A Hybrid Attention-Guided U-Net++ Architecture for Diabetic Foot Ulcer Segmentation

Rajdeep Chatterjee
School Of Computer Engineering
Kalinga Institute of Industrial Technology, Deemed to be University (KIIT)
Bhubaneswar, India
rajdeepfcs@kiit.ac.in

Adrija Das
School Of Computer Engineering
Kalinga Institute of Industrial Technology, Deemed to be University (KIIT)
Bhubaneswar, India
2205267@kiit.ac.in

Swaralipi Samanta
School Of Computer Engineering
Kalinga Institute of Industrial Technology, Deemed to be University (KIIT)
Bhubaneswar, India
2205338@kiit.ac.in

Vivek Singh
School Of Computer Engineering
Kalinga Institute of Industrial Technology, Deemed to be University (KIIT)
Bhubaneswar, India
22052868@kiit.ac.in

Saumya Tayal
School Of Computer Engineering
Kalinga Institute of Industrial Technology, Deemed to be University (KIIT)
Bhubaneswar, India
2205326@kiit.ac.in

***Abstract*—Diabetic Foot Ulcers (DFUs) are one of the most common complications of diabetes, often leading to infection and even amputation if not properly monitored. It affects approximately 18.6 million people worldwide each year. Automated segmentation of DFUs from clinical images is essential for accurate wound assessment and timely treatment. However, obtaining such segmentation results manually is time-consuming and prone to inconsistency, highlighting the need for an efficient and accurate automated approach. Correct segmentation of Diabetic Foot Ulcers (DFUs) plays an important role in wound assessment, treatment planning, and monitoring of healing progress. Current deep learning models, such as U-Net and U-Net++, repeatedly struggle to predict fine ulcer boundaries and suppress irrelevant background details due to difficult or complex variations in wound appearance, size, and texture. To solve these challenges, this study presents a hybrid deep learning architecture known as AG-U-Net++, which combines the dense skip connections of U-Net++ with attention gate mechanisms. The attention gates selectively spread the needed boundary precision. The dataset used for training and evaluation of the model is the FUSeg. The dataset comprises 1210 annotated DFU images, utilizing a combination of binary cross-entropy and dice loss functions. The result of our experiments demonstrated that the model we proposed achieved a Dice score of 0.81, an IoU score of 0.73, and an accuracy of 99.91% which results in better performance with other architectures such as VGG16, MobileNetV2, EfficientNetB0, and ResNet34. The results ensure that the AG-U-Net++ model effectively and efficiently improves feature extraction, segmentation quality, and computational efficiency, making it acceptable for real-time DFU analysis and clinical wound monitoring applications.**

***Index Terms*—Attention Mechanism, Clinical Wound Analysis, Deep Learning, Diabetic Foot Ulcers, Convolutional Neural Networks, Segmentation**

I. Introduction

Diabetic individuals often develop foot ulcers due to poor blood flow and nerve damage. If these wounds are not taken care of in time, they may appear minor at first but could soon become serious. According to research, diabetic foot ulcers are often neglected by the health care sector [1]. Doctors are required to carefully examine the images of the wounds to treat them appropriately and monitor their healing. However, obtaining consistent results in hospitals or large screening programs is challenging because doing this by hand takes a considerable amount of time and can vary depending on who is performing the task.

In recent years, deep learning has become a powerful approach in medical image segmentation [2], making it possible to identify and outline diseased areas with high performance and minimal manual effort. However, segmenting diabetic foot ulcers (DFUs) remains a tough task because real clinical images often differ mainly in terms of ulcer shape, colour, texture, and lighting conditions.

II. Motivation:

Accurate segmentation of diabetic foot ulcers is very important in all wound assessments aided by computers, making the doctor's job easier. However, the existing Deep Learning architectures like the U-Net, U-Net++ often face many limitations even when combined with models like VGG16, MobileNetV2, EfficientNetB0 and ResNet34. The DFU datasets contain various colours, wound sizes, and background textures, making boundary decisions very difficult. These challenges motivated us to develop a more refined segmentation architecture that is capable of enhancing the boundary precision and focusing on the more relevant areas of interest.

