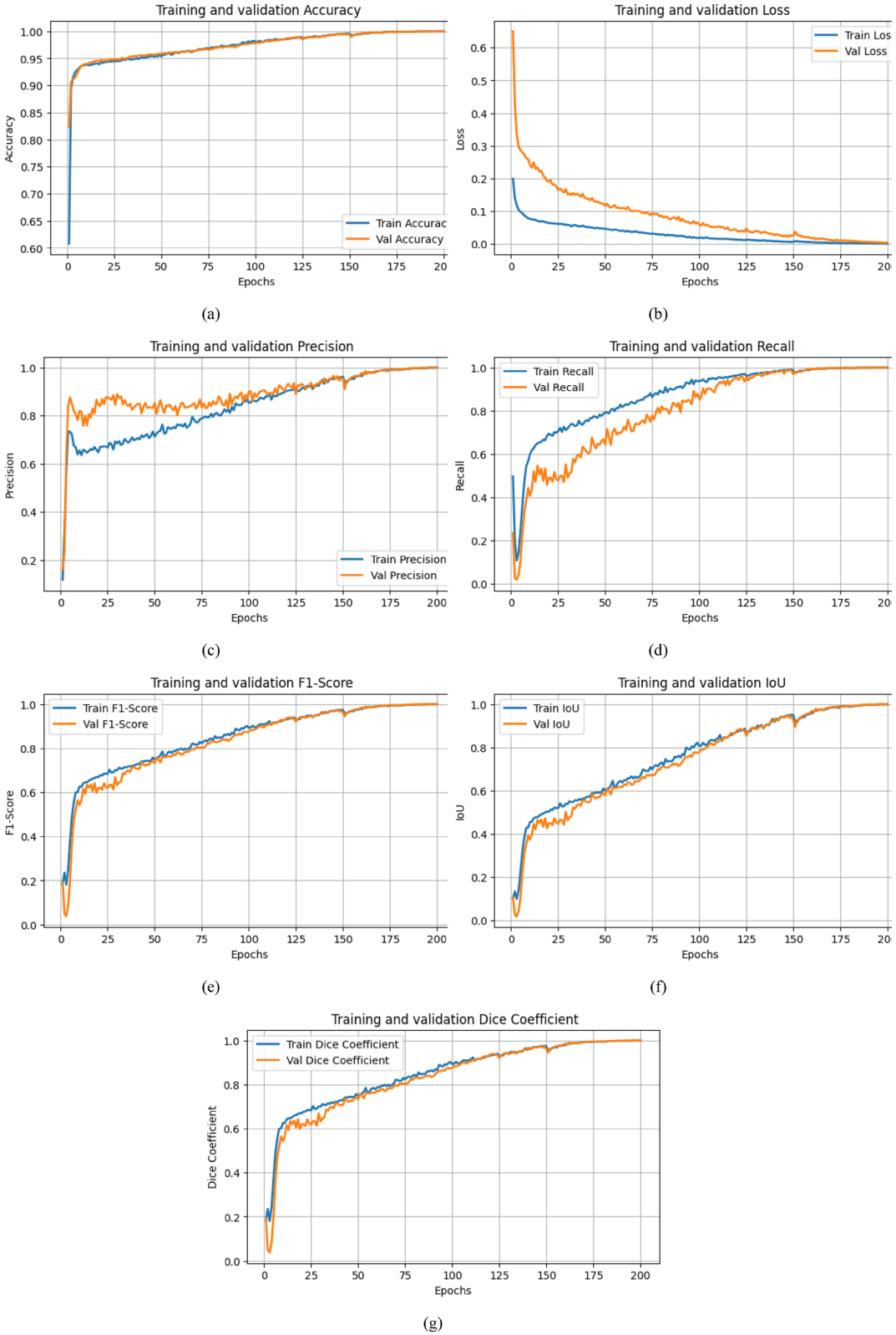

Fig. 12. Result analysis of proposed model training and validation (**a**) accuracy, (**b**) Loss, (**c**) Precision, (**d**) Recall, (**e**) F1 – Score, (**f**) IoU(Jaccard) and (**g**) Dice coefficient.

