IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 September 2023, accepted 26 October 2023, date of publication 6 November 2023,
date of current version 13 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3330465

RESEARCH ARTICLE

An Explainable AI System for Medical Image Segmentation With Preserved Local Resolution: Mammogram Tumor Segmentation

AYA FARRAG[1], GAD GAD[2], (Student Member, IEEE),
ZUBAIR MD. FADLULLAH[2], (Senior Member, IEEE),
MOSTAFA M. FOUDA[3], (Senior Member, IEEE),
AND MAAZEN ALSABAAN[4]
[1]Department of Computer Science, Lakehead University, Thunder Bay, ON P7B 5E1, Canada
[2]Department of Computer Science, Western University, London, ON N6G 2V4, Canada
[3]Department of Electrical and Computer Engineering, Idaho State University, Pocatello, ID 83209, USA
[4]Computer Engineering Department, College of Computer and Information Sciences, King Saud University, Riyadh 11451, Saudi Arabia

Corresponding author: Mostafa M. Fouda (mfouda@ieee.org)

This work was supported by King Saud University, Riyadh, Saudi Arabia, through Researchers Supporting, under Project RSPD2023R636.

ABSTRACT Medical image segmentation aims to identify important or suspicious regions within medical images. However, many challenges are usually faced while developing networks for this type of analysis. First, preserving the original image resolution is of utmost importance for this task where identifying subtle features or abnormalities can significantly impact the accuracy of diagnosis. While introducing the dilated convolution improves the resolution of the convolutional neural network (CNN), it is not without shortcoming, i.e., the loss of local spatial resolution due to increased kernel sparsity in checkboard patterns. To address this shortcoming, we conceptualize a double-dilated convolution module for maintaining local spatial resolution while improving the receptive field size. Then, this approach is applied, as a proof-of-work, to tumor segmentation task in mammograms. In addition, our proposal also tackles the class imbalance problem, originating at the pixel level of the mammogram screenings, by identifying and selecting the best candidate among a number of potential loss functions to facilitate mass segmentation. We also carry out quantitative and qualitative evaluations of the interpretability of our proposal by leveraging Grad-CAM (Gradient weighted Class Activation Map). We also present a comparative performance evaluation with existing explainable techniques tailored for segmenting images. Moreover, an empirical assessment on lesion segmentation is conducted on mammogram samples from the INBreast dataset, both with and without incorporating our envisaged dilation module into CNN. The obtained results elucidate the effectiveness of our proposal based on mass segmentation performance measures, such as Dice similarity and Miss Detection rate. Our analysis also promotes using the Tversky Loss function in training pixel-imbalanced data and integrating Grad-CAM for explaining image segmentation results.

INDEX TERMS Medical image analysis, tumor segmentation, mammograms, CAD, deep learning, CNN, explainable AI, spatial resolution, pixel-level class imbalance.

I. INTRODUCTION

Recently, the wide adoption of data-driven models has led to a significant increase in the exploration and advancement of AI (Artificial Intelligence)-assisted Computer-Aided Diagnosis (CAD) systems. Radiologists, leveraging AI-assisted CAD systems, may be able to combine computer insights with their expertise, enabling more precise and prompt assessment. Such smart systems often work in tandem with biomedical imaging techniques, such as X-rays, CT

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

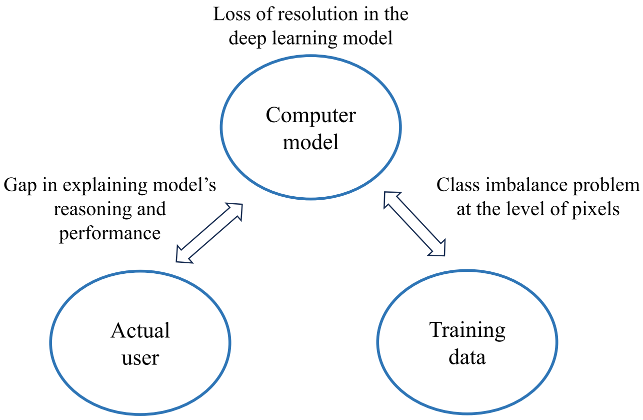


FIGURE 1. The key challenges of biomedical image segmentation addressed in tour research involving interdisciplinary understanding aiming to explain the obtained analytic results in the lens of a domain expert or a familiar user.

scans, and MRI samples [1], [2], [3]. A key element of emerging CAD systems is medical image segmentation, wherein the target input image is split into unique regions, i.e., segments, each of which represents a specific anatomy-related or disease-specific feature. For instance, in oncological CAD frameworks, tumor segmentation, a specialized branch of image segmentation application, plays a critical role in pinpointing and outlining tumor boundaries within the images. This aids in evaluating numerous tumor attributes, including its size, contour, and volume [4]. Such insights prove instrumental in ensuring accurate diagnoses, monitoring tumor progression, and assessing therapeutic efficacy.

Despite its transformative potential in healthcare, medical image segmentation remains among the more intricate facets of computational health analysis [5], [6]. Some inherent complexities stem from the unique and intricate anatomical representations in medical images, which can showcase vast inter-individual variations [7]. The segmentation process is further compounded when tissue or organ boundaries blur, or when tissue presentation alters due to conditions like inflammation or pathology. Additionally, there are hurdles tied to the learning mechanisms and the inherent opaqueness of commonly-employed Deep Learning (DL) models for this task. A conventional medical image segmentation setup encompasses three intertwined components: a *model* educates itself from *data* to instruct *human* medical practitioners. In our research, we tackle three prevalent challenges in DL-centric systems involving biomedical image segmentation tasks. These issues are not only associated with the DL model structures but also may emerge from model understandability or interpretability as outlined in Figure 1. In other words, the key challenges in this domain revolve around the resolution degradation in deep learning models (such as CNN), pixel-level class discrepancies in segmentation tasks, along with the inherent absence of data-driven model transparency (i.e., explainability).