While developing the model, we were inspired by:
1. U-Net - provides a strong encoder-decoder foundation, but it faces difficulties in preserving the fine boundaries.
2. U-Net++ - Introduces nested deep skip Connections, which improves the multi-scale feature, but still the background redundant information is transmitted.
3. Attention U-Net: it has a spatial attention mechanism, it allows selective focus but lacks the multi-level feature aggregation

To overcome the limitations and use the relevant features, we designed a model that combines the dense Connectivity of the U-Net++ architecture and the selective focus capability of the attention gates. Our main aim was not only to create a hybrid model (U-Net++ & Attention U-Net) that not only uses the information gathered from the various levels of contextual understanding but also helps suppress the irrelevant background features, which ensures the accurate identification and mapping of the particular boundary of the ulcer.

III. Background Concepts:

A. Overview:

The proposed AG-U-Net++ model is inherently based on the basic U-Net++ architecture, which in itself is an enhanced version of the U-Net architecture. Before we get into U-Net++, let us go over the basic U-Net structure itself. The U-Net architecture comes from the basic convolutional neural network. A simple convolution architecture comprises three important elements: convolutional layers, pooling layers and fully connected layers. The convolution layer applies a filter to the input images, thus enabling feature extraction. The pooling layers reduce the spatial dimension of the feature maps in order to reduce complexity, and the fully connected layers integrate the extracted features for classification or regression.

B. U-Net Architecture:

The U-Net architecture,introduced by Ronneberger et al.(2015) for biomedical image segmentation [3] became very widely adopted due to its symmetric design and skip connections which preserves fine-grained spatial details. The structure gets its name from its U-shaped structure of the architecture which consists of the encoder and decoder. The encoder also known as the contracting path captures the features and context of the image using a repetition of convolutional layers and max pooling layers. As the image passes through the encoder, the number of feature channels increases while the spatial resolution decreases. The second component decoder, also known as the expansive path, reconstructs the image through up-sampling operations. Here, the size of the image gradually increases and the depth slowly decreases. A unique feature of U-Net is the concept of skip connection, which directly links corresponding layers in the encoder and decoder paths. It transfers the feature representations from each encoder layer to its corresponding decoder layers to battle the loss of spatial details. These feature maps are then concatenated with the up-sampled decoder features. This enables the network to generate better output with sharper segmentation boundaries.

C. UNet++:

UNet++ is an enhanced version of the U-Net architecture introduced in the research paper "UNet++:A Nested Architecture for Medical Image Segmentation" [4]. The unique feature of UNet++ is the dense skip connections and deep supervision. The main idea was to reduce the gap between the resolution and feature abstraction of the images. Unlike the U-Net architecture, where the encoder and decoder are connected directly, the UNet++ architecture introduces a series of convolutional blocks between them to form nested skip pathways. These blocks refine the features of the encoders before connecting to the features of the decoder. Additionally, UNet++ supports deep supervision, where outputs of multiple decoder layers can be used to compute the segmentation loss.

The model that we are proposing is built on the fundamental concepts of UNet++, along with the implementation of attention gates to enhance the feature learning process and focus on the most important regions of the image, which in our case are the ulcers on the foot. The attention gates integrated into the skip connections further improve the segmentation accuracy. The attention gates are designed to automatically learn which regions of the image are most important to the target structure, enabling the network to focus on the most relevant features while suppressing irrelevant and noisy background information.

In the context of foot ulcer image segmentation, the attention gates improve the ability to identify and segment ulcer regions accurately which is also observed in previous research [5], even when they appear with irregular shapes, sizes, varying backgrounds, discoloured skin and lighting variations.

D. Evaluation Metrics:

In order to evaluate the performance of our model, we used the following performance metrics:

The Intersection over Union(IoU), also known as the Jaccard Index, measures the overlap percentage between the ground truth mask and the predicted mask. It is defined as the ratio of the intersection to the union of the two regions. It is expressed by:

$$IoU = \frac{|PredictedMask \cap GroundTruthMask|}{|PredictedMask \cup GroundTruthMask|} \tag{1}$$

The other evaluation metrics used are the Dice Similarity Coefficient(DSC) or simply the Dice score. It measures the harmonic mean of precision and recall, emphasizing the overlap between predicted truths and ground truths. It is expressed by:

$$Dice = \frac{2(|PredictedMask \cap GroundTruthMask|)}{|PredictedMask| + |GroundTruthMask|} \tag{2}$$

IV. Proposed Model:

A. Overview:

The proposed architecture is a hybrid encoder-decoder convolutional network, which is designed to improve the Segmentation IoU matrices for complex medical images like the diabetic foot Ulcer. The model extends the traditional U-Net++ architecture, which in itself is an enhanced version of U-Net."The U-Net architecture achieves very good performance on very different biomedical segmentation applications" [3] by integrating attention gates into the dense skip connection pathways, which allows the network to focus on the foot ulcer more effectively.