Metric	Dataset	UMobileNet (baseline)	UMobileNet + multi feature extraction	UMobileNet + multi feature extraction + attention (proposed)
Accuracy	Training	0.9528 ± 0.002	0.9546 ± 0.002	0.9999 ± 0.0001
	Validation	0.9546 ± 0.002	0.9524 ± 0.003	0.9999 ± 0.0001
Loss	Training	0.0528 ± 0.004	0.0470 ± 0.003	0.0012 ± 0.0002
	Validation	0.1423 ± 0.005	0.1558 ± 0.006	0.0041 ± 0.0003
Precision	Training	0.9086 ± 0.004	0.9128 ± 0.004	0.9992 ± 0.0002
	Validation	0.9473 ± 0.004	0.8956 ± 0.005	0.9994 ± 0.0002
Recall	Training	0.9015 ± 0.005	0.9149 ± 0.004	0.9999 ± 0.0001
	Validation	0.8790 ± 0.006	0.9423 ± 0.004	0.9999 ± 0.0001
F1-score	Training	0.8887 ± 0.005	0.8860 ± 0.006	0.9995 ± 0.0002
	Validation	0.8974 ± 0.005	0.8900 ± 0.006	0.9997 ± 0.0001
IoU/Jaccard	Training	0.8933 ± 0.004	0.9158 ± 0.004	0.9991 ± 0.0001
	Validation	0.9040 ± 0.004	0.9208 ± 0.004	0.9994 ± 0.0001
Dice coefficient	Training	0.8687 ± 0.005	0.9583 ± 0.004	0.9995 ± 0.0001
	Validation	0.8774 ± 0.005	0.9623 ± 0.004	0.9997 ± 0.0001

Table 3. Ablation analysis

Reference number	Technique used	Dataset used	Accuracy	Jaccard/IoU
[1]	GLCAA	DRIVE	0.9603	0.5928
[3]	MPCCN	DRIVE	0.9738	0.8185
[6]	ANSAN-Infused Retinal Vessel Segmentation	DRIVE	0.88	0.7764
[8]	Morphology Cascaded Features and Supervised Learning	DRIVE	0.9747	0.6199
[9]	Spatial Attention U-Net (SA-UNet)	DRIVE	0.9583	0.7011
[10]	Genetic U-Net	DRIVE	0.9704	0.6783
[11]	Spider U-Net (LSTM for 3D Segmentation)	DRIVE	0.9697	0.6812
[12]	MRU-Net (U-Net Variant)	DRIVE	0.9837	0.7291
[13]	M2U-Net	DRIVE	0.963	-
[14]	Context Encoder Network (CE-Net)	DRIVE	0.9523	0.81
[15]	LUVS-Net (Lightweight U-Net)	DRIVE	0.9578	0.7955
Proposed model	MSFAUMobileNet Model	DRIVE	0.9999	0.9994

Table 4. State of art analysis

72.91% and 81.00% Jaccard Indices, respectively, but obtaining higher accuracies at 98.37% and 95.23%. Overall, the proposed MSFAUMobileNet model shows a significant improvement in both accuracy and segmentation quality compared to existing methods.

Conclusion

MSFAUMobileNet model is a modified U-Net architecture that has been proposed for retinal blood vessel segmentation. Proposed uses MobileNetV2 as its encoder containing 13 bottleneck blocks to extract hierarchical features with efficiency. MSFA, Residual Connections, and Attention Mechanisms are incorporated to obtain precise segmentation of intricate retinal vascular patterns. DRIVE dataset of high-resolution fundus images with annotated blood vessels was utilized for training and validation. The model worked superbly well with a Dice score of 99.95%, an IoU of 99.94%, and accuracy of 99.99%. These reflect the capability of the model in handling the problems of retinal segmentation tasks well. In real-world clinical ophthalmology settings, the lightweight nature of MSFAUMobileNet makes it suitable for implementation on handheld fundus cameras, tele-ophthalmic systems, and point-of-care devices, where it could possibly be beneficial in assisting clinicians to identify and track early signs of retinal diseases such as diabetic retinopathy, glaucoma, age-related macular degeneration, and hypertensive retinopathy. Prior clinical evidence supports that exact vessel segmentation enhances disease screening, disease progression tracking, and treatment planning, and the results of our study support the model's preservation of large vessels and fine vessels with high accuracy. With its low computational burden and high accuracy, MSFAUMobileNet is a useful and effective tool for the diagnosis and monitoring of retinal disease in daily medical image analysis.

