IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 August 2023, accepted 22 August 2023, date of publication 28 August 2023, date of current version 7 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3309158

RESEARCH ARTICLE

EMED-UNet: An Efficient Multi-Encoder-Decoder Based UNet for Medical Image Segmentation

KASHISH D. SHAH[1], DHAVAL K. PATEL[1], (Member, IEEE), MINESH P. THAKER[2], (Member, IEEE), HARSH A. PATEL[1], MANOB JYOTI SAIKIA[3], (Member, IEEE), AND BRYAN J. RANGER[4], (Member, IEEE)

[1]School of Engineering and Applied Science, Ahmedabad University, Ahmedabad 380009, India
[2]Indus Institute of Technology and Engineering, Indus University, Ahmedabad 382115, India
[3]Department of Electrical Engineering, University of North Florida, Jacksonville, FL 32224, USA
[4]Boston College, Chestnut Hill, Newton, MA 02467, USA

Corresponding author: Kashish D. Shah (kashish.s2@ahduni.edu.in)

ABSTRACT Many current and state-of-the-art deep learning models for accurate image segmentation are based on the U-Net architecture, a convolutional neural network designed for biomedical applications. Despite its widespread adoption in the medical imaging community, U-Net has two major limitations. First, due to its deep structure and the large number of filters used, the number of parameters and Floating-Point Operations Per Second (FLOPS) are high. This results in high computational complexity and demands a large memory size, making real-time implementation and deployment of U-Net models challenging. Second, the base U-Net model only uses a single kernel type (i.e., 3×3) throughout the network for all convolution operations. Feature extraction using a single spatial extent throughout the network is suitable if the size, location, and shape of the salient regions remain static in all the dataset images, which is not necessarily the case in medical imaging. To address these two limitations, we propose an Efficient Multi-Encoder-Decoder based UNet (EMED-UNet), a novel architecture for efficient medical image segmentation. We evaluated our network on four medical imaging datasets: Montgomery County, Shenzhen CXR, COVID-19 CT LS, and the BraTS (Brain Tumor Segmentation) dataset. EMED-UNet outperforms U-Net and its variants in terms of accuracy, with around 77% reduction in parameters, a 60% reduction in FLOPS, and a 79.2% reduction in memory usage (all as compared to U-Net). The results demonstrate that EMED-UNet is a lightweight and accurate model for image segmentation that substantially improves upon the U-Net base model and is more feasible to deploy, given its decreased computational cost.

INDEX TERMS Deep learning, U-Net, bottleneck layer, semantic segmentation, medical image segmentation, encoder-decoder networks.

I. INTRODUCTION

A. BACKGROUND

Medical image segmentation involves partitioning an image into multiple regions based on the spatial and pixel-level semantic information of the image. The shape, size, and location of the salient features in medical images often vary depending on the dataset, making it challenging to train a model to perform the semantic segmentation task. Deep learning (DL)-based medical image segmentation has addressed some of these challenges and is useful in many clinical applications, such as identifying or isolating an anatomical structure, diagnosis and treatment planning in radiology, image-guided interventions, and surgeries, quantitative analysis, computer-aided diagnosis, and many more. With that goal, Li et al. proposed a deep convolution neural network (CNN)-based framework for liver tumor segmentation [1], which was expanded upon with further computed tomography (CT) studies by Vivanti et al. [2]. In 2014, Menze et al. developed the multi-modal brain

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

tumor segmentation benchmark [3], and in 2017 targeting the same challenge, Cherukuri et al. demonstrated segmentation to postoperative hydrocephalic scans [4]. Long et al. [5] developed fully convolutional networks for semantic segmentation.

Recent advancements in medical image segmentation are formulated on encoder-decoder-based DL architectures [6], [7]. Such DL approaches have been very effective in learning the spatial and semantic information from the image. However, DL models with denser layers often lose low-level spatial information and tend to learn more complex high-level semantic information. This challenge is addressed by encoder-decoder-based models, which we explore further in this paper and is further improved using our proposed architecture.

B. LITERATURE REVIEW

1) U-NET AND ITS VARIANTS

The U-Net architecture was proposed by Ronneberger et al. in 2015 [8]. A detailed review of U-Net and its variants was published by Siddique et al. in 2021 that outlines the theory and current applications of these methods [9]. A brief summary, as relevant to our proposed EMED-UNet architecture, is presented here.

The U-Net architecture contains an encoder (contracting path) and a decoder (expanding path). U-Net has been very successful in part due to the incorporation of skip connections [10], [11], [12], [13], which can retrieve full-scale spatial resolution from the encoder side and merge them with the semantic and coarser features at the decoder side.

Since 2015, numerous variants of U-Net have emerged. For example, Res-UNet [14], Dense-UNet [15], UNet++ [16], and UNet3+ [17] were developed to improve segmentation performance compared to the base U-Net model. The U-Net++ network architecture proposed by Zhao et al. [12] solved the problem of finding the optimal depth of U-Net for the best result and also provided a redesign of the skip connections for the flexible fusion of features, unlike the restrictive fusion of the same-scale feature maps of U-Net. UNet++ was further improved by Lei et al. [18] for congestive heart failure diagnosis. Later, Zou et al. employed a Multi-scale Residual UNet++ [19] for the same purpose. Implementation of the U-Net architecture for 3D image segmentation was developed as 3D U-Net [12], where all operations and functions, including convolution, max-pooling, and up-sampling, were performed in 3D. Adding attention-based learning, Oktay et al. proposed the Attention U-Net [20], which gave the model the ability to focus only on certain key features. Further, attention-gated networks [21] were introduced to leverage salient regions of medical images.

Recently, COVID TV-UNet was proposed by Saeedizadeh et al. [22] in 2021 for segmenting COVID-19 chest CT images. It was further extended by Saeedizadeh et al. for retinal pigment epithelium layer detection [23] in order to develop a shape-preserving retinal optical CT image alignment method.

Introducing transformers and HarDNet Structures, Shen and Xu proposed a dual-encoder image segmentation network [24], solving the limitation of losing the interaction of inter-channel information during the upsampling of the decoder. Thus, they brought more strength to the local connection between the adjacent chunks. In [25], Chen et al. proposed an unsupervised tilewise autoencoder (T-AE) pretraining architecture for transferable knowledge. They also proposed a sub-framework based on a reconstruction network regularized segmentation model for fine-tuning.

Substantial work has been done with COVID-19 CT imaging in segmenting lesions and detecting lung infection from the CT scan. Das, in 2021, proposed an Adaptive Activation Function-based U-Net model [26] and ensemble learning with the CNN deep features with a novel approach called EL-CNN-DF. In 2022, the Cascaded 3D U-Net [27] concept was presented by Aswathy et al. working on the COVID-19 CT segmentation task. They proposed a cascade of two 3D U-Nets, both fulfilling different purposes. The first 3D U-Net extracts the lung parenchyma from the CT volume and the second works on extracting the infected 3D modules. Malik et al. in [28] proposed CDC-Net, a multi-classification CNN model for the detection of chest-related infections of COVID-19, pneumothorax, pneumonia, lung cancer, and tuberculosis. The proposed architecture involved residual connections and dilated convolutions. Yin et al. proposed another segmentation framework: SD-UNet [29] targeting CT imaging modality to detect lung infections. They modified the U-Net with Squeeze-and-Attention and dense ASPP (atrous spatial pyramid pooling) modules in order to fuse the global and multi-scale features.

Later came the concept of deeply learned vectors' formation [30], proposed by Naeem et al., mainly based on implementing a CNN with auto-correlation, gradient computation, scaling, filter, and localization coupled with state-of-the-art content-based image retrieval methods.

2) WORKS ON EFFICIENCY

To decrease computational costs, many research groups have focused on developing lightweight networks. These networks aim to make the models deployable on resource-constrained hardware and for real-time applications. For example, InceptionNet proposed by Szegedy et al. [31] has a bottleneck layer (i.e., 1×1 convolutions) which is used for dimensional reduction to decrease the number of parameters significantly. In this network, a kernel of the same size could not be used in every convolution throughout the network if the salient regions in every image vary in sizes. Therefore, they showcased the importance of using different spatial extents (i.e. kernel sizes) in the convolution layers. Based on this network, Singh et al. developed Inception U-Net [32] with the aim of automating the selection of the numerous layers in the U-Net architecture. It contains the inception layers in

the nodes of the standard U-Net. Cahall et al. [33] expanded on this to use the dilated convolution in the Inception U-Net to capture both the local structural as well as global contextual information and proposed the Dilated Inception U-Net. Following that, Zhang et al. [34] proposed the Dense Inception U-Net integrating the Inception-Res module and the densely connected convolution module. In addition, Beheshti and Johnsson proposed a lightweight U-Net, the ''Squeeze U-Net'' [35] in 2020, with reduced memory requirement and improved efficiency. As another approach towards creating lightweight models for medical image segmentation came the knowledge distillation technique [36]. They used a technique of distilling knowledge from another well-trained segmentation network. Finally, in 2023, He et al. in [37] proposed an efficient hierarchical hybrid vision Transformer (H2Former), getting the aggregating potential of CNNs, multi-scale channel attention, and Transformers. More transformer-based Medical Image Analysis techniques are elaborated in [38]. This states that the transformers can play a good role in enhancing the current state-of-the-art techniques developed for medical image analysis.