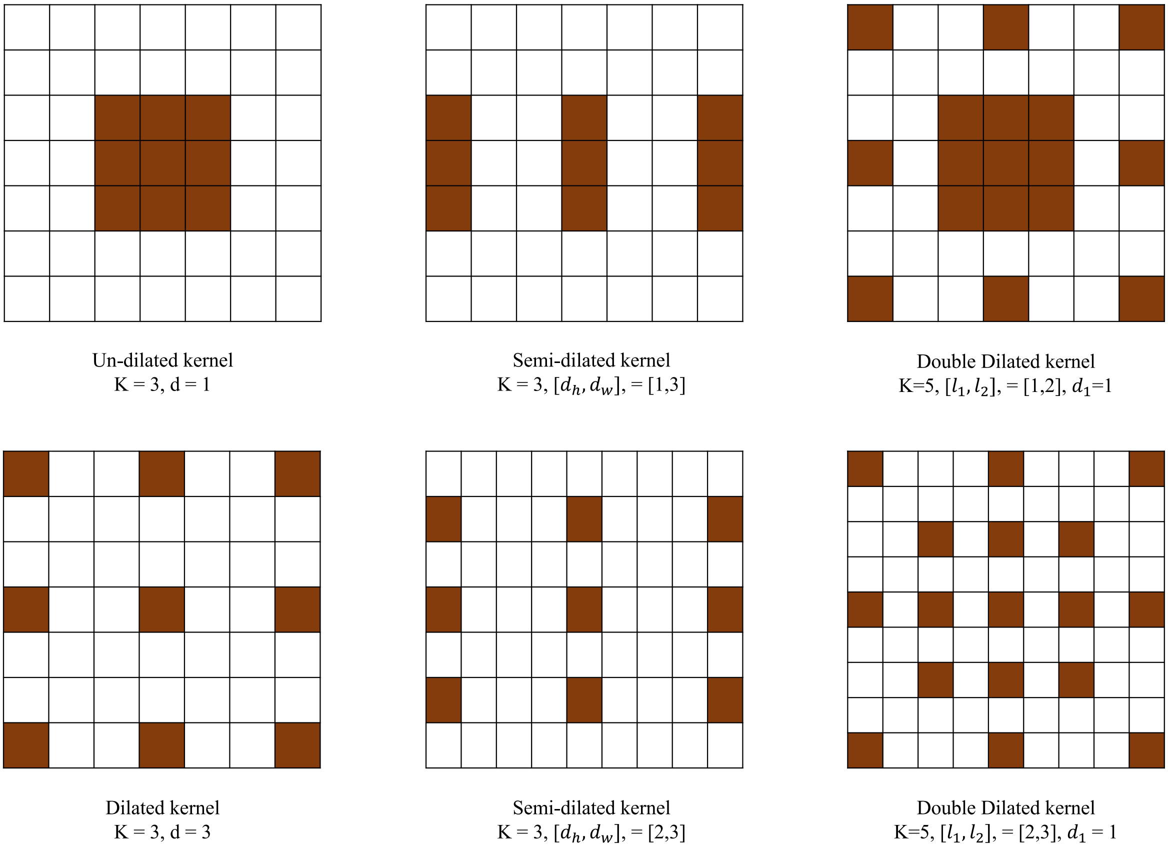
A. RESOLUTION LOSS IN CNN

The application of CNN-based structures has demonstrated commendable performances in areas including image classification and semantic segmentation. In the image classification domain, the objective is to typically identify samples belonging to appropriate classes or cateogries, whereas segmentation involves classification at the pixel level to detect the interest worth of interest from the point of view of radiologists and caregivers. Traditionally, hierarchical deep learning architectures which were initially conceptualized to handle classfication tasks are nowadays customized to handle segnmentations tasks also.

In an image classification system, a sequence of pooling modules is considered for expanding the network's receptive field and obtain information providing the context of the classification task [9]. Semantic segmentation systems [10], [11], however, suffers from resolution loss when the same technique is considered because of the hierarchical (i.e., pyramid-like) shape. As a remedy to this problem, researchers have resorted to up-sampling techniques for reproducing pre-pooling resolutions in such semantic segmentation systems. Despite their wide adoption int he literature, researchers in [12] raised their shortcoming in large-scale iamge segmentation tasks since the degradation of input resolution, inherently a problem associated with image classification systems, adversely impacts the image segmentation performance.

As a solution to the aforementioned problem, we consider introducing the dilated convolution technique in dense prediction systems to effectively address the pixel resolution issue. Dilated convolution was initially conceptualized for efficiently deriving the wavelet transform [12], [13]. By extending the original concept of dilated convolution, it may be possible to exponentially grow the kernel's receptive field to reduce or eliminate the need for pooling layers in the underlying CNN model. However, to realize this, the kernel's sparsity also grows exponentially, which may adversely affect the local spatial resolution [14].

Therefore, while carrying out the segmentation task involving biomedical images, it is essential to preserve the originally achievable spatial resolution. Otherwise the complex nature of such data may not be reflected in the trained model. In our research, we attempt to investigate this issue. In this vein, while exponentially expanding the receptive field, we envision a customized, dilated convolution technique, consisting of a dilation parameter which accounts for multiple factors to flexibly control both the inner and outer kernel resolution. Figure 2 depicts illustrations of various kernel shapes facilitating dilated convolution that include standard dilation (bottom left), semi-dilation [8] (middle), and the proposed double-resolution dilation (right). The figure demonstrates how our module can assume unique kernel shapes with a dense kernel but large receptive field to eliminate the "gridding" problem [14]. This phenomenon is observed because of the large dilation factors causing increased sparsity of the underlying kernels.



Un-dilated kernel
K = 3, d = 1

Semi-dilated kernel
K = 3, $[d_h, d_w]$, = [1,3]

Double Dilated kernel
K=5, $[l_1, l_2]$, = [1,2], d_1=1

Dilated kernel
K = 3, d = 3

Semi-dilated kernel
K = 3, $[d_h, d_w]$, = [2,3]

Double Dilated kernel
K=5, $[l_1, l_2]$, = [2,3], d_1 = 1

FIGURE 2. Different convolution dilation modules are compared. A 3 × 3 kernel's standard dilation and undilated state are displayed in the left column. The semi-dilation suggested by [8] is displayed in the center column. Our suggested dilation surgery is displayed in the right column.

B. PIXEL-LEVEL CLASS IMBALANCE

The discrepancy in pixel-level classification represents a predominant challenge for biomedical image segmentation tasks. Notably, in tasks like mass segmentation, there exists a pronounced imbalance in the pixel count across varied classes, such as tumor versus non-tumor regions. For instance, during organ-specific screenings, the pixel count designated as "mass" is frequently dwarfed by those labeled as "normal". Such marked disparities in class distributions compromise the efficacy of Deep Learning models [5].