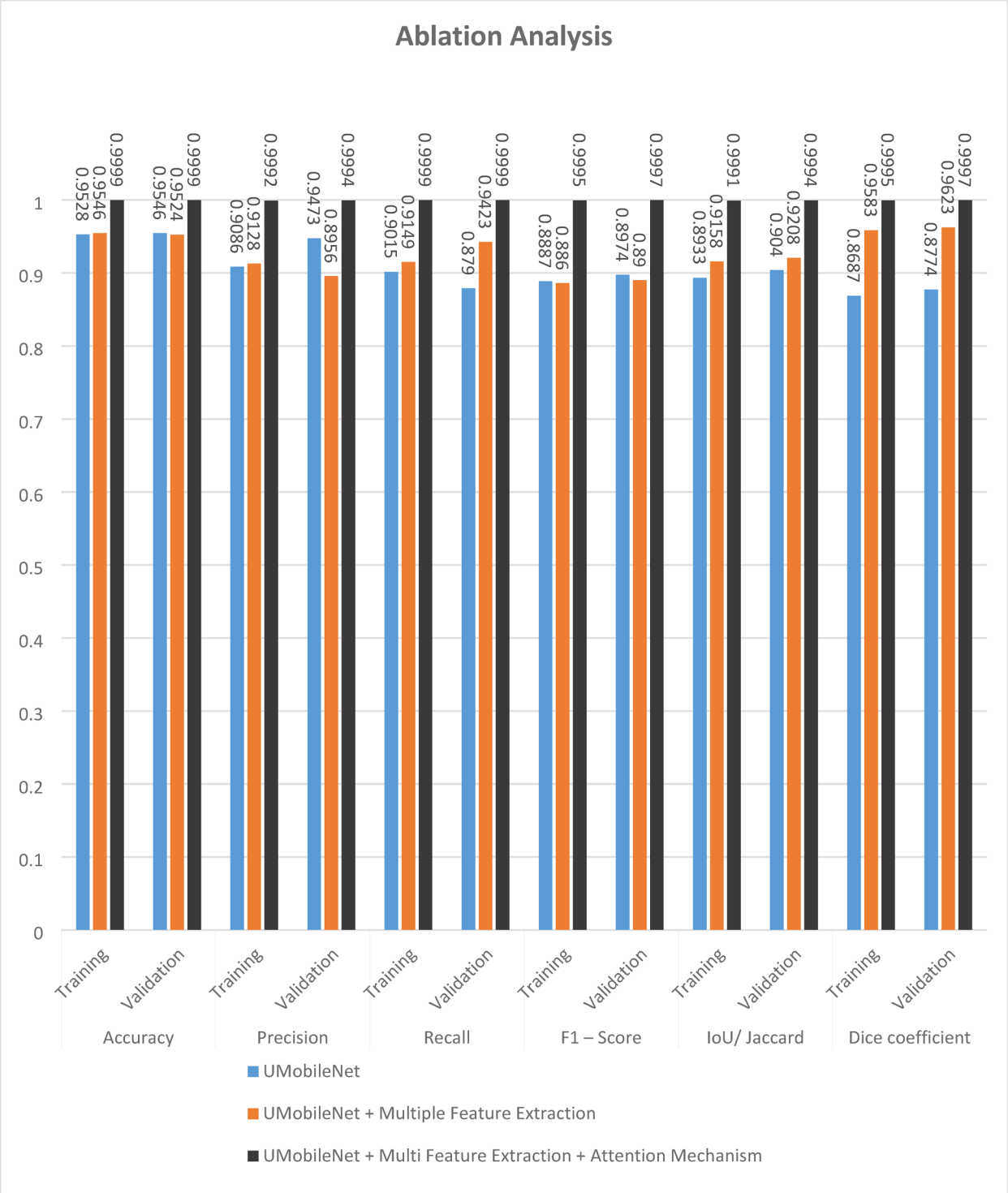

Fig. 13. Ablation analysis.