C. OUR PREVIOUS WORK

We presented the first iteration of EMED-UNet as part of the IEEE Region-10 Symposium (TENSYMP) 2022 Conference [39]. To strengthen that preliminary work, we have further improved our model architecture, incorporated a deep supervision technique in our model for performance improvement, and validated the EMED-UNet model with additional datasets. In addition, instead of using large 5×5 or 7×7 convolutions, we used dilated 3×3 convolutions to reduce the number of trainable parameters and FLOPS in the model. In our initial work, we also observed that certain mid-level features need to be extracted before passing the images into the EMED-UNet network. Therefore, two convolution layers were added as pre-processing steps before the input layers of the EMED-UNet network. We also improvised the model accuracy by applying Deep Supervision to the EMED-UNet model.

D. MOTIVATION AND CONTRIBUTIONS

Despite its wide adoption in the medical imaging community, the U-Net architecture [8] has significant limitations, particularly when it comes to the need to deploy models in real-time systems such as mobile devices. Though U-Net and its variants have shown high accuracy, the primary challenge is that the U-Net models generally have complex neural architectures that employ a large number of filters and require a high amount of trainable parameters and many Floating Point Operations Per Second (FLOPS). This results in a high computational cost and requires a significant amount of onboard memory.

Another challenge faced by the U-Net architecture is that it limits the type and size of the kernel used for the convolutions in the architecture. Specifically, there is only one type of kernel used throughout the U-Net architecture in all the convolutions: the standard 3×3 kernel. Using only one kernel type throughout the model is suitable if the size and shape of the salient regions in all the images in the dataset remain static, which is generally not in the case of medical imaging. In addition, as there is only one encoder-decoder pair in the U-Net model with only one outlet to collect the output, a single effective receptive field is formed over the input image. Due to this, the feature extraction happens at a single effective receptive field generated over the input at the end of the network.

To address these limitations, we present the EMED-UNet network architecture. The proposed architecture consists of an Efficient Feature Extraction (EFE) module, which is employed at each encoder and decoder in the network. The EFE module reduces the number of parameters at the bottleneck layers, thereby reducing overall computational complexity and memory usage. The module consists of dilated convolution layers, which increase the receptive field without increasing the kernel size. The design of EMED-UNet consists of multiple encoders and decoders (i.e. multiple U-Nets), all capturing information at different receptive fields. A specific encoder-decoder pair (i.e. one U-Net) focuses on capturing the feature maps at a specific receptive field during convolution in the EFE module. Therefore, using multiple encoder-decoder pairs, EMED-UNet can capture features at multiple receptive fields. This allows multiple gradients to back-propagate simultaneously and multiple outputs to collect at the end. Furthermore, we implemented a deep supervision technique in the EMED-UNet to enable a collaborative learning experience between different U-Nets. Inspired by the UNet++ [16] proposed by Zhou et al., we redesigned the skip connections to achieve a flexible fusion of the saliency maps extracted by different U-Nets at varying resolution levels in the EMED-UNet.

We evaluated the EMED-UNet architecture on four medical imaging datasets: Montgomery County (MC), Shenzhen CXR dataset, COVID-19 CT lesion segmentation dataset, and the Brain Tumor Segmentation (BraTS) dataset. Our results demonstrate that the proposed model is able to match, or in some cases, outperform, the accuracy of U-Net while significantly reducing the number of parameters, FLOPS, and memory usage. With our EMED-UNet network architecture, we make the following four key contributions:

- We propose the EMED-UNet architecture, which contains multiple U-Nets, each capturing information at different receptive fields. Thus, the model can learn features of multiple receptive fields and multiple segmented outputs (one from each U-Net) and can be collected at the end. A deep supervision technique is also incorporated to enable collaborative learning among multiple U-Nets.
- To reduce computational complexity and memory usage, we introduce an Efficient Feature Extractor (EFE) module. The EFE contains dilated convolutions that increase the receptive field, per the network's requirement,

TABLE 1. Effect of 3 × 3 and 5 × 5 kernels on the U-Net architecture performance.

Kernel size	Dataset	Dice Loss	Dice Coefficient	IoU	Precision
5 × 5	Montgomery County	**0.0757**	**0.9243**	**0.8601**	**0.9752**
3 × 3	Montgomery County	0.0840	0.9160	0.8462	0.9733
5 × 5	Shenzhen Dataset	0.0630	0.9370	0.8825	0.9655
3 × 3	Shenzhen Dataset	**0.0577**	**0.9423**	**0.8917**	**0.9605**

without a significant increase in the number of parameters. It also contains multiple convolutional bottleneck layers, which contribute to parameter reduction.

- We evaluated the EMED-UNet network architecture on four different medical image segmentation datasets and proved its proficiency over the standard U-Net and its variants in terms of reduced computational complexity and memory size.
- We present two studies to independently showcase the strength of the EMED-UNet network and the EFE module. At first, we removed the EFE module and added the ResNet backbone to study the strength of the EFE module. In the second study, we studied the effects of the ResNet backbone on the proposed network and the U-Net network, comparing the learning abilities of both networks. Hence, studying the strength of the network.

The upcoming sections of this paper are organized as follows: In Section II, the EMED-UNet Network Architecture in multiple subsections, covering the motivation, the proposed module, technical specifications of the proposed EMED-UNet architecture, network connectivity, and the deep supervision enhanced collaborative learning. Datasets, Experiments, Results, and the two ablation studies are presented in Section III. Section IV provides the inferences from the work and possible future works. Conclusions are discussed in Section V.

II. PROPOSED ARCHITECTURE

A. MOTIVATION BEHIND THE NEW ARCHITECTURE

The base U-Net model contains a large number of filters in each of the convolution layers. Due to this, the total number of parameters in the model increases by a considerable amount (i.e. the number of parameters in the U-Net model goes over 31 million). A large number of filters demands a substantially high number of matrix multiplications, resulting in high FLOPS. Specifically, there are 386 G FLOPS in the base U-Net model, and the overall model size of the architecture is 373 MB.

Any CNN model, particularly those developed to analyze medical images, should be created such that its receptive field can cover the entire salient region of the input image. In medical imaging datasets that contain images from various modalities, the salient regions of the input images are of varying sizes. Therefore, the receptive field should be such that all the salient regions are covered. This requires multiple sets of kernels in the convolution layers of the models, all with different receptive fields. However, the base U-Net model carries a single type of kernel in its entire encoder-decoder structure, which is not feasible if the images in the dataset have salient parts of varying size and shape and are located in different spatial locations. Therefore, a deep learning model that can extract feature maps at multiple receptive fields is desirable.

We carried out a study to analyze the effect of different types of kernels in the convolution box of U-Net architecture. Two types of kernels were applied in the convolution box of the base U-Net architecture: 3 × 3 and 5 × 5. Table 1 shows that the U-Net (5 × 5) performed better when the experiment was conducted using the Shenzhen dataset, but U-Net (3 × 3) outperformed it when the experiment was conducted on the Montgomery County dataset. From Table 1, we conclude that in some cases, the U-Net with the 3 × 3 kernels work better, and in some, the 5 × 5 one. Thus, exactly what kernel to use depends on the resolution of the images, along with the type of latent space to be learned. This demonstrates that a model created for the medical imaging segmentation task should be able to focus on multiple receptive fields. A fixed kernel convolution throughout the network cannot learn all types of salient regions if the size and location of salient features in the image are not common in all the images.

The challenges described above motivated our development of an efficient architecture that can learn all types of features based on multiple receptive fields during the convolution operation while reducing computational complexity and memory usage.

B. EFE MODULE

We propose an EFE module to improve memory efficiency and computational complexity in the EMED-UNet architecture. As shown in Fig. 2, the EFE module consists of 1 × 1 convolutions, 3 × 3 dilated convolutions (with varying dilation rate), and a 2 × 2 MaxPooling2D layer. The 1 × 1 convolution reduces the depth dimension of the feature maps and filters only the channels suitable for the next 3 × 3 convolution to apply on. The purpose of using dilated 3 × 3 convolutions is to increase the receptive field during the convolution operation. For example, 3 × 3 convolutions with a dilation rate equal to 2 will have a receptive field of 5 × 5. Those with a dilation rate equal to three will have a receptive field of 7 × 7.

The EFE module comprises three hyper-parameters to achieve optimal outputs in terms of accuracy and computational complexity. The hyper-parameters are as follows:

- K: Denotes the receptive field to be developed in the module using the 3 × 3 convolutions.