A widely-adopted remedial approach to this challenge is sample re-weighting. In this method, enhanced weights are ascribed to pixels from underrepresented classes during the training phase [5]. The efficacy of this strategy is modulated by the chosen objective function, which influences the loss computation while training the underlying model. In spite of an appreciable amount of research work on designing deep learning loss functions to circumvent this imbalance issue [15], [16], [17], selecting the optimal loss function with regard to segmenting highly skewed medical image classes remains an open research challenge. As such, our research undertakes a comparative evaluation of the contemporary loss functions, aiming to systematically identify the one with superior performance.

C. LACK OF EXPLAINABILITY

In medical scenarios, it is important for DL models to offer clear explanations behind their predictions, ensuring radiologists and healthcare experts grasp the logic underpinning model outcomes. When delving into medical image segmentation, while these models don't produce direct diagnoses like classification tasks do, they still need to clearly elucidate their performance rationale. It's vital to communicate how and why the model makes specific annotations. If medical professionals cannot comprehend the reasoning behind a system's results, their trust in the system diminishes, leading them to lean more on their individual assessments. Hence, this explainability gap acts as a significant barrier to the wider acceptance and adoption of AI-assisted CAD (Computer-Aided Diagnosis in caregiving setting [18].

In recent times, strides have been made to ensure the transparency of model-driven solutions, particularly in medical contexts. A number of these revamped models have found their place in disease categorization networks, offering justification for predictions made by such opaque DL models [19]. Nevertheless, the quest to augment transparency in segmentation networks is still in its infancy [20], [21], although these also are often impacted by the "black-box" frameworks. Additionally, post-inference model interpretation may be instrumental for researchers to discern whether the model is pinpointing relevant patterns or merely over-optimizing based on training images' irrelevant attributes. Such insights can be crucial in fine-tuning model and hyperparameter tuning to achieve a more robust performance that mirrors real-world data [19]. Thus, this research ventures into the realm of explainability in segmentation, gauging the efficacy of transparency-driven techniques within the domain of medical image segmentation. Our goal is to incorporate explainable AI frameworks within our proposed segmentation model, to provide a transparent and reliable system for seamless incorporation with medical image acquisition devices.

D. MAMMOGRAM SEGMENTATION USE-CASE

In this section, we delineate the importance of tumor segmentation for mammogram screening. According to recent statistics [22], breast cancer is a top contributor to female mortalities globally. Therefore, early identification and accurate diagnosis are critical in providing the patients with appropriate treatment pathways. The integration of AI systems in medical applications is instrumental in reducing mammogram reading times for radiologists, enhancing diagnostic precision [23]. In this paper, we utilize the public INBreast dataset [24] for mammogram analyses to gauge our model's segmentation performance. Our objective is, therefore, to come up with an interpretable, as well as accurate tumor segmentation solution. In this vein, we first introduce an innovative dilated convolution component to retain the input's spatial consistency, then we investigate the issue of class imbalance by systematically identifying and adopting a viable loss function, and integrating explainable AI techniques into our considered image segmentation frameworks.

II. RELATED WORK

In this section, we provide a comprehensive literature survey. First, we provide a review of the state-of-the-art techniques for performing biomedical medical image segmentation. Then, we present the contemporary research work on dilated convolution to address the resolution degradation issue in image segmentation tasks, and describe the research gap. We then provide the recent work on semi-dilated convolution and provide the needed preliminaries.

A. MEDICAL IMAGE SEGMENTATION

Medical images' regions of interest (ROI) delineation is crucial for diagnosis. Numerous CAD solutions for this have been suggested in the literature. Earlier methods involved thresholding, boundary and region-specific image segmentation tasks, as well as the so-called template matching [25]. The Fully Convolutional Neural Networks (FCNN) [26], has surpassed traditional techniques that used manual feature extraction [27], [28], [29]. This advancement has enabled large, trainable models to deduce the best features for segmentation.

In medical imagery, the targeted anatomy often occupies a minor portion of the image. Even if these minor features are vital, training with such data can lead networks to favor the background, resulting in the majority class label often being assigned to all pixels [5]. Recent work tackled this by using balanced pixel batching from both major and minor classes during training [30]. However, this can lead to a loss of vital geometric details. An alternative approach is sampled loss training, wherein only random pixels are considered for loss computation. A limitation of this method is the unpredictability of the chosen pixels [5].

Sample re-weighting, where lesion pixels are weighted more during loss calculation, is another popular solution for class imbalance [31]. Studies like [32] and [33] used a weighted customization of the cross-entropy loss function. Region-specific loss functions were taken into consideration by several researchers [17], [34]. A prominent example of this is the Dice and Tversky loss function. However, it remains uncertain if a universally optimal loss function exists for medical image segmentation. To address mammogram image imbalance, we assess various loss functions to find the most effective for tumor segmentation.

Regarding explainability, various explainable AI (XAI) models emerged to demystify deep neural networks' operations. For example, LIME [35], DTD [36], and LRP [37], have been employed in medical image classification [19], [38], [39]. But, many such techniques necessitate a global classification layer at a CNN's end, making them less feasible for pixel-level tasks.

While much of the XAI work is centered around CNN classifications, only a few integrate this into semantic segmentation networks. In [20], the SHAP method is used for explanations in oil slick segmentation. For autonomous driving systems, [21] utilized the second-order neuron activation derivative to offer attention maps. In medical imagery, a sole study provided explanations for liver tumor CT image segmentation using activation maximization [40]. With regard to the contemporary work, our research in this paper may be regarded as a pioneering effort that integrates explainability with tumor segmentation for mammography, thereby evaluating the efficiency of XAI (explainable AI methods) through entropy and pixel-flipping graphs, inspired by coauthors' earlier investigation in [38].

B. DILATED CONVOLUTION

Drawing from biological investigations, CNNs typically utilize a hierarchical pyramid-shaped structure. In this arrangement, pooling layers are applied following convolutional layers to systematically reduce the dimensions of feature maps generated from each convolutional operation. This structural configuration has demonstrated remarkable efficacy in image processing, as well as computer vision tasks. However, when such classical CNN models are directly applied for segmenting images, the original image resolution deteriorates in a significant manner, particularly in biomedical imaging segmentation and analytics [41].

Previous works have addressed the resolution degradation in segmentation tasks due to the use of the hierarchical pyramid-shaped CNN architecture. Noh et al. [42] apply a deconvolution network following the convolution network composed of deconvolution and up-pooling layers to reconstruct a segmentation map of the original spatial size. Another approach generates different versions of the input with different sizes and finally uses the attention mechanism to combine all outputs into one refined segmentation map [43]. While in global prediction CNN tasks (e.g. image classification) invariance to the transformations in the input image is advantageous, segmentation tasks require accurate localization of spatial details. This has motivated the introduction of the dilation convolution [12] to expand the receptive field of the CNN as the input image moves through the network while keeping the number of parameters from exploding. This further cuts down the required number of downsampling and pooling elements across the CNN structure.