B. Encoder Network:

The encoder consists of five hierarchical stages that help extract the main features (down-sampling). Following previous work [6], we apply a convolution - Batch Normalization - ReLU activation function in each stage, these are which is applied twice, followed by the MaxPooling to downsample the spatial dimensions by doubling the feature maps. The convolutional filters vary from 32 to 512 across the encoder, which helps the model capture low-level textures and high-level semantics easily and effectively.

C. Attention Gate Mechanism:

From previous research, we know the huge success of the attention mechanism in many visual classifications and segmentations. [7].In the field of medical image segmentation, attention mechanisms have become a crucial tool in directing the model's focus towards important regions of the image [8], [9]. These were integrated into each skip connection layer between the encoder and the decoder, which transmits the important features. This ensures that only the important features are transmitted to the decoder, which improves the segmentation around the irregular edges of the wound.

D. Decoder Network:

The decoder reconstructs the segmentation mask through a series of up-sampling and convolutional blocks [6]. In each stage, the upsampled decoder feature maps and the corresponding attention-refined encoder features are concatenated. This helps the model to combine high-resolution spatial information with very deep semantic context.

E. Dense skip Connections:

Inspired by the U-Net++ architecture, dense skip connection [10] pathways are established between the encoders and the decoders. These nested connections help us improve the multi-scale feature reusability and consistent gradient flow during the entire training. This allows intermediate feature aggregation, which improves generalization and converges stability. In the proposed model, dense connections pass through the attention gates before concatenation.

F. Output Layer:

The final layer employs a 1*1 convolution, which is followed by a sigmoid activation function, which helps us to produce a binary segmentation map which represents the ulcer and non-ulcer regions.

G. Loss Function:

The network uses a hybrid loss, which combines binary cross-entropy [11] and dice loss [12]. This combination ensures both pixel-wise accuracy and region-level overlap consistency, which is very important for medical image segmentation with an imbalanced class.

V. Model Architecture:

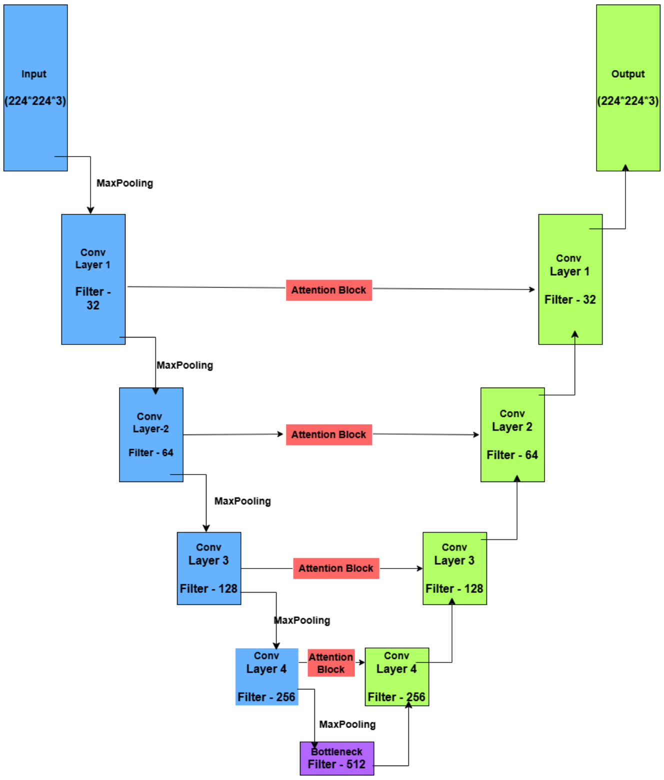

Fig. 1. AG-U-Net++ Model Architecture

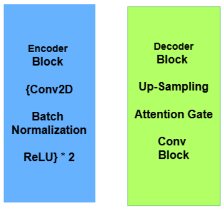

Fig. 2. Encoder and Decoder Blocks

As shown in the diagram, the proposed mode (AG-U-Net++) architecture consists of an encoder-decoder framework with attention-enhanced dense skip-connections. The blue blocks represent the encoder layers (convolution - batch normalization - ReLU activation function and max pooling). The green blocks refer to the decoder layers (UpSampling - Attention Gate Integration - Batch Normalization). The purple block represents a bottleneck. It captures the deepest semantic representation before the decoder progressively reconstructs the segmentation map.