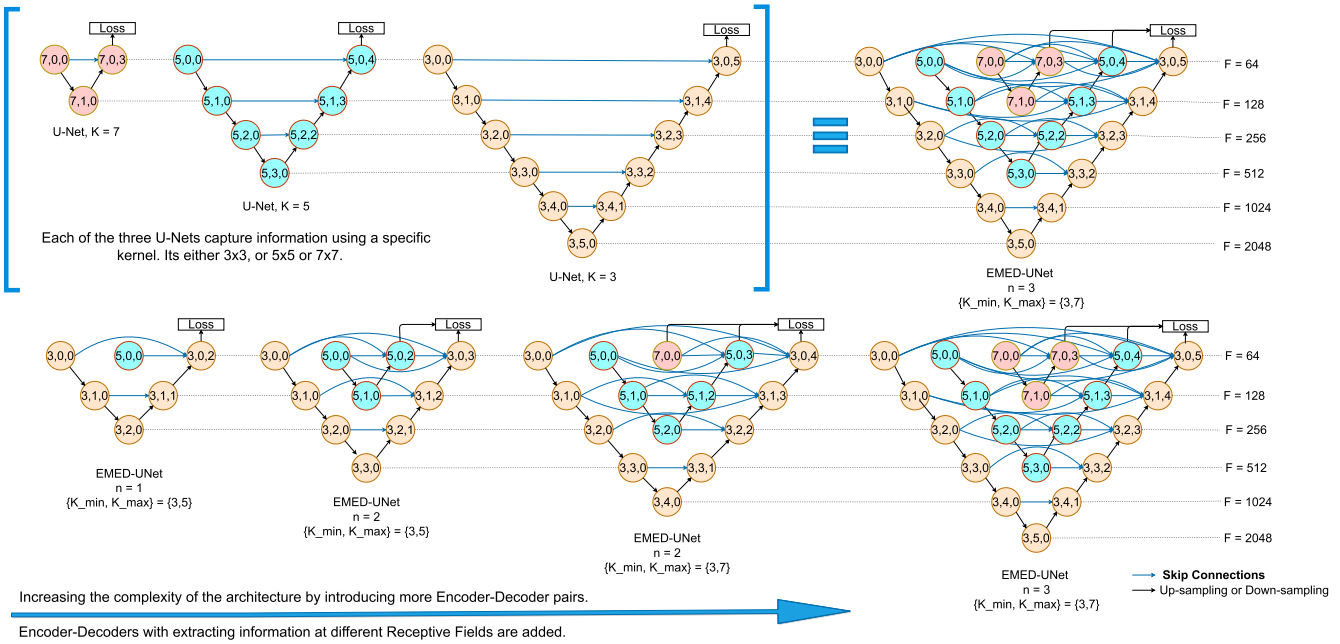


FIGURE 1. ***Top*: EMED-UNet architecture with comparison on how features extracted by EMED-UNet are equivalent to the features extracted by different U-Nets, using different spatial extents. *Bottom*: The EMED-UNet network with different configurations of {n,K_{min},K_{max}} is shown.**

- λ_1: The reduction factor to be applied to limit the number of 1×1 filters(depth dimension of the 1×1 filters).
- λ_2: The reduction factor to be applied to limit the number of 3×3 dilated filters (depth dimension of the 3×3 filters).

There are three parallel paths in the EFE module. One path goes with the 2×2 MaxPooling2D layer followed by a 1×1 convolution. The second path contains a 1×1 convolution bottleneck followed by the 3×3 dilated convolution, and the third path contains only the 1×1 convolution bottleneck. All three parallel paths concatenate at the end as the output of the EFE module. K, λ_1, and λ_2 are the three controls to the EFE module and directly impact the model's receptive field and the total number of trainable parameters.

The EFE module is used in every encoder (before the final down-sampling layer) and decoder (before the final up-sampling layer) node. The following analysis describes the step-by-step operations performed by the EFE module and drives the final output. First, consider F_1 and F_2 to be the reduced number of filters after the reduction caused by the factors λ_1 and λ_2, respectively. 'F' is the input number of filters to the EFE module based on the Network structure as described in Fig. 1. Therefore,

$$F_1 = \frac{1}{\lambda_1} \times F \tag{1}$$

$$F_2 = \frac{1}{\lambda_2} \times F \tag{2}$$

Now, as described in Fig. 2, Z_1, Z_2, and Z_3 are the respective outputs of the three parallel paths. M is the weight matrix of the 2×2MaxPooling2D operation. W_1, W_2, W_3, and W_4 are the weight matrices of the convolution layers, also described in Fig. 2. Let's assume the input image matrix is Z. Then

$$Z_1 = ZW_1, \quad \therefore Z_1 \in R^{H \times W \times F_1} \tag{3}$$

$$Z_2 = (ZW_2)W_3, \quad \therefore Z_2 \in R^{H \times W \times F_2} \tag{4}$$

$$Z_3 = (ZM)W_4, \quad \therefore Z_3 \in R^{H \times W \times F_1} \tag{5}$$

$$Z_{out} = ZW_1 \oplus (ZW_2)W_3 \oplus (ZM)W_4 \tag{6}$$

Finally, $Z \in R^{H \times W \times (F_{out})}$, where F_{out} denotes the number of channels in the output of the module. The expression for F_{out} is as follows:

$$F_{out} = 2 \times F_1 + F_2 \tag{7}$$

$$F_{out} = 2 \times \frac{1}{\lambda_1} \times F + \frac{1}{\lambda_2} \times F \tag{8}$$

where λ_1 and λ_2 can be any number from the sequence {2,4,6,8,..}.

In place of the EFE module, the base U-Net model had two 3×3 convolution layers with an "F" number of filters. The final output will be as $Z_{out} \in R^{H \times W \times (2F)}$. The final number of channels in the U-Net module is significantly more than that of the EFE module.

The above analysis describes the reduction in the number of channels in the model after applying the EFE module. As the EFE module is present in every encoder and decoder node, this will eventually result in an overall parametric reduction in the architecture.

Fig. 1 demonstrates that the proposed EMED-UNet contains multiple embedded U-Nets. The feature extraction

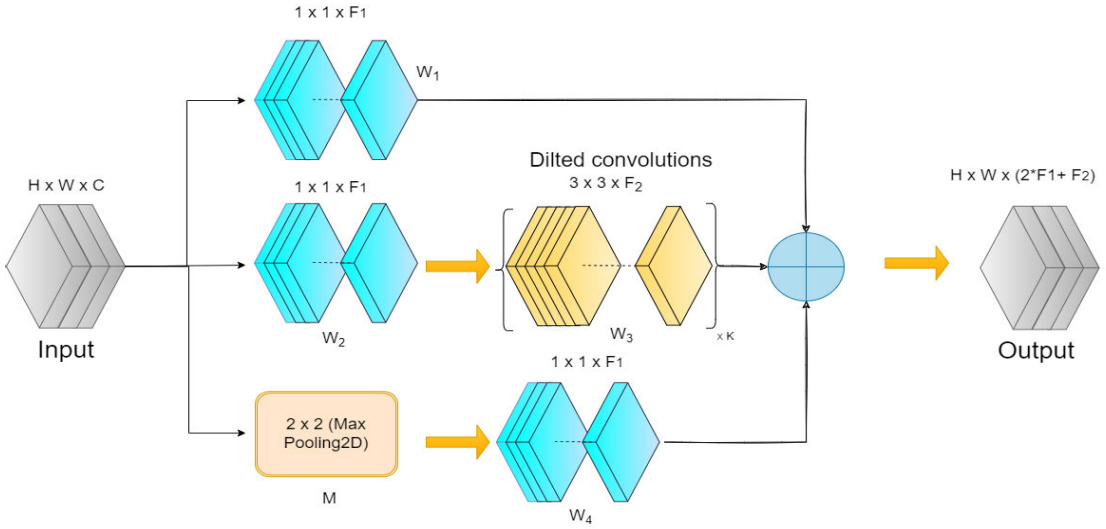


FIGURE 2. The EFE module is used in both the encoders and decoders.

in each U-Net is done using unique kernels. Therefore, using different spatial extents, the features at different receptive fields are extracted by different U-Nets. In this way, the Multiple Encoder-Decoder structures of EMED-UNet can extract information at multiple receptive fields. Fig. 1 describes EMED-UNets with different configurations based on three parameters:

- n: Number of U-Nets or a number of encoder-decoder pairs in the architecture.
- K_{min}: Minimum receptive field formed by any convolution layer in the entire architecture.
- K_{max}: Maximum receptive field formed by any convolution layer in the entire architecture.

C. TECHNICAL SPECIFICATIONS OF THE EMED-UNet ARCHITECTURE

For example, in Fig. 1, the EMED-UNet with specifications: $n, K_{min}, K_{max} = 2, 3, 7$ contains two U-Net models (two encoder-decoder pairs), one with the receptive field of 3×3 (K=3 and therefore dilation rate = 1) in its EFE module, and the other with the receptive field of 5×5 ($K = 5$ and therefore dilation rate = 2) in its EFE module. The single-node 7,0,0 contains only the EFE module with a receptive field of 7×7 ($K = 7$ and, therefore, dilation rate = 3). This node is not a part of any U-Net structure (no up-sampling or down-sampling operations involved), but it is connected with the other U-Nets using skip connections. Thus, the variables $K_{min} = 3$ and $K_{max} = 7$ means the convolutions with receptive field set of {3,5,7} are used in the architecture. In this way, the feature learning in the model will be more effective as feature maps are extracted using convolutions focusing on different receptive fields.