The dilated convolution technique [12] exploits a sparse kernel, which exponentially grows, thereby attaining a much broader receptive field. As mentioned earlier, this drastically reduces the need for downsampling/pooling layers, and as a consequence, was an attractive choice for various semantic segmentation models [14], [44]. The conventional dilated convolution architectures, however, are not without shortcoming. For instance, they are associated with the "gridding" [14] phenomenon whereby the convolutional kernel are padded with zero's at a static dilation rate. As a result, a checkboard-shaped receptive field emerges, and only the non-zero locations are considered for sampling contributing to neigboring information loss. This problem intensifies when the dilation rate propagates across the higher layers in the CNN structure. This renders the convolutional kernel useless to capture local information due to its over-sparse nature comprising widely-spaced non-zero weights.

While a number of solutions emerged in the literature to deal with the aforementioned gridding issue, Wang et al. [14] introduced a sequential hybrid convolution, which is worth noting. Their approach is essentially a stack of convolutions with varying dilation rates. A similar strategy is followed by the segmentation model family Deeplab, such as Deeplabv2 [45], Deeplabv3 [46], and Deeplabv3+ [47]. These models demonstrated reasonable performance compared with benchmarks (e.g., PASCAL VOC), and thus have proven to be superior to other models not exploiting dilation [48].

C. SEMI-DILATED CONVOLUTION

Perhaps the most relevant work to our research is semi-dilated convolution [8], which is specifically proposed to leverage the non-uniform shape consisting of images arranged in a rectangular grid referred to as the scalograms configuration. For instance, 100 × 6000 scalograms are considered in [8] to facilitate the exponential growth of the receptive field along a unique direction/dimension of the underlying image. Instead of a single dilation factor, Hussein et al. [8] used a dilation vector that has two dilation factors (d_h, d_w) denoting the dilation factor in the image height direction and the image width direction (i.e. each dilation factor determines how the kernel expands in a certain direction). Semi-dilation convolution for a $K \times K$ kernel can be described mathematically as:

$$\begin{aligned}&(K+(d_h-1)(K-1))\times(K+(d_w-1)(K-1))\\&\quad=((K-1)d_h+1)\times((K-1)d_w+1).\end{aligned}\tag{1}$$

We suggest a refined version of the dilated convolution module called the *double* − *dilated convolution*. This modified version uses a dilation factor for the kernel's core and another dilation factor for the kernel's edges, specifically targeting medical image segmentation tasks. The rationale behind the proposed *double* − *dilated* convolution module is to have a sparse edge, enabling a rapidly expanding receptive field, and a denser core, capturing small details at the core. Our main contributions in this research work may be highlighted as below. First, we conceptualize the *double* − *dilated* convolution technique by taking into account two dilating elements that differentiate the dilation at the kernels' core from that at the edges. Next, we integrate the *double* − *dilated* convolution technique with the appropriate biomedical image segmentation models, which would otherwise use the standard dilation convolution. Then, we demonstrate the superiority of our developed module on a publicly available breast cancer dataset [24] through extensive, comparative analytics.

III. DATA PREPARATION

We have chosen the INBreast dataset [24] from various available mammogram datasets for our analysis. This decision is supported by several factors. In contrast to other publicly accessible datasets that utilize digitized film-screen mammograms, the INBreast dataset stands out as the sole publicly accessible collection of Full-field Digital Mammogram (FFDM) data. As a result, it contains high-resolution images that remain consistent, as they avoid any inconsistencies that could arise during the digitization process. Furthermore, this dataset includes meticulously drawn pixel-level outlines of lesions by radiologists, rather than offering just circular regions of interest (ROIs) like most other databases do. This distinction can be of paramount importance in the diagnostic process, as the shape of a mass is a strong indicator of its malignancy [49]. The reason behind selecting this dataset is the additional benefit it provides in terms of images with multiple views, namely MLO (mediolateral oblique) and CC (craniocaudal). Our considered dataset comprises 410 dual-view images (with resolution of 3328 × 4084 or 2560 × 3328 pixels), obtained during 04/2008-07/2010, based on the FFDM (Full-field Digital Mammogram) system provided by MammoNovation Siemens. The resolution varied due to the use of compression

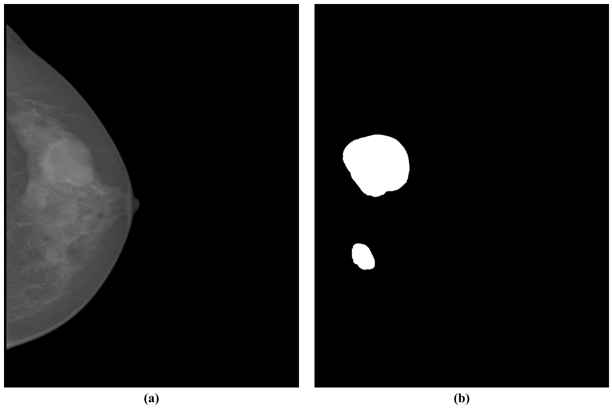


FIGURE 3. An example from the INBreast dataset is shown, which includes a mammography showing the left breast picture from the craniocaudal (CC) perspective. This is accompanied with a produced mask that shows the presence of masses. Panels (a) and (b), respectively, demonstrate the original DICOM-based image and our processed image, i.e., the constructed mask.

plate while acquiring the image. The images captured a myriad of observations are present, including normal breast tissue, calcifications, asymmetries, presence of masses, multiple findings, and artifcats/architectural distortions. For the scope of our research, we concentrate on accurately outlining masses, which are three-dimensional structures characterized by outwardly convex borders, following the definition outlined in the BI-RADS (Breast Imaging Reporting and Data System). Specifically, among the images, 107 contained at least one mass lesions instance.

Note that the mammogram samples (i.e., the acquired images through the earlier mentioned acquisition system) were formatted with the widely adopted Digital Imaging and Communications in Medicine (DICOM) standard. On the other hand, the annotations for each image were stored in an XML file that outlines the Regions of Interest (ROIs) present within an image. These annotations are represented as contour point lists with reagrd to every ROI, differentiated by labels, such as "calcification" and "mass". To facilitate this process, we developed a Matlab script. This script first reads through the data from a XML file. Then, the script retrieves the annotation data, which are relevant to the masses. Finally, it generates contour outlines. Subsequently, these mask images are converted and saved as PNG files for ease of visualization along with the original images. Figure 3 depicts an example of the original DICOM image and its respective mask image. For effective management, we employ Matlab's ImageDatastore and PixelLabelDatastore classes to load both the original and mask images. We then convert all image files into a uniform dimension of 512 × 512 pixels that renders minimal loss of information. For our conducted experiments, we adopt a five-fold split for model validation that will be described in detail in the following section.