1) Qualitative Comparison Analysis:: To further validate the effectiveness of the proposed model, a qualitative analysis of the segmentation results was performed across different models The image shows the Original Validation Image, Ground Truth and the Predicted Mask of the different models (Proposed Model, VGG16, MobileNetV2, ResNet34, EfficientNetB0 Models)

VI. Experiments and Result Analysis:

A. Dataset Used:

1) Data Source: The dataset I used was taken from the Foot Ulcer Segmentation (FUSeg) Challenge was originally presented by the AZH Wound and Vascular Centre, Milwaukee, USA [13]. It contains 1210 clinical images. All the images were of 512 * 512 pixels. The dataset is divided into 810 training, 200 validation images and 200 testing images.

2) Annotation Details:: The dataset includes manually annotated binary segmentation masks which show the wounds and the non-wound regions.

B. System:

For training the model, we used Kaggle notebooks. It provides a powerful cloud-based environment that is very good for deep learning and data science problems. We have used the GPU-P100 accelerator. It has 13 GB RAM and a disk space of about 20 GB. We used the Adam optimizer and a learning rate of 1e-4. This was selected to ensure gradual convergence and to prevent oscillations or divergence during training. The loss function was binary cross-entropy and dice loss combined. For evaluation, we have used the IoU score and the Dice Coefficient. The experiments were carried out for 50 epochs and at a batch size of 2.

C. Result Analysis and Discussion

Accuracy alone may give an overestimated result because most pixels in medical images are background (non-ulcer) while dice is more robust for medical segmentation due to foreground sparsity [14]. A model might show high accuracy but low IoU and Dice, which means that it correctly identifies most of the background pixels but poorly segments the ulcer. Thus, relying on IoU and Dice ensures that the evaluation reflects how well the model detects the actual ulcer area, not just the overall image.

1) Training Performance:

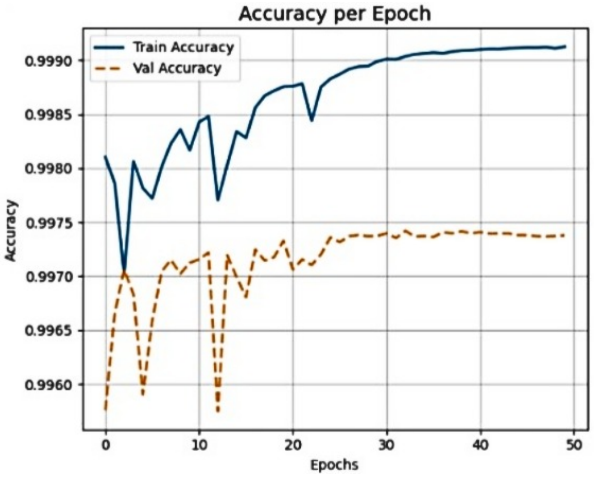

Fig. 3. Accuracy Per Epoch

Accuracy: Both the training and the validation accuracy had reached approximately 99% after completing the training.

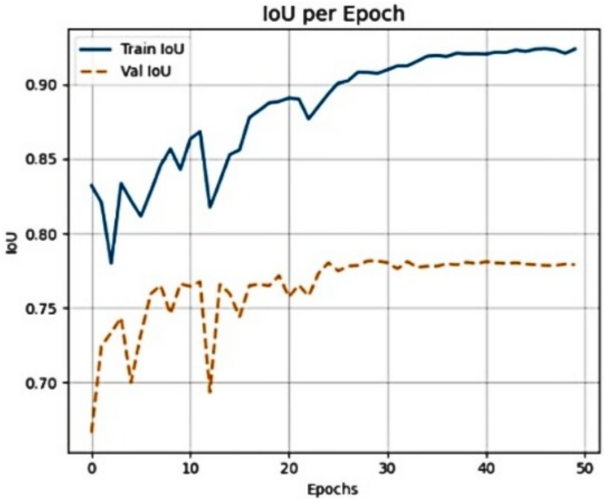

Fig. 4. IoU Per Epoch

IoU: The validation IoU increased steadily and reached about 0.73, which indicates a precise overlap between the predicted and actual ulcer regions.