D. CONNECTIVITY OF THE NETWORK

The EMED-UNet network architecture has different connectivity for varying configurations based on three variables: n, K_{min}, and K_{max}. Again, consider the EMED-UNet with parameters $n, K_{min}, K_{max} = 2, 3, 7$ described in Fig. 1, showing the connectivity of the network. Let us denote any node in the network as $X_{K,i,j}$ (denoted in the figure as K, i, j in the circle), where K is the parameter of the EFE module present in that node, i is the level of down-sampling of the node, and j denotes the number of skip connections coming as inputs to the node. Note that in the case of up-sampling, $i = i - 1$, and in the case of down-sampling, $i = i + 1$. For example, if we consider any node $X_{3,1,3}$, the EFE module will have a parameter $K = 3$. The node is part of the outer U-Net sub-network (denoted by the light brown shade), reached after down-sampling the input four times and up-sampling thrice. It can also be verified that the number of skip connection inputs given to the node is three.

E. COLLABORATIVE LEARNING USING DEEP SUPERVISION

The EMED-UNet architecture contains multiple U-Nets. Therefore, there are several segmentation outputs generated at the end of every decoder. This motivates us to apply deep supervision in the model by back-propagating through all the decoder ends and to make all the U-Nets learn collaboratively. This feature ensembles the U-Net sub-components into one model to leverage enhanced collaborative learning between them at multiple receptive fields. The learning will happen in such a way that it learns features within the context of features learned by other encoder-decoder sub-components.

At the time of output collection, we are collecting the best output of those generated at multiple ends. All the

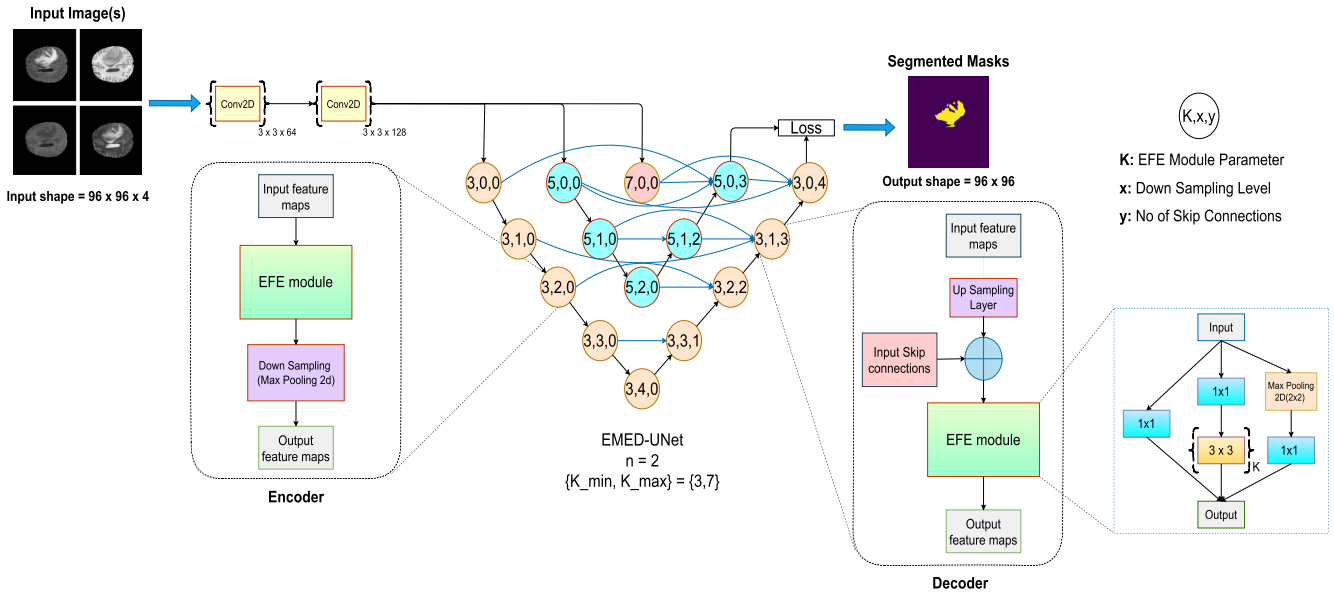


FIGURE 3. Overview of the deeply supervised EMED-UNet architecture, showing the unfolding of Encoder-Decoders of the EMED-UNet and the EFE module.The images shown in the figure are of the BraTS Dataset just to create a case example.

upcoming results of EMED-UNet are generated after applying deep supervision. Fig. 3 shows the deeply supervised EMED-UNet network, getting multiple output masks while back-propagating the loss simultaneously in all the U-Nets. Fig. 3 describes the overall bird's eye view and unfolding of the EMED-UNet architecture.

III. EXPERIMENTS

A. DATASETS

1) MONTGOMERY COUNTY - CHEST X-RAY DATABASE

Montgomery County chest X-ray dataset [40], [41] was created for the diagnosis of Tuberculosis (TB) disease. It was a collaborative work of the National Library of Medicine (NLM) with the Department of Health and Human Services, Montgomery County, Maryland, USA. The dataset contains 138 chest X-rays, out of which 58 are from infected individuals and the rest are from non-infected individuals. This relatively small data set created a challenge to make the model learn all the critical features using a limited number of samples. The images were resized to $512 \times 512 \times 3$ before passing them into the model to train.

2) TUBERCULOSIS CHEST X-RAY DATA(SHENZEN)

The Shenzhen Tuberculosis dataset [40], [41], [42] is a chest X-ray dataset created to detect Tuberculosis disease. The dataset resulted from the collaborative work of NLM and Shenzhen No.3 People's Hospital, Guangdong Medical College, Shenzhen, China. It contains 662 images, 336 of which are from infected cases and 326 from normal cases. The 566 lung masks for this dataset were proposed by the Computer Engineering Department, Faculty of Informatics and Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine. Lung masks are available here. The images are resized to $512 \times 512 \times 3$ during processing.

3) COVID-19 CT LESION SEGMENTATION DATASET

This is a large lung CT scan dataset [43], [44], [45] which was developed by combining data from three public datasets on COVID-19. It contains 2,729 images with their corresponding ground truth lesion masks. Each CT scan file of a patient consists of multiple slices, and thus every image is a slice from the CT scans of the patients. These scans are converted to NumPy arrays and then to png files for further processing. The resolution of every image in the dataset is 512×512. We converted the images to the resolution of 256×256 before feeding the images to train in the model. The data set is publicly available at Kaggle.

4) BRAIN TUMOR SEGMENTATION (BraTS) DATASET

This dataset focuses on the evaluation of state-of-the-art methods for the segmentation of brain tumors in multimodal magnetic resonance imaging (MRI) scans. It contains volumes with multimodal scans: T1, post-contrast T1-weighted T1Gd, T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR). The dataset contains over 66,348 slices. We converted the images to the resolution of 96×96 before training.

TABLE 2. Effect of choosing different configurations of EMED-UNet on parameters, FLOPS, and, model size in EMED network.

$\{n, K_{min}, K_{max}, \lambda_1, \lambda_2\}$	No of Trainable parameters (M)	FLOPS ($\times 10^{02}$ G)	Model Size (MB)
{2,3,7,4,1}	10.39	2.16	126
{2,3,5,4,1}	2.615	0.382	32.3
{2,3,7,4,2}	6.35	1.52	77.5
{2,3,5,2,1}	4.65	0.658	56.7

TABLE 3. Training parameters.

Train: Val: Test(split ratio)	70:10:20
number of epochs	60
learning rate	$1e^{-5}$
Batch size	4
epochs patience	5
optimizer	Adam
min learning rate	$1e^{-7}$

TABLE 4. High-Performance computing specifications of the Paramshavak Supercomputer.

Specifications (Supercomputer Paramshavak)	
CUDA cores	3072
Operating System	Ubuntu 18.4.3 LTS
GPU (on board)	Nvidia Quadro P4000
GPU (add on)	Nvidia Quadro RTX 5000
RAM	96 GB

B. IMPLEMENTATION

1) EMED-UNet

As described in Section – II, the EMED-UNet model uses the hyperparameters: n, K_{min}, K_{max}, λ_1, and λ_2. We iterated over different model configurations to find the optimal set of parameters. Increasing the values of λ_1 and λ_2 will reduce the number of channels in the model by a significant amount. Therefore, the model's performance will start deprecating after a critical value of λ_1 and λ_2. The critical values were found to be $\lambda_1 = 16$ and $\lambda_2 = 4$. The following are the settings where good performance was achieved with the model:

$$\{n, K_{min}, K_{max}, \lambda_1, \lambda_2\} = \{2, 3, 7, 4, 1\} \quad (9)$$
$$\{n, K_{min}, K_{max}, \lambda_1, \lambda_2\} = \{2, 3, 5, 4, 1\} \quad (10)$$
$$\{n, K_{min}, K_{max}, \lambda_1, \lambda_2\} = \{2, 3, 7, 4, 2\} \quad (11)$$
$$\{n, K_{min}, K_{max}, \lambda_1, \lambda_2\} = \{2, 3, 5, 2, 1\} \quad (12)$$

The effect of change in parameters with different configurations mentioned above is shown in Table 2.