IV. PROPOSED METHOD

With regard to the processed dataset involving the breast cancer segmentation use-case, appropriate image segmentation methods/models need to be designed. In this vein, in this section, we present our envisioned methodologies to effectively carry out the breast cancerous tumor segmentation. First, we delineate our envisioned double-dilated convolution concept and underscore its uniqueness from the contemporary dilation methods. Next, we outline the techniques under consideration for achieving balanced class distribution at the pixel level. Then, we detail the procedural framework employed for experimentation and expound upon the interpretability approaches.

A. DOUBLE-DILATED CONVOLUTION

First, we elucidate the alteration we propose for the dilated convolution module originally presented in the work by Yu and Koltun [12] to enhance the local spatial resolution. For a kernel sized $K \times K$, the conventional convolution operation is defined as:

$$y[i] = \sum_{k=0}^{K} x[i+k]w[k] \tag{2}$$

Here, the output feature map is denoted as y, the input feature map as x, and the kernel as w. In this context, K signifies the kernel's dimensions, while k serves as an iterating variable across the kernel's positions.

The concept of dilated convolution enhances the conventional convolution mechansims to the following,

$$y[i] = \sum_{k} x[i+l.k]w[k], \tag{3}$$

where l denotes a dilation parameter/factor. It is worth noting that the kernel remains unchanged to accommodate the sparse effect brought about by dilation. The operation skips a specific range over the input by a factor of l.

Next, we propose a variant of dilated convolution where we adopt a piece-wise function as follows,

$$y[i] = \begin{cases} \sum_{k} x[i+l_1.k]w[k] & j \leq d_1 \\ \sum_{k} x[i+l_2.k]w[k] & j > d_1. \end{cases} \tag{4}$$

In this context, $j = |\frac{K}{2} - k|$, and d_1 refers to the radius of the inner kernel. Instead of using l as in the classical dilated convolution mechanism, our dilated convolution variant, referred to as *double dilated* or *double atrous*, employs two distinct dilation factors, l_1 and l_2, enabling greater control over the kernel's sparsity rate.

The receptive field of our proposed double dilation module can be derived from equation 1 by [8] as:

$$\begin{aligned} &(K + (l_2 - 1)(K - 1)) \times (K + (l_2 - 1)(K - 1)) \\ &= ((K - 1)l_2 + 1) \times ((K - 1)l_2 + 1) \end{aligned} \tag{5}$$

The double dilated convolution utilizes d_1 to establish the size of the inner (central) kernel, which spans an area of $(2d_1 + 1)^2$, and employs a dilation factor of l_1. Conversely, the dimension and dilation factor of the outer kernel are $(2K + 1)^2$ and l_2, respectively. We rationalize this modification to address the intrinsic issue of *gridding* as outlined in [14] in traditional dilation techniques. Since the non-zero weights are widely distanced (i.e., sparse), the receptive field of the dilated kernel only covers regions with checkerboard patterns. This effect becomes particularly pronounced as the dilation factors assume high values. Let us refer to Figure 2, whereby the coverage of the receptive field is related to its size that comprises the non-zero weight values. On the other hand, with growing dilation rates, sparsity occurs rendering the coverage ineffective even though the receptive field expands exponentially. Admittedly this may be justified near kernel-edge(s) because of the larger coverage area dominating the leakage (i.e., loss) of adjacent/local information. On the other hand, keeping a uniform dilation rate to cause such an information outage through all the segments of the kernel, specifically at its core, is not viable. This limitation restricts the capacity of the underlying CNN model extracting the relevant features. This problem may be mitigated by employing a more densely packed kernel core.

B. PIXEL-LEVEL CLASS BALANCING

The primary objective in deep learning models is to perform minimization on a selected loss function. This minimization is required to mitigate the gap between the model's predicted and actual (i.e., true) outputs. Loss functions adopted in neural network-based learning in a large number of layers (i.e., in deep learning structures) may be either distribution-based or group-based. In case of the former, the dissimilarity between predicted and actual output probability distributions is evaluated over the entire input domain. The latter, on the other hand, asseses the output gap within a particular input space zone [50].

In situations where there is a significant disparity in class distribution at the pixel level, such as the problem we are considering in this work, employing a conventional loss function like binary cross-entropy may yield subpar results. This is because the model tends to exhibit bias towards the more prevalent class, resulting in high and low levels of accuracy for the normal and tumor classes, respectively. Given that choosing the loss function is subject to the nature of the specific task and the characteristics of the training data, we empirically evaluate the two types of loss functions (i.e., region- and distribution-specific) to select the optimal one for model training.

To ascertain the most suitable loss function for addressing the class imbalance challenge observed at the pixel level in our chosen dataset, a comprehensive review of contemporary loss functions is provided that are typically employed in segmentation tasks. Based on our finding, particularly from [50], we selected two distribution-based and two region-specific loss functions. The former category consists of the binary and weighted cross-entropy while the latter type comprises dice loss and Tversky loss, respectively. Additionally, these four loss functions have demonstrated effectiveness in the context of skull segmentation tasks as documented in [50].

Next, we provide the necessary preliminaries of these loss functions. Let the total number of pixels in a given batch be denoted by N. g_i and p_i indicates the ground truth value and the predicted probability of a pixel associated with a given lesion. Note that g_i can assume a binary value and is set to be 1 (one) for lesion pixels alone.