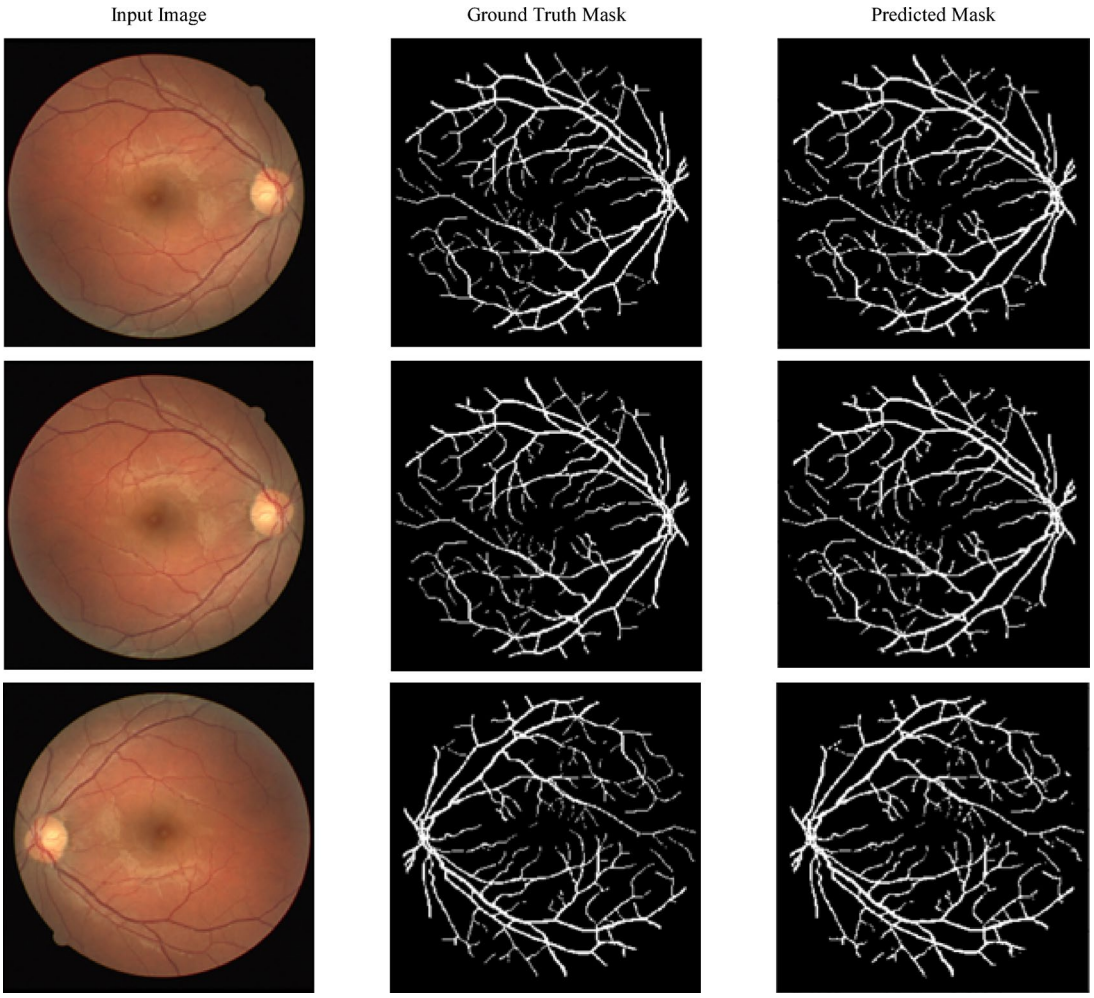

Fig. 14. Visual analysis.