2) COMPARISON OF MODELS

We compared the results of the EMED-UNet architecture with two U-Net models, each resulting in a different receptive field. One U-Net contains the 3×3 filters in every convolution layer, and the other has 5×5 filters in each of its convolution layers. Table 2 describes the set of filters used throughout the architectures. Detailed information on the U-Net implementation is presented in the later sections.

We took the Train:Val:Test split ratio of 70:10:20 in the case of MC, SHZ, and C-19 CT datasets. The EMED-UNet was trained at a learning rate of $1e^{-5}$ with an automatic reduction in learning during validation, extending to the minimum until $1e^{-7}$. The Adam optimization was used in the training to optimize learning. In the pre-processing stage, images were resized to 512×512.

Table 3 shows the hyperparametric settings while training the model. The model was trained on a Paramshavak Supercomputer, the specifications of which are outlined in Table 4.

Every model mentioned in the results section is trained once on each dataset. Training and testing on one data set are done independently from all other data sets.

C. PERFORMANCE METRICS AND LOSS FUNCTIONS

The Performance Metrics and Loss Functions used in evaluating the model are as follows:

- **Dice-coefficient**:
  The Dice Coefficient [46] is a similarity metric used to measure the similarity between two images. The mathematical formulation of the Dice-coefficient metric function is:
  $$DC(A, B) = \frac{2\times \mid A \cap B \mid}{\mid A \mid + \mid B \mid} \quad (13)$$
  where A is the prediction, and B is the ground truth.
- **Intersection over Union**:
  The Intersection over Union (IoU) [47], also known as Jaccard Index, is a standard metric to evaluate segmentation performance. This metric shows the overlap ratio between the segmented output and the ground truth.
- **Precision**:
  Precision shows, out of all the predicted masks, how many of those objects had matched or overlapped with the ground truth annotation.
- **Dice Loss**:
  The Dice Loss [46] is also one of the most commonly used loss functions in segmentation tasks. It is based on the dice coefficient metric. The good characteristic of this function is that it can compute loss at global as well as local scale.
  $$DL(y, \hat{p}) = 1 - \frac{2y\hat{p} + 1}{y + \hat{p} + 1} \quad (14)$$
  Above is the mathematical formulation of the dice loss where y is the ground truth and $\hat{p}$ is the predicted output.
- **Metrics Comparing Efficiency**:
  The metrics that compare the efficiency of the models

TABLE 5. Results of segmentation on Montgomery County dataset.

Model	Params(M)	FLOPS ($\times 10^{02}$G)	Loss	DC	IoU	Precision	Model Size (MB)
U-Net (3x3) [8]	31.043	3.86	0.0840	0.9160	0.8462	0.9733	373
U-Net (5x5) [8]	81.247	10.4	0.0757	0.9243	0.8601	**0.9752**	975
U-Net++ (3x3) [16]	33.241	4.05	0.0542	0.9458	0.8751	0.9744	429
EMED-UNet{2,3,7,4,2}	6.35	1.52	0.0600	0.9400	0.8871	0.9730	77.5
EMED-UNet{2,3,5,2,1}	4.65	0.658	**0.0422**	**0.9578**	**0.9195**	0.9604	56.7
EMED-UNet{2,3,5,4,1}	2.615	0.382	0.0455	0.9545	0.9133	0.9540	32.3
PSPNet	11.884	1.84	0.0831	0.9169	/	/	799
PSPNet T-AE [25]	$\geq$ 11.894	$\geq$ 1.84	0.0426	0.9574	/	/	$\geq$799
Rec-PSPNet [25]	$\geq$ 11.894	$\geq$ 1.84	0.0424	0.9576	/	/	$\geq$799
Rec-UNet [25]	$\geq$ 31.043	$\geq$3.86	0.0425	0.9575	/	/	$\geq$ 373
DeepLab V3+	26.0	4.35	0.1296	0.8704	/	/	503.06

TABLE 6. Results of segmentation on Shenzhen CXR dataset.

Model	Params(M)	FLOPS ($\times 10^{02}$G)	Loss	DC	IoU	Precision	Model Size (MB)
U-Net (3x3) [8]	31.043	3.86	0.0577	0.9423	0.8917	0.9605	373
U-Net (5x5) [8]	81.247	10.4	0.0630	0.9370	0.8825	0.9655	975
U-Net++ (3x3) [16]	33.241	4.05	0.0624	0.9376	0.8834	**0.9660**	429
EMED-UNet{2,3,7,4,2}	6.35	1.52	0.0522	0.9478	**0.9021**	0.9654	77.5
wide U-Net [16]	33.23	4.04	0.0734	0.9266	0.8740	0.9597	420
PSPNet	11.884	1.84	0.1674	0.8326	/	/	799
PSPNet (TAE) [25]	$\geq$ 11.884	$\geq$ 1.84	0.0426	0.9574	/	/	$\geq$799
Rec-PSPNet [25]	$\geq$ 11.884	$\geq$ 1.84	0.0425	0.9575	/	/	$\geq$ 799
Rec-UNet [25]	$\geq$ 31.043	$\geq$ 3.86	**0.0424**	**0.9576**	/	/	$\geq$ 373
DeepLab V3+	26.0	4.35	0.1351	0.8649	/	/	503.06

TABLE 7. Results of segmentation on C-19 CT LS dataset.

Model	Params(M)	FLOPS ($\times 10^{2}$G)	Loss	DC	IoU	Precision	Model Size (MB)
U-Net (3x3) [8]	31.043	3.86	0.2426	0.7571	0.6332	0.8386	373
U-Net (5x5) [8]	81.247	10.4	0.2299	0.7699	0.6489	0.8480	975
U-Net++ (3x3) [16]	33.241	4.05	0.2270	0.7730	0.6662	0.8510	429
EMED-UNet{2,3,5,2,1}	4.65	0.658	0.2572	0.7424	0.6120	0.8536	56.7
EMED-UNet{2,3,7,4,2}	6.35	1.52	0.2320	0.7666	0.6299	0.8504	77.5
EMED-UNet{2,3,7,4,1}	10.39	2.16	**0.1969**	**0.8018**	**0.6769**	**0.8565**	126
wide U-Net [16]	33.23	4.04	0.2281	0.7719	0.6654	0.8459	420

are the number of parameters and FLOPS (Floating Point Operations Per Second). The overall inference time of a system depends on the computing device (CPU, GPU cores, level of parallel processing, and more). Therefore, we use the number of FLOPS for evaluating the efficiency of the models. FLOPS are proportional to the inference time of any system, which is directly proportionate to the system's efficiency.

D. RESULTS

Table 5 compares the performance of the U-Net base model, its other variants, other state-of-the-art models, and the EMED-UNet model on the Montgomery County dataset. This is a relatively small dataset where the total number of images is 138. As shown in the table, the EMED-UNet model with parameters {2,3,5,2,1} performed better than the U-Net base model, its other variants, PSPNet, other compared models, and its other EMED-UNet counterparts with an IoU of 0.9195 and dice coefficient of 0.9578, with a significantly decreasing in the number of trainable parameters from 31.043 (in UNet(3×3)) million to 4.65M and FLOPS from 386 G to 65.8 G. It also significantly outperforms UNet++ in terms of efficiency and stays inline in terms of accuracy.

Table 6 also shows the comparison between the U-Net base model and other state-of-the-art models on the Shenzen dataset. The precision, IoU, and dice coefficient increased from 0.9655, 0.8825, and 0.9370 in the case of U-Net(5×5) to 0.9654, 0.9021, and 0.9478 in EMED-UNet with configuration {2,3,7,4,2}. Thus, on a larger dataset with more training images, the EMED-UNet performs better than U-Net. The UNet++ achieves slightly better precision, but it is outperformed by EMED-UNet{2,3,7,4,2} in terms of Dice Coefficient, IoU, and Loss, with much better efficiency.

Table 7 demonstrates the comparison between the U-Net base model, its other variants, and the EMED-UNet network

TABLE 8. Results of segmentation on BraTS dataset. Note: the input image size used is 96 × 96.