1) CANDIDATE 1 LOSS FUNCTION: BINARY CROSS-ENTROPY (BCE)

BCE, introduced in [51], measures the disparity between a pair of probability distributions associated with a random variable. This method is widely used in classification tasks where binary labels are involved. Given that image segmentation essentially entails pixel-level classification, BCE is a commonly employed loss function in segmentation tasks as well [50], [52]. Note that the sample reweighting is not considered in the original version of BCE. However, we consider sample reweighting in BCE as a reference point for the sake of performance comparison. The reason behind this is its adoption in deep learning model training. BCE is formulated as:

$$L_{BCE} = -\frac{1}{N}\sum_{n=i}^{N}(g_i log(p_i) + (1 - g_i) log(1 - p_i)). \quad (6)$$

2) CANDIDATE 2 LOSS FUNCTION: WEIGHTED CROSS-ENTROPY (WCE)

WCE is a refined version of BCE, which was described earlier. The positive class examples are assigned specific weights in WCE through a coefficient to address the imbalanced distribution present in the training data, as discussed in [15]. WCE may be formulated as follows,

$$L_{WCE} = -\frac{1}{N}\sum_{n=i}^{N}(\beta * g_i log(p_i) + (1 - g_i) log(1 - p_i)), \quad (7)$$

where β is a hyper-parameter, which can be tuned for amplifying the weights of positive examples by assigning values greater than 1 to it.

3) CANDIDATE 3 LOSS FUNCTION: DICE LOSS

Derived from the Dice coefficient, which is a widely recognized metric for quantifying the similarity between two images, the Dice Loss was introduced in [16] with a specific focus on segnmenting images. The Dice coefficient is comparable to the F1-score, i.e., the harmonic mean of recall and precision. Thus, the Dice coefficient can be regarded as a performance measure for a binary classifier. Therefore, if we consider a binary image segmentation task, the Dice Loss function can be expressed as,

$$L_{Dice} = \frac{1}{N}\sum_{n=i}^{N} 1 - \frac{2p_i g_i}{p_i^2 + g_i^2}. \quad (8)$$

4) CANDIDATE 3 LOSS FUNCTION: TVERSKY LOSS (T)

The Tversky loss (T), as described in [17], can be viewed as an extended form of the Dice loss function, which we covered in the preceeding section. In this case, α and β parameters are included for controlling the contribution (i.e., importance) of false positives and false negatives, respectively. The Tversky loss function may be written as follows,

$$L_{Tversky} = \frac{1}{N}\sum_{n=i}^{N} 1 - \frac{p_i g_i}{p_i g_i + \alpha p_i(1-g_i) + \beta(1-p_i)g_i}. \tag{9}$$

Note that $\alpha + \beta = 1$ leads to a $F\beta$ scores set, allowing for the adjustment of the balance between precision and recall, as outlined in [17]. Also note the condition $\alpha = \beta = 0.5$ that simplifies it to the Dice Loss function. By considering β values of [0.5, 1], it is possible to more strongly control the false negatives, and thereby optimize lesion segmentation.

It is worth mentioning that we conduct hyper-parameter tuning for all the candidate loss functions discussed earlier.

C. EXPERIMENTAL METHODOLOGY

From hereon, we empirically evaluate the performance of proposed techniques. All our reported experiments were conducted using Matlab R2022b.

1) BASELINE MODEL

To formulate a baseline model, we first adopt two popular image segementation frameworks, namely the U-Net [10] and DeepLabV3+ [47], which are tailored for similar task, i.e., brain tumor segmentation [53]. While the former primarily employs a traditional CNN structure, the latter leverages the ASPP (Atrous Spatial Pyramid Pooling) module on top of a CNN architecture, and makes use of the dilated convolution technique. Although DeepLabV3+ was not explicitly designed for medical image segmentation, it demonstrated superior performance with our breast cancer data. Consequently, DeepLabV3+ with the default cross-entropy loss function was selected as our baseline model. Furthermore, the encoder-decoder structure of DeepLabV3+ [47] appears useful, since the encoder and decoder components utilize multi-scale features and retrieve spatial resolution, respectively. Figure 6 depicts the encoder element of the vanilla DeepLabV3+ framework. Here, an output stride of (8) is considered for illustration purpose. We incorporate our envisioned dilation module to the convolution layers residing at the encoder of this framework. Moreover, the atrous convolution comprising varying rates is applied to the cascaded layers, as well as the spatial pyamid pooling layers [46]. This results in the expansion of the receptive field and inclusion of the multi-scale context without a significant loss of the resolution of the image. Thus, our modified framework is able to produce output maps with a resolution downsampled only by a factor of 8 in contrast with a downsampling ratio of 256 required by conventional pyramid-shaped convolution structures.

2) DOUBLE-DILATED CONVOLUTION

We outline a proof-of-concept implementation in figure 5 for the proposed $double - dilation$ convolution. Unlike the standard convolution operation which applies a single dilation factor across the kernel, our proposed $double - dilated$ convolution differentiates the dilation factor at the kernel's core from that along its edges. Our envisioned $double - dilated$ convolution hinges upon the distributed property of the convolution process as follows. Assume that $h_1[n]$, $h_2[n]$, and $x[n]$ represent three distributed functions such that,

$$x * h_1 + x * h_2 = x * (h_1 + h_2). \tag{10}$$

Figure 5 depicts our implementation of the $double{-}dilated$ convolution operation which deploys two unique kernels having their corresponding dilation factors. Note that both these kernels opearte on the same input. Also notice in the figure that the resultant outputs from each operation are merged to derive the final outcome.

Originally, DeeplabV3, which utilizes the Resnet18 as a backbone model, introduces the ASPP module which combines simultaneous convolutions with multiple dilation factors. To evaluate the segmentation efficiency of $double - dilated$ convolution and quantify the enhancement resulting from applying this convolution, it was integrated into DeeplabV3+ by substituting standard convolutions in both the backbone model and the ASPP with $double - dilated$ convolution modules as displayed in Figure 6. We call the modified ASPP module as Double Atrous Spatial Pyramid Pooling (DASPP). For training the modified Deeplabv3+, we follow the training setup in [46]. The model was trained with the learning rate of 1e-2 and a batch size of twelve over thirty epochs.

3) PIXEL-GRAINED CLASS BALANCING

Here, we comparatively assess various candidate loss functions in terms of their robustness against pixel-level imbalanced classes. In this vein, we generated distinct variants of our model for each candidate loss function, each of which incorporates the corresponding loss function in the pixel-level classification layer of the network. In cases where the function includes hyper-parameters, such as class weights, we initially refined the model using a five-fold validation approach to identify the most optimal parameter combination for that specific loss function.