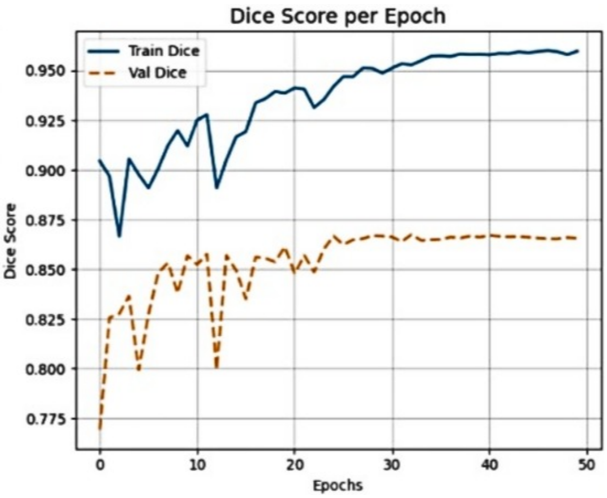

Fig. 5. Dice Score Per Epoch

Dice Score: The Dice coefficient achieved was 0.79, which demonstrated that the model can detect ulcers effectively.

2) Qualitative Comparisons:: Qualitative Comparison: represents visual segmentation results across different models, showing that the proposed AG-U-Net++ achieves more accurate and complete wound masks compared to existing architectures such as VGG16, MobileNetV2, ResNet34, and EfficientNetB0.

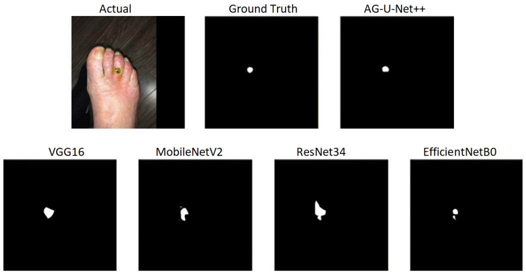

Fig. 6. Sample 1

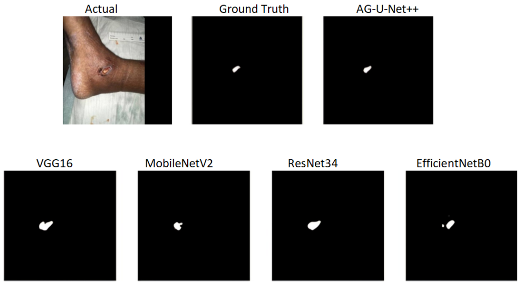

Fig. 7. Sample 2

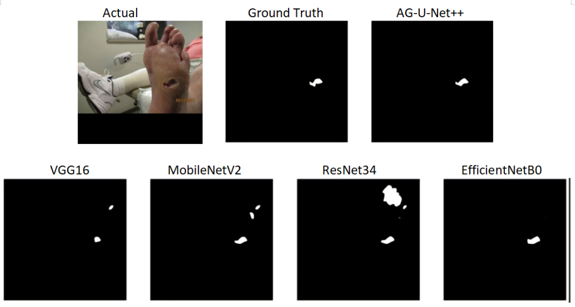

Fig. 8. Sample 3

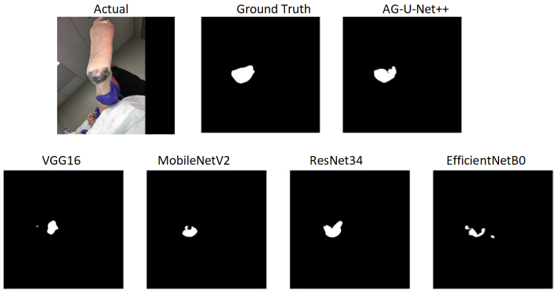

Fig. 9. Sample 4

In example 1 (Fig. 5): The image shows that our model is the closest to the ground truth mask, and the other models did not perform that well. The Ulcer Boundaries are clearly defined. Unlike the other models, which fail to capture the complete ulcer region.

In example 2 (Fig. 6): This is another image that shows our model accurately predicted the mask, whereas VGG16 and the MobileNetV2 miss many crucial boundaries, while ResNet34 and EfficientNet display false detection of the non-ulcer areas.

In example 3 (Fig. 7): In this image, the model shows almost perfect segmentation; the ground truth and the mask overlap almost completely. Other models produce distorted ulcer masks.

In example 4 (Fig. 8): In this image, we can see that our model shows the accurate segmentation, while the other models fail to capture the Ulcers correctly.