Model	Params(M)	FLOPS ($\times 10^{02}$G)	Loss	DC	Model Size (MB)
U-Net (3x3) [8]	7.8	0.97	0.29	0.71	91
U-Net (5x5) [8]	20.32	2.61	0.28	0.72	238
U-Net++ [16]	9.0	1.02	0.264	0.736	104.6
EMED-UNet{2,3,5,2,1}	1.35	0.165	0.272	0.728	13.83
EMED-UNet{2,3,7,4,1}	1.84	0.38	**0.26**	**0.74**	18.9
wide U-Net [16]	9.1	1.01	0.276	0.724	102

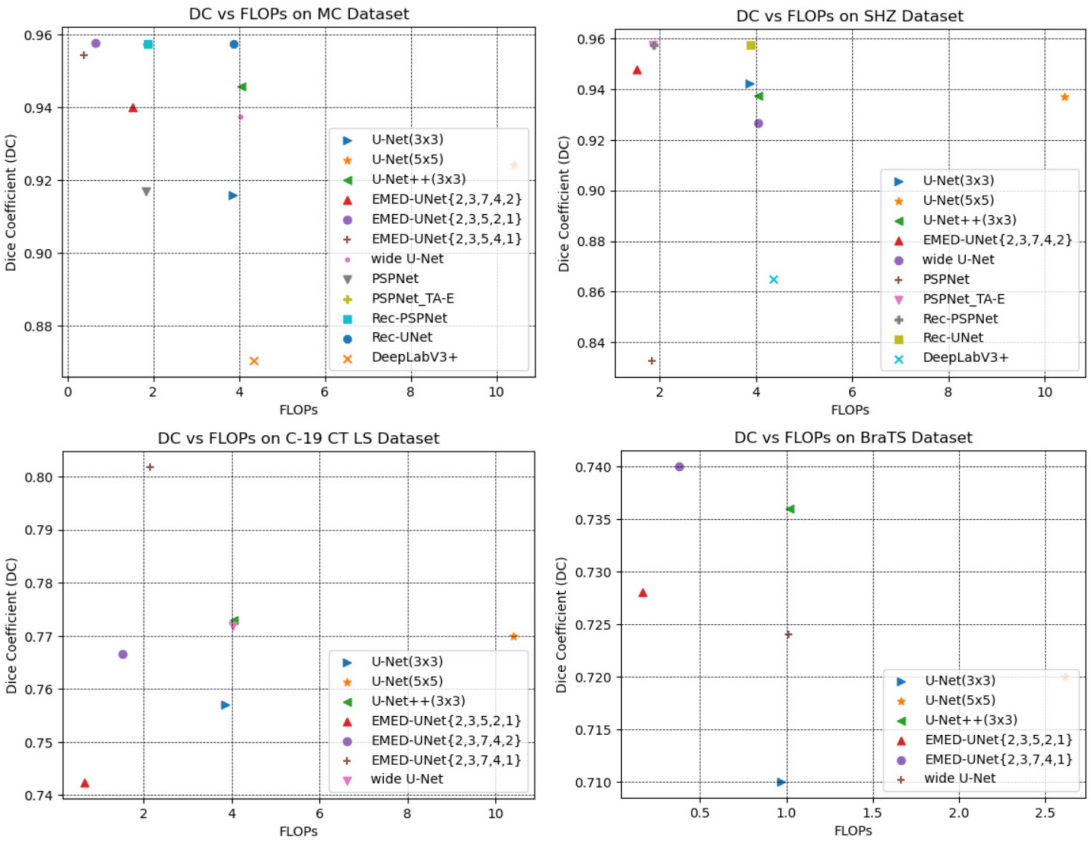


FIGURE 4. Dice Coefficient vs. FLOPS on Montgomery County, Shenzhen CXR, Covid-19 CT Lesion segmentation, and the BraTS datasets.

on the COVID-19 CT Lesion Segmentation dataset. The EMED-UNet{2,3,7,4,2} outperforms the U-Net(3 × 3) in terms of both accuracy and efficiency. But, the U-Net(5 × 5) performs better in terms of accuracy. Still, an EMED-UNet configuration like EMED-UNet{2,3,7,4,1}, where the model loses some efficiency, can outperform the U-Net(5 × 5), with around 4% improvement in the dice coefficient, 4.3% increment in the IoU, and also 1% increment in the precision. Similarly, a significant increment is observed in the accuracy and efficiency when compared to UNet++.

Table 8 shows that the performance of the EMED-UNet models outperforms all the other models in terms of both accuracy (DC) and efficiency (FLOPs and Params(M)). The EMED-UNet{2,3,7,4,1} outperforms the other models with a DC of 0.74 with only 1.84 M parameters and 0.38G FLOPs. It turns out that the EMED-UNET{2,3,5,2,1} is the lightest model, which also gives a decent accuracy of 0.272. Thus, it could be useful for specific purposes.

The above results prove that the embedded U-Net structure of the EMED-UNet Network is efficient in learning parameters as compared to the dense architecture of U-Net++, as it takes fewer parameters for learning the salient features of the medical images.

Fig. 4 shows the Dice Coefficient vs. FLOPS on four medical imaging datasets. It can be observed that the EMED-UNet model consistently outperformed the U-Net and its other variants in terms of efficiency. EMED-UNet model achieved a gain in accuracy on the MC, Shenzhen, and BraTS datasets, while it lost some accuracy on the COVID-19 CT dataset.

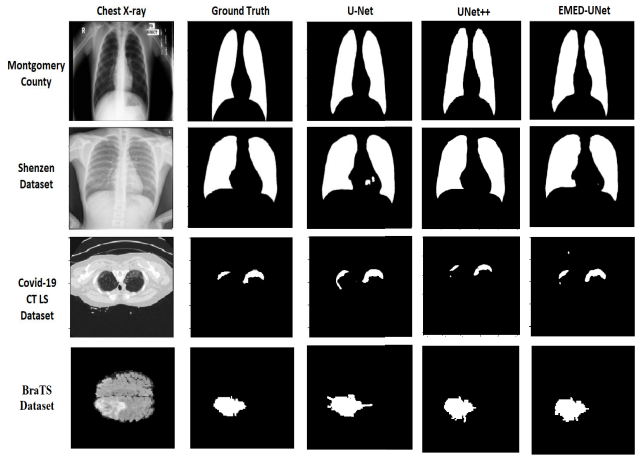


FIGURE 5. Qualitative results comparison between output samples generated with U-Net and the EMED-UNet models on all the datasets. The EMED-UNet configuration used is {2,3,7,4,2} for the Shenzhen and C-19 Datasets, and {2,3,5,2,1}, on the Montgomery County dataset.

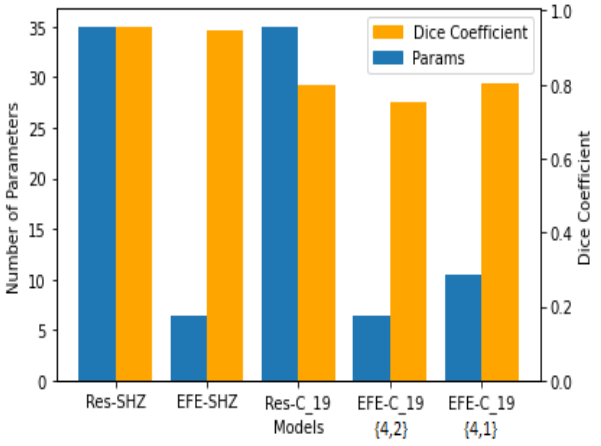


FIGURE 6. Results of Ablation study. The Model 'Res-XX' denotes the residual backbone on the multi-encoder-decoder network, and 'EFE-XX' denotes the multi-encoder-decoder network with the EFE module, which is basically the EMED-UNet architecture.

Fig. 5 demonstrates a visual output comparison between U-Net and EMED-UNet segmentation results. In terms of accuracy, both the U-Net and EMED-UNet are performing well, but in some edge cases, EMED-UNet performs better. For example, as shown in Fig. 5, the U-Net predicted the '1' (in binary image) in some places, while EMED-UNet has a false prediction. The accuracy of the results can be verified from the numerical results in Table 4.

E. ABLATION STUDIES

We conducted two ablation studies: First, to compare the EFE module with another backbone; Second, to compare the proposed Network architecture (Multi-Encoder-Decoder network) with the base U-Net architecture. We chose the strong residual backbone [48] and evaluated its effects on the proposed Multi-Encoder-Decoder network. We replace the EFE module with a ResNet module and compare the efficiency and accuracy in both cases. For the purposes of this section, let us refer to EMED-UNet without the EFE module as MED-UNet. The MED-UNet with the residual module is denoted by MED-UNet(res).

The residual module used in place of the EFE module consists of two convolution layers, one 1×1 convolution (to match the third dimension) and the addition operation. The residual module takes three inputs: input feature maps, the number of filters (F), the parameter 'K', and the type of filter used in convolution (i.e., the 3×3 filter or 5×5). For example, the node $X_{5,1,2}$ has the number of filters $F = 128$ and $K = 5$, and the input to that residual module is the output of the node $X_{5,2,0}$. Further, the two convolution layers used have the kernel size of 3×3 with a dilation rate of 2 to maintain the receptive field of 5 (as $K = 5$).