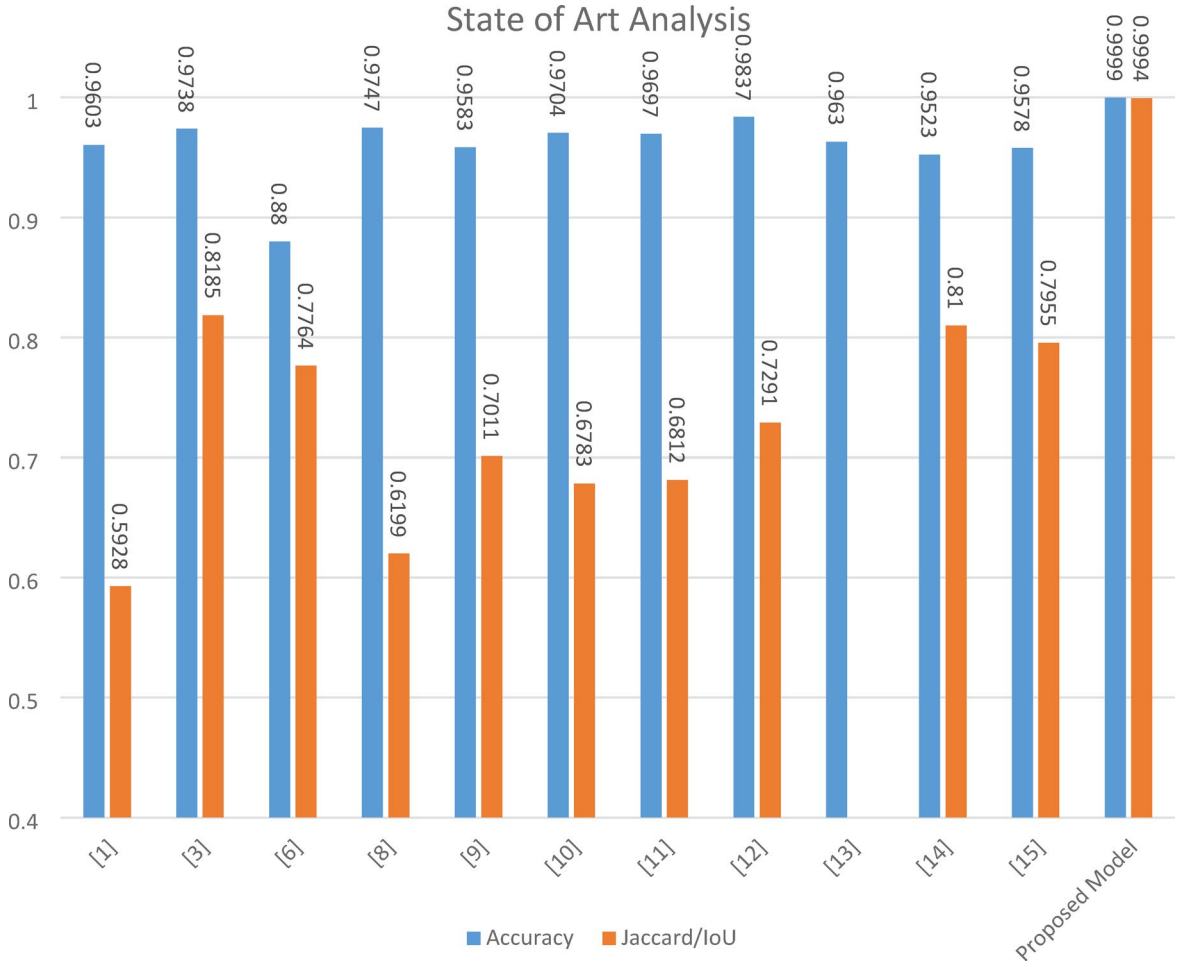

Fig. 15. State of art analysis.

Data availability

The dataset used in this study is publicly available at the following link: https://www.kaggle.com/datasets/andrewmvd/drive-digital-retinal-images-for-vessel-extraction.