Table 1 lists the hyper-parameters for the considered loss functions. For BCE, we did not modify the original classification layer found in the Deeplabv3+ framework. For WCE, the parameter range was derived by taking into account the ratio of the number of pixels in the two classes without and with lesion, respectively. This revealed that pixels describing no-lesion class occurred approximately 70 times more frequently than pixels describing lesions or masses. Consequently, the minority class weight was varied

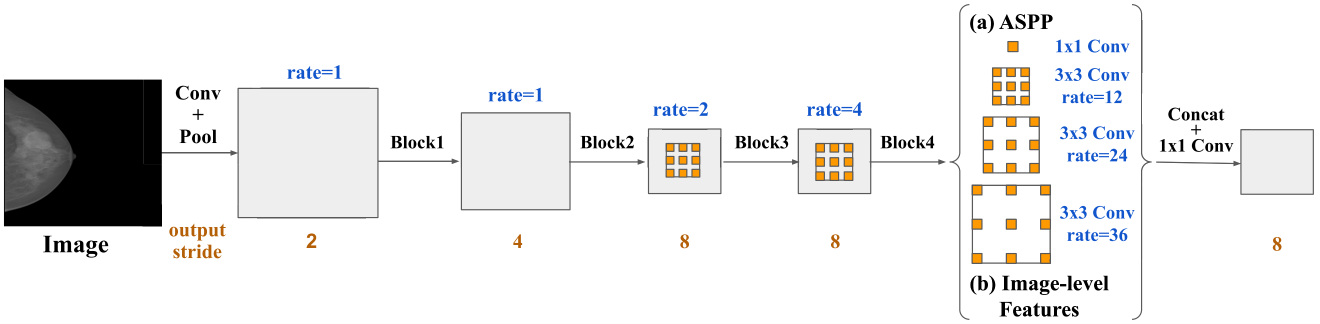


FIGURE 4. An illustration of the original Deeplabv3+ encoder architecture applied in this work. Note that the last two blocks of the ResNet backbone network and the parallel modules of the Atrous Spatial Pyramid Pooling (ASPP) module also use dilated convolution, albeit at different speeds. The resulting feature maps are produced by adding image-level features to the ASPP output. This figure shows the applied dilation rates with an output stride of 8, which has been adjusted based on the work in [46].

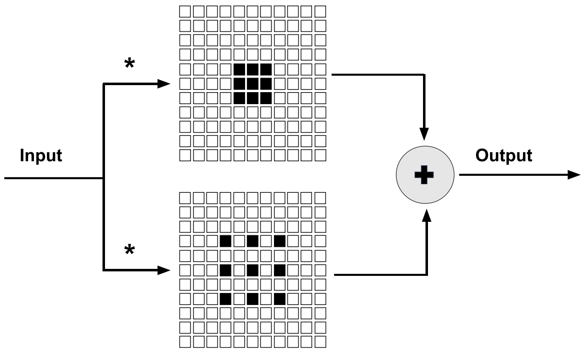


FIGURE 5. A straightforward method that employs two parallel convolution branches. The first branch is for the sparse outer kernel (rate=2) and the other is associated with an inner dense kernel (rate=1). This method aims to produce a double-dilated convolution with 1 and 2 dilatation rates. The output obtained from their feature maps is equivalent to performing the double-dilated convolution when added together.

TABLE 1. The parameters used in various loss functions to address the problem of uneven classes at the pixel level. Class2 represents a mass lesion, whereas Class1 represents the non-lesion class. The weights assigned to false positives and false negatives in the Tversky loss function are controlled by alpha and beta, respectively.

	Weighted Cross-Entropy	Tversky
Parameters	[class1_weight, class2_weight]	[alpha, beta]
Range	[1, 10], [1,20].. [1,100]	[0.1, 0.9], [0.2,0.8], .. [0.5,0.5]

from ten to one hundred in steps of ten. This allowed us to validate the varying weighting ratios and thererby choose the optimal value.

Regarding both the Dice and Tversky Losses, where the latter is an extended version of the former, a customized pixel classification layer was formulated to exploit the Tversky loss from eq. 9. For Dice Loss and Tversky loss functions, $\alpha = \beta = 0.5$ and $\beta = [0.5, 1], \alpha = (1 - \beta)$, respectively. The latter permits the Tversky Loss function to prioritize the minority class.

D. EXPLAINABLE AI (XAI) METHODS

To offer the actual explanation that corresponds to the obtained breast cancerous tumor segmentation results in our envisioned models, we integrated explainable AI techniques into our proposal. Initially, a simple yet effective method was developed, termed as the ''activation visualization'' in the remainder of the paper. The aim of ''activation visualization'' is to inspect the activation state of the last feature extraction layer in the image segmentation model, and thus corroborate the performance yielded by our model. In other words, this approach reveals the features learned by the network. This is made possible by checking and comparing the points/regions of activation with regard to the original (input) image. It is worth noting that the final six channels of the last deconvolution layer within the DeepLabV3+ provide interesting insights into the class prediction (i.e., decision making) process, and therefore, we draw particular attention to those channels. Next, we extend our ability to explain the model performance by applying two popular xAI techniques, namely Gradient-weighted Class Activation Mapping (Grad-CAM) [54] and Occlusion Sensitivity [55], both adhering to the aforementioned principle of visualization approaches.

1) GRAD-CAM

Grad-CAM [54] allows users to visualize the zones in the input image which carry most contributions toward the underlying deep learning model's prediction for a specific class. When a CNN makes a prediction, it performs a series of convolution operations to extract features from the input image. Grad-CAM uses the gradients of the predicted class score with respect to the final convolutional layer's feature maps to determine the importance of each spatial location in the feature maps for that prediction. As a consequence, a heatmap is generated comprising the most relevant regions within the input image toward the model outcome/decision.

The main idea behind Grad-CAM is that the final convolutional layers are expected to capture high-level semantics while preserving spatial information. Typically, the

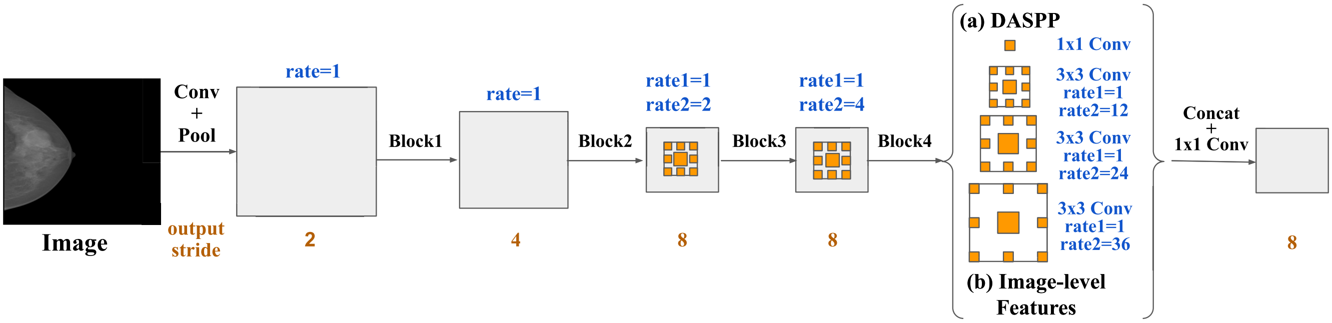


FIGURE 6. After the double-dilated convolution module is plugged in, the Deeplabv3+ architecture is updated. Two parallel convolutions are used to replace each dilated convolution from the original network: one is an undilated convolution (rate1), and the other is a dilated convolution (rate2) with the same rate as the original layer.