1) ABLATION STUDY - 1

In the MED-UNet, the configuration will be $\{n, K_{min}, K_{max}\}$. There are no reduction factors involved as there is no EFE module present. We used the configurations: MED-UNet{2,3,7}(res) (with residual module), EMED-UNet{2,3,7,4,2} (with EFE), and EMED-UNet{2,3,7,4,1} (with EFE). In Fig. 6, the Res-SHZ refers to the MED-UNet with a residual backbone (i.e. MED-UNet(res)), and the EFE-SHZ refers to the original EMED-UNet, both on the Shenzhen Dataset and the Covid-19 CT Lesion Segmentation dataset.

Fig. 6 and Table 8 describe the results of the ablation study. The MED-UNet(res) used 81% more parameters as compared to the EMED-UNet{2,3,7,4,2}. It achieves a negligible accuracy gain on the Shenzhen dataset and a small gain on the COVID-19 CT dataset. This demonstrates that the EFE modules work well in finding the balance between accuracy and efficiency. The detailed results comparison is shown in Table 8.

2) ABLATION STUDY - 2

In this study, we tested the effects of different backbones on both the MED-UNet and the U-Net architectures. We considered the ResNet backbones. The MED-UNet network consists of multiple embedded U-Nets, with each U-Net extracting information at different receptive fields. We demonstrate that the MED-UNet network outperforms U-Net in terms of accuracy by using a common backbone in both networks.

The residual module used in MED-UNet is the same as the one used in the ablation study. Here we are considering the same model of MED-UNet(res) with a residual module as the one we considered in the ablation study - 1. In the case of U-Net, we use regular 3×3 kernels in the residual module rather than using dilated convolutions. The U-Net model with residual backbone is denoted by U-Net(3×3)(res).

Again, we conducted the experiment on the COVID-19 CT LS dataset. The results in Table 8 show the strength of the Multi-Encoder-Decoder model architecture. It is

TABLE 9. Results of studies from Section-IV subsections-E and F.

Model	Dataset	params(M)	DC
U-Net(3 × 3)(res)	C-19	32.427	0.7353
EMED-UNet{2,3,7,4,1}	C-19	10.39	0.8018
EMED-UNet{2,3,7,4,2}	C-19	6.35	0.7666
MED-UNet{2,3,7}(res)	C-19	36	0.7992
MED-UNet{2,3,7}(res)	Shenzhen	36	0.9574
EMED-UNet{2,3,7,4,2}	Shenzhen	6.35	0.9478

observed that the MED-UNet{2,3,7,4,2}(res) outperforms the U-Net(3 × 3)(res) with the dice coefficient of 0.7992, with U-Net having a dice coefficient of 0.7353. This shows the strength of the EMED-UNet Network architecture.

IV. DISCUSSION

The ablation study demonstrates the impact of the EFE module on the overall model efficiency. Using the reduction factors as controls, we can find the best trade-off between accuracy and efficiency. In the ablation study, we observed that the EMED-UNet{2,3,7,4,2} consisting of 6.35M parameters nearly underperforms compared to MED-UNet{2,3,7}(res), consisting of 36M parameters, in terms of accuracy. However, if we consider a different configuration of EMED-UNet like {2,3,7,4,1} consisting of 10.39M parameters, it slightly crosses MED-UNet{2,3,7} with the residual module with a dice coefficient of 0.8018, on the C-19 dataset. Thus, by compromising on efficiency, one can get better accuracy.

While increasing the value of the reduction factors in the EFE module, the model is losing accuracy as we reduce the number of channels in the feature maps. However, we can recover the accuracy using the proposed structure of the EMED-UNet network. The network consisting of multiple encoders and decoders can capture the salient features from the images using multiple receptive fields. The strength of the EMED-UNet network is shown in the study described in Section E. Thus, the EMED-UNet network with the EFE module, i.e. the EMED-UNet network, is the best fit.

Based on the detailed study of the proposed network, we can see that a major improvement in results is arising due to the network structure. This work might be further uplifted by observing the effect of a transformer backbone on the same network structure while testing multiple network configurations.

V. CONCLUSION

The proposed EMED-UNet network architecture was demonstrated as very effective at segmenting biomedical images with improved accuracy and significantly reduced parameters, FLOPS, and model size as compared to the standard U-Net base model. By employing the EFE module, our network decreases the computational complexity and memory required for the semantic segmentation task. The EMED-UNet can learn all the salient features in the image by introducing multiple encoder-decoders to extract features at multiple receptive fields, incorporating collaborative learning using the deep supervision technique. We evaluated our multi-encoder-decoder network on four medical image segmentation datasets and observed that it matches or exceeds the accuracy of the standard U-Net base, measured using multiple performance metrics, on a very low parametric cost, significantly fewer FLOPS, and low memory consumption. Iterating through different parameters in the model, one can find the best configuration for the segmentation task on any medical image modality that maintains a good trade-off between accuracy and efficiency.

REFERENCES

[1] W. Li, F. Jia, and Q. Hu, "Automatic segmentation of liver tumor in CT images with deep convolutional neural networks," *J. Comput. Commun.*, vol. 3, no. 11, pp. 146–151, 2015.

[2] R. Vivanti, A. Ephrat, L. Joskowicz, O. Karaaslan, N. Lev-Cohain, and J. Sosna, "Automatic liver tumor segmentation in follow-up CT studies using convolutional neural networks," in *Proc. Patch-Based Methods Med. Image Process. Workshop*, vol. 2, 2015, pp. 1–9.

[3] B. H. Menze, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015.

[4] V. Cherukuri, P. Ssenyonga, B. C. Warf, A. V. Kulkarni, V. Monga, and S. J. Schiff, "Learning based segmentation of CT brain images: Application to postoperative hydrocephalic scans," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 8, pp. 1871–1884, Aug. 2018.

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[6] K. Zhou, H. Greenspan, and D. Shen, *Deep Learning for Medical Image Analysis*. New York, NY, USA: Academic Press, 2017.

[7] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[9] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, vol. 9, pp. 82031–82057, 2021.

[10] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2017, pp. 1–4.

[11] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5168–5177.

[12] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "ICNet for real-time semantic segmentation on high-resolution images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 405–420.

[13] N. Tajbakhsh, B. Lai, S. P. Ananth, and X. Ding, "ErrorNet: Learning error representations from limited data to improve vascular segmentation," in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2020, pp. 1364–1368.

[14] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 94–114, Apr. 2020.

[15] S. Cai, Y. Tian, H. Lui, H. Zeng, Y. Wu, and G. Chen, "Dense-UNet: A novel multiphoton in vivo cellular image segmentation model based on a convolutional neural network," *Quant. Imag. Med. Surgery*, vol. 10, no. 6, pp. 1275–1285, Jun. 2020.

[16] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.

[17] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "UNet 3+: A full-scale connected UNet for medical image segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1055–1059.

[18] M. Lei, J. Li, M. Li, L. Zou, and H. Yu, "An improved UNet++ model for congestive heart failure diagnosis using short-term RR intervals," *Diagnostics*, vol. 11, no. 3, p. 534, Mar. 2021.

[19] L. Zou, Z. Huang, X. Yu, J. Zheng, A. Liu, and M. Lei, "Automatic detection of congestive heart failure based on multiscale residual UNet++: From centralized learning to federated learning," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–13, 2023.

[20] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[21] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Med. Image Anal.*, vol. 53, pp. 197–207, Apr. 2019.

[22] N. Saeedizadeh, S. Minaee, R. Kafieh, S. Yazdani, and M. Sonka, "COVID TV-Unet: Segmenting COVID-19 chest CT images using connectivity imposed Unet," *Comput. Methods Programs Biomed. Update*, vol. 1, Jan. 2021, Art. no. 100007.

[23] N. Saeedizadeh, M. Tajmirriahi, A. Haghani, Z. Amini, E. K. Pour, H. Riazi-Esfahani, K. Fadakar, R. Kafieh, and H. Rabbani, "A device-independent, shape preserving retinal optical coherence tomography image alignment method applying TV-Unet for RPE layer detection," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–8, 2022.

[24] T. Shen and H. Xu, "Medical image segmentation based on transformer and HarDNet structures," *IEEE Access*, vol. 11, pp. 16621–16630, 2023.

[25] Y. Chen, H. Zhang, Y. Wang, L. Liu, Q. M. J. Wu, and Y. Yang, "TAE-Seg: Generalized lung segmentation via tilewise AutoEncoder enhanced network," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–13, 2022.

[26] A. Das, "Adaptive UNet-based lung segmentation and ensemble learning with CNN-based deep features for automated COVID-19 diagnosis," *Multimedia Tools Appl.*, vol. 81, no. 4, pp. 5407–5441, Feb. 2022.

[27] A. L. Aswathy and S. S. V. Chandra, "Cascaded 3D UNet architecture for segmenting the COVID-19 infection from lung CT volume," *Sci. Rep.*, vol. 12, no. 1, p. 3090, Feb. 2022.