Received: 21 May 2025; Accepted: 12 November 2025
Published online: 09 December 2025

References

1. Wang, Y. & Li, H. A Novel Single-Sample Retinal Vessel Segmentation Method Based on Grey Relational Analysis. *Sensors* **24**(13), 4326 (2024).
2. Javed, S., Khan, T. M., Qayyum, A., Sowmya, A. & Razzak, I. *Region guided attention network for retinal vessel segmentation*. ArXiv Preprint at https://*arXiv.org/abs/*240718970. (2024).
3. Xia, C. & Lv, J. MPCCN: a symmetry-based multi-scale position-aware cyclic convolutional network for retinal vessel segmentation. *Symmetry* **16**(9), 1189 (2024).
4. Giesser, S. D. et al. Evaluating the impact of retinal vessel segmentation metrics on retest reliability in a clinical setting: A comparative analysis using automorph. *Investig. Ophthalmol. Vis. Sci.* **65** (13), 24–24 (2024).
5. Jahan, N., Dhrubo, N. S. & Islam, M. R. April. U-Net Deep Learning Model for the Retinal Vessel Segmentation of Latest Imaging Technology OCTA Images. *In 2024 3rd International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)*. 1–5 (IEEE, 2024).
6. Keerthivasan, E. et al. April. Early Glaucoma Detection through ANSAN-Infused Retinal Vessel Segmentation. *In 2024 International Conference on Inventive Computation Technologies (ICICT)*. 1212–1218 (IEEE, 2024).
7. Ramezanzadeh, E. et al. *A high-accuracy segmentation hybrid method for retinal blood vessel detection in fluorescein angiography images of real diabetic retinopathy patients*. (2024).
8. Devi, Y. A. S. & Kamsali, M. C. Retinal blood vessel segmentation through morphology cascaded features and supervised learning: RETINAL BLOOD VESSEL SEGMENTATION THROUGH SUPERVISED LEARNING. *J. Sci. Industrial Res. (JSIR)*. **83** (3), 264–273 (2024).
9. Guo, C. et al. January. Sa-unet: Spatial attention u-net for retinal vessel segmentation. *In 2020 25th international conference on pattern recognition (ICPR)*. 1236–1242 (IEEE, 2021).
10. Wei, J. et al. Genetic U-Net: automatically designed deep networks for retinal vessel segmentation using a genetic algorithm. *IEEE Trans. Med. Imaging*. **41** (2), 292–307 (2021).

11. Lee, K., Sunwoo, L., Kim, T. & Lee, K. J. Spider U-Net: Incorporating inter-slice connectivity using LSTM for 3D blood vessel segmentation. *Appl. Sci.* **11**(5), 2014 (2021).
12. Ding, H., Cui, X., Chen, L. & Zhao, K. MRU-Net: a U-shaped network for retinal vessel segmentation. *Appl. Sci.* **10**(19), 6823 (2020).
13. Laibacher, T., Weyde, T. & Jalali, S. M2u-net: Effective and efficient retinal vessel segmentation for real-world applications. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. (2019).
14. Gu, Z. et al. Ce-net: context encoder network for 2d medical image segmentation. *IEEE Trans. Med. Imaging*. **38** (10), 2281–2292 (2019).
15. Islam, M. T. et al. LUVS-Net: A Lightweight U-Net Vessel Segmentor for Retinal Vasculature Detection in Fundus Images. *Electronics* **12**(8), 1786 (2023).
16. Cao, J., Chen, J., Gu, Y. & Liu, J. MFA-UNet: A vessel segmentation method based on multi-scale feature fusion and attention module. *Front. Neurosci.* **17**, 1249331 (2023).
17. Ross, T. Y. & Dollár, G. K. H. P. July. Focal loss for dense object detection. *In proceedings of the IEEE conference on computer vision and pattern recognition*. 2980–2988 (IEEE, 2017).

Acknowledgements

Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R138), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Author contributions

Tanishq Soni: Conceptualization, methodology design, software development, data curation, writing—original draft preparation.Sheifali Gupta: Investigation, validation, formal analysis, writing—review and editing. Salil Bharany: Supervision, project administration, resources, writing—review and editing.Ateeq Ur Rehman: Methodology enhancement, formal analysis, visualization, writing—review and editing.Rania M. Ghoniem: Validation, supervision, resources, critical revision of the manuscript.Belayneh Matebie Taye: Software testing, comparative analysis, manuscript proofreading, and cross-institutional collaboration support.

Declarations

Competing interests

The authors declare no competing interests.

Ethics approval

No animals or human subjects were involved in this study. The study utilized publicly available datasets, and all methods were carried out in accordance with relevant guidelines and regulations.

Additional information

Correspondence and requests for materials should be addressed to T.S. or B.M.T.