3) Quantitative Comparisons:: To measure the performance of the proposed model (AG-U-Net++), we performed a detailed comparison. We trained models like VGG16. MobileNetV2, ResNet34, EfficientNetB0, where each model was used with the baseline of the encoder-decoder architecture. Each model was evaluated for 50 epochs, learning rate 0.0002, the Adam optimizer and the same dataset.

Based on the other performance, we can conclude that: Our model achieved the highest IoU Score (0.73), which shows improved overlap between the predicted and the ground truth masks. We achieved a Dice Score of 0.81, which proved that the model can detect ulcers effectively. We also achieved an accuracy of 99.91% which is higher than the other models used. Although MobileNetV2 was trained much faster than our model, the overall accuracy is lower than our model. The proposed model achieves a balanced trade-off between accuracy and speed.

TABLE I
COMPARISON OF MODEL PERFORMANCE METRICS

Model	**IoU Score**	**Dice Score**	**Accuracy**	**Time to Train (s)**
AG-U-Net++	0.73	0.81	0.9991	1433
VGG16	0.70	0.79	0.9961	3355
MobileNet-V2	0.59	0.69	0.9955	453
EfficientNet-B0	0.64	0.74	0.9963	679
ResNet-34	0.62	0.72	0.9958	1760
U-Net++	0.68	0.77	0.9987	1303
Attention U-Net	0.56	0.64	0.9981	797

The Proposed model performs better in both quantitative metrics and computational efficiency. This makes it highly suitable for real-time medical image segmentation tasks, such as Foot Ulcer Detection.

ACKNOWLEDGMENT

We gratefully acknowledge the organizers of the Foot Ulcer Segmentation (FUSeg) Challenge for providing access to the dataset used in this study. We also express our sincere appreciation to our mentor, Dr. Rajdeep Chatterjee at Kalinga Institute of Industrial Technology, deemed to be University (KIIT-DU), for his valuable guidance, encouragement, and constructive feedback throughout the course of this research. We also acknowledge the support provided by the Department of Computer Science and Engineering for access to computational facilities and technical resources that helped us successfully implement and evaluate the proposed model.

REFERENCES

[1] W. J. Jeffcoate and K. G. Harding, "Diabetic foot ulcers," *The Lancet*, vol. 361, no. 9368, pp. 1545–1551, 2003.

[2] X. Liu, L. Song, S. Liu, and Y. Zhang, "A review of deep-learning-based medical image segmentation methods," *Sustainability*, vol. 13, no. 3, p. 1224, 2021.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Cham: Springer, 2015, pp. 234–241.

[4] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Proc. Int. Workshop Deep Learning in Medical Image Analysis*. Cham: Springer, 2018, pp. 3–11.

[5] Z. Zhu, Y. Yan, R. Xu, Y. Zi, and J. Wang, "Attention-unet: A deep learning approach for fast and accurate segmentation in medical imaging," *J. Comput. Sci. Softw. Appl.*, vol. 2, no. 4, pp. 24–31, 2022.

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2018.

[7] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, "Attention mechanisms in computer vision: A survey," *Comput. Visual Media*, vol. 8, no. 3, pp. 331–368, 2022.

[8] S. Umirzakova, S. Ahmad, L. U. Khan, and T. Whangbo, "Medical image super-resolution for smart healthcare applications: A comprehensive survey," *Inf. Fusion*, vol. 103, p. 102075, 2024.

[9] S. Saifullah, R. Dreżewski, A. Yudhana, M. Wielgosz, and W. Caesarendra, "Modified u-net with attention gate for enhanced automated brain tumor segmentation," *Neural Comput. Appl.*, vol. 37, no. 7, pp. 5521–5558, 2025.

[10] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*, 2017, pp. 4799–4807.

[11] U. Ruby and V. Yendapalli, "Binary cross entropy with deep learning technique for image classification," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 10, 2020.

[12] R. Zhao, B. Qian, X. Zhang, Y. Li, R. Wei, Y. Liu, and Y. Pan, "Rethinking dice loss for medical image segmentation," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*. IEEE, 2020, pp. 851–860.

[13] C. Wang, A. Mahbod, I. Ellinger, A. Galdran, S. Gopalakrishnan, J. Niezgoda, and Z. Yu, "Fuseg: The foot ulcer segmentation challenge," *Information*, vol. 15, no. 3, p. 140, 2024.

[14] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. Int. Conf. 3D Vision (3DV)*. IEEE, 2016, pp. 565–571.