[28] H. Malik, T. Anees, M. Din, and A. Naeem, "CDC_Net: Multi-classification convolutional neural network model for detection of COVID-19, pneumothorax, pneumonia, lung cancer, and tuberculosis using chest X-rays," *Multimedia Tools Appl.*, vol. 82, pp. 13855–13880, Sep. 2022.

[29] S. Yin, H. Deng, Z. Xu, Q. Zhu, and J. Cheng, "SD-UNet: A novel segmentation framework for CT images of lung infections," *Electronics*, vol. 11, no. 1, p. 130, Jan. 2022.

[30] A. Naeem, T. Anees, K. T. Ahmed, R. A. Naqvi, S. Ahmad, and T. Whangbo, "Deep learned vectors' formation using auto-correlation, scaling, and derivations with CNN for complex and huge image retrieval," *Complex Intell. Syst.*, vol. 9, pp. 1–23, Oct. 2022.

[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[32] N. S. Punn and S. Agarwal, "Inception U-Net architecture for semantic segmentation to identify nuclei in microscopy cell images," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 1, pp. 1–15, Feb. 2020.

[33] D. E. Cahall, G. Rasool, N. C. Bouaynaya, and H. M. Fathallah-Shaykh, "Dilated inception U-Net (DIU-Net) for brain tumor segmentation," 2021, *arXiv:2108.06772*.

[34] Z. Zhang, C. Wu, S. Coleman, and D. Kerr, "DENSE-INception U-net for medical image segmentation," *Comput. Methods Programs Biomed.*, vol. 192, Aug. 2020, Art. no. 105395.

[35] N. Beheshti and L. Johnsson, "Squeeze U-Net: A memory and energy efficient image segmentation network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 1495–1504.

[36] D. Qin, J.-J. Bu, Z. Liu, X. Shen, S. Zhou, J.-J. Gu, Z.-H. Wang, L. Wu, and H.-F. Dai, "Efficient medical image segmentation based on knowledge distillation," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3820–3831, Dec. 2021.

[37] A. He, K. Wang, T. Li, C. Du, S. Xia, and H. Fu, "H2Former: An efficient hierarchical hybrid transformer for medical image segmentation," *IEEE Trans. Med. Imag.*, early access, Apr. 5, 2023, doi: 10.1109/TMI.2023.3264513.

[38] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," *Med. Image Anal.*, vol. 88, Aug. 2023, Art. no. 102802.

[39] K. D. Shah, D. K. Patel, H. A. Patel, and H. R. Nagrani, "EMED-UNet: An efficient multi-encoder–decoder based UNet for chest X-ray segmentation," in *Proc. IEEE Region 10 Symp. (TENSYMP)*, Jul. 2022, pp. 1–6.

[40] S. Candemir, S. Jaeger, K. Palaniappan, J. P. Musco, R. K. Singh, Z. Xue, A. Karargyris, S. Antani, G. Thoma, and C. J. McDonald, "Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration," *IEEE Trans. Med. Imag.*, vol. 33, no. 2, pp. 577–590, Feb. 2014.

[41] S. Jaeger, A. Karargyris, S. Candemir, L. Folio, J. Siegelman, F. Callaghan, Z. Xue, K. Palaniappan, R. K. Singh, S. Antani, G. Thoma, Y.-X. Wang, P.-X. Lu, and C. J. McDonald, "Automatic tuberculosis screening using chest radiographs," *IEEE Trans. Med. Imag.*, vol. 33, no. 2, pp. 233–245, Feb. 2014.

[42] S. Stirenko, Y. Kochura, O. Alienin, O. Rokovyi, Y. Gordienko, P. Gang, and W. Zeng, "Chest X-ray analysis of tuberculosis by deep learning with segmentation and augmentation," in *Proc. IEEE 38th Int. Conf. Electron. Nanotechnol. (ELNANO)*, Apr. 2018, pp. 422–428.

[43] S. P. Morozov, A. E. Andreychenko, N. A. Pavlov, A. V. Vladzymyrskyy, N. V. Ledikhova, V. A. Gombolevskiy, I. A. Blokhin, P. B. Gelezhe, A. V. Gonchar, and V. Yu. Chernina, "MosMedData: Chest CT scans with COVID-19 related findings dataset," 2020, *arXiv:2005.06465*.

[44] M. Jun et al., "COVID-19 CT lung and infection segmentation dataset," Apr. 2020. [Online]. Available: https://zenodo.org/record/3757476

[45] *COVID-19 CT Segmentation Dataset*. Accessed: Dec. 23, 2020. [Online]. Available: http://medicalsegmentation.com/covid19/

[46] S. Jadon, "A survey of loss functions for semantic segmentation," in *Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol. (CIBCB)*, Oct. 2020, pp. 1–7.

[47] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 658–666.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.



KASHISH D. SHAH received the B.Tech. degree in information and communication technology from the School of Engineering and Applied Sciences, Ahmedabad University, Gujarat, India, in 2022. He is currently a full-time Ph.D. Scholar with the Computer Science and Engineering Department, Ahmedabad University. He is also a Junior Research Fellow of the 5G-ITS Project financed by the Department of Science and Technology (DST), GUJCOST, India. His current research interests include computer vision-based image analysis, AI-based coexistence of 5G NR-V2X and WiFi systems under the unlicensed spectrum, bio-medical image segmentation, computer vision, reinforcement learning, deep learning, wireless networks, and cellular-V2X.



DHAVAL K. PATEL (Member, IEEE) received the B.E. and M.E. degrees (Hons.) in communication systems engineering from Gujarat University, Ahmedabad, India, in 2003 and 2010, respectively, and the Ph.D. degree in electronics and communications from the Institute of Technology, Nirma University, Ahmedabad, in 2014. He was a Visiting Faculty with the Franklin W. Olin College of Engineering, Needham, MA, USA. From 2011 to 2014, he was a Junior Research Fellow with the Post Graduate Laboratory for Communication Systems, Nirma University. Since 2014, he has been with the School of Engineering and Applied Science, Ahmedabad University, India, where he is currently an Associate Professor. He has completed the research projects funded by the Department of Science and Technology, U.K.–India Education and Research Initiative (UKIERI), the Association of Southeast Asian Nations (ASEAN) India Collaborative Research and Development Project, and the Gujarat Council on Science and Technology (GUJCOST), as a Principal Investigator. His current research interests include vehicular cyber-physical systems, 5G wireless networks, non-parametric statistics, and physical layer security. He is a Reviewer of various international conferences, including IEEE ICASSP, IEEE VTC, and IEEE PIMRC.



MINESH P. THAKER (Member, IEEE) has been a Cisco Instructor with the Cisco Networking Academy, since 2005. He is currently an Assistant Professor with the Indus Institute of Technology and Engineering, Indus University, India. He has published many research papers in international journals and international conferences. His current research interests include fault tolerance in MPLS, network layer in MANET, VANET, and xG technologies Internet of Things. He is a member of professional bodies, like ISTE, IE, and IETE.



HARSH A. PATEL received the B.E. degree in information and communication technology from the School of Engineering and Applied Sciences, Ahmedabad University, Gujarat, India, in 2022. His current research interests include computer vision, human pose estimation, and medical image analysis.



MANOB JYOTI SAIKIA (Member, IEEE) received the B.E. degree in electronics and communication engineering from Visvesvaraya Technological University, Belgaum, India, in 2009, the M.Tech. degree in bioelectronics from Tezpur University, Tezpur, Assam, India, in 2013, and the Ph.D. degree in electrical engineering from The University of Rhode Island, Kingston, RI, USA, in 2019. He was a Research Associate with the School of Engineering, Tufts University, from September 2019 to August 2022. He was a Senior Research Associate with the Department of Engineering, Boston College, from May 2022 to August 2022. From January 2016 to July 2019, he was a Research Assistant of a project funded by the National Science Foundation, USA. From 2012 to 2015, he was awarded a Senior Research Fellowship from the Ministry of Science and Technology, India, working at the Indian Institute of Science, Bengaluru, India. He is currently a tenure-track Assistant Professor with the Electrical Engineering Department, University of North Florida. His current research interests include biomedical instrumentation, sensors, neuroimaging (fNIRS and EEG), signal processing, machine learning, and the Internet of Things.



BRYAN J. RANGER (Member, IEEE) received the B.S.E. and M.S.E. degrees in biomedical engineering from the University of Michigan, in 2011 and 2012, respectively, and the Ph.D. degree in medical engineering and medical physics from the Massachusetts Institute of Technology (MIT), in 2018. At MIT, he was a NSF Graduate Research Fellow, and his research consisted of developing ultrasound imaging methods to improve and facilitate lower limb prosthesis design. From 2018 to 2021, he was a Program Officer with the Bill and Melinda Gates Foundation, where he managed research and investments in technologies for maternal, newborn, and child health. He was an Assistant Professor and a Founding Faculty Member with the Department of Engineering, Boston College, in 2021. His current research interests include AI/ML-based image analysis and user guidance, with a particular focus on ultrasound imaging and biomedical instrumentation development.