kernels learned in these layers are specialized in detecting class-specific features within the image. Hence, Grad-CAM leverages gradient information flowing into these last convolutional layers of the CNN to determine the significance of a feature map with respect to a particular class c. Assuming that the last convolution layer produces K feature maps, where A^k is the k^{th} activation map, with each element of the map having a location (i, j), the formula to calculate the class-specific importance weights (α_k^c) for the feature map is provided as follows:

$$\alpha_k^c = \frac{1}{z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k}, \tag{11}$$

where y^c is the score for class c and A^k is a feature map activation produced by the last convolution layer. Finally, a weighted combination of forward activation maps followed by a ReLU is performed to calculate the class-specific Grad-CAM localization map M^c:

$$M^c = ReLU(\sum_k \alpha_k^c A^k) \tag{12}$$

In their paper, the authors of Grad-CAM explained how the output of the CNN-based network y^c does not need to be a global class score for a successful application of their approach. This allows the integration of Grad-CAM in different types of networks, other than image classification networks, without making any architectural changes, including pixel-level classification problems such as image segmentation. In our case, the class of interest (c) is the lesion class and the output (y^c) is a pixel-level score that determines whether or not a pixel in the mammogram screening belongs to a mass region. In our experiments, we adopted an off-the-shelf implementation of Grad-CAM provided by Matlab by using the *gradCAM()* function to explain the output of the segmentation network.

2) OCCLUSION SENSITIVITY

The Occlusion Sensitivity [55], on the other hand, is a perturbation-based technique that systematically occludes different parts of an input image. By doing so, this technique attempts to investigate the influence of occlusion on the deep learning model's decision. When the classification score for a certain class significantly changes due to occluding a specific patch of the input image, this indicates the significance of this patch in the decision-making process.

In our image segmentation problem, assuming the function learned to classify a lesion in an input image x is denoted as $f(x)$, then the occlusion sensitivity score (S^i) for the i^{th} patch in an input image can be expressed as follows,

$$S^i = f(x) - f(x \odot \Omega^i), \tag{13}$$

where Ω^i is a mask that occludes the patch number i in the image by estimating the element-wise product of the mask and the input image. The scores of different patches of the image can be utilized to construct a heatmap, which highlights the areas where the function experiences the most significant decrease due to occlusion.

Although there are different shapes of masks that can be integrated into the occlusion process, we used the default settings adopted by Matlab's *occlusionSensitivity()* function which we utilized in our interpretability experiments. These occluding masks covered 20% of the input size with a step size of 10% of the input size and replaced the occluded pixels with the channel-wise mean of the input image.

3) EXPLAINABILITY EXPERIMENTS

In addition to employing the activation visualization technique, we experiment with integrating both of these methods into our networks to generate heatmaps indicating the regions in the image that represent mass segmentation during the process of mammography. We apply these techniques to the original Deeplabv3+ framework and also to our modified deep learning model comprising double-dilated convolution so that we may comparatively validate the explanations for the carried out segmentation by both models. Subsequently, we conduct a quantitative assessment of all the explanatory methods by assessing image entropy and constructing pixel-flipping graphs, following a similar approach as seen in prior research [38], [56], [57]. Note that the entropy metric provides an idea on the complexity associated with the xAI method in terms of uncertainty or randomness within the produced explanation map. On the other hand, the pixel flipping approach may also be useful in assessing if skipping the features contributing highly toward the model decision actually affects the network's predictive performance or not.

To compute entropy values, we used Matlab's built-in function called "entropy". This function calculates the entropy values of the image for various explanation maps produced across all validation set images. Then, the average of these entropy values was calculated that quantifies the complexity. On the other hand, Algorithm 1 was developed for deriving the pixel flipping effect on the model outcome that estimates the similarity scores during various stages of pixel-flipping. This involves the iterative removal of input features, starting from the most relevant and progressively moving towards the least relevant. We continue this process until the distortion of 10% of the image pixels. Throughout this process, we monitor changes in the segmentation model's predictive performance. Next, we plot the drop in similarity scores whereby a sharp drop demonstrates a reliable explanation method, reflecting the model's prediction performance. By averaging the plots over the whole validation dataset, we then present a robust evaluation of the reliable explanation offered by our proposed Algorithm 1.

Algorithm 1 Similarity Scores Derivation Via Pixel Distortion (Flipping) Technique.

```
Input: Segmentation network ($net$), explanation map ($map$), input imaging data ($img$), and ground truth image ($gt$).
Output: A list of similarity scores for varying rates ($scores$) of pixel flipping.
procedure pixelFlip($map, net, img, gt$)
    $img_{segmented}$ is assigned to $net(img)$
    $score_{initial}$ is set to $similarityScore(imgsegmented, gt)$
    $scores$ is initialized to []
    $idx_{sorted} \leftarrow arg$ (sorted elements in $map$)
    $idx_{start}$ is set to 1
    for i in 1:10 do
        $idx_{end} \leftarrow idx_{start} + i/100 * size(img)$
        $img(idx_{sorted}(idx_{start} : idx_{end})) \leftarrow radndom()$
        $img_{segmented} \leftarrow net(img)$
        $scores(i) \leftarrow similarityScore(img_{segmented}, gt)$
    end for
    $scores_{normalized} = scores/score_{initial}$
    return $scores_{normalized}$
end procedure
```

V. PERFORMANCE EVALUATION

In this section, we first describe the performance metrics for validating our proposed technique. Then, we report the experimental results and discuss our findings in detail.

A. METRICS FOR EVALUATING THE SEGMENTATION PERFORMANCE

To offer a robust performance analysis, we evaluate the performance of the tumor segmentation task in both pixel and lesion levels. First, the pixel-level evaluation measures the model's accuracy in predicting individual pixels within the image, whereas the lesion-level evaluation provides high-level metrics to assess the capability of the underlying model to predict lesions within the imaging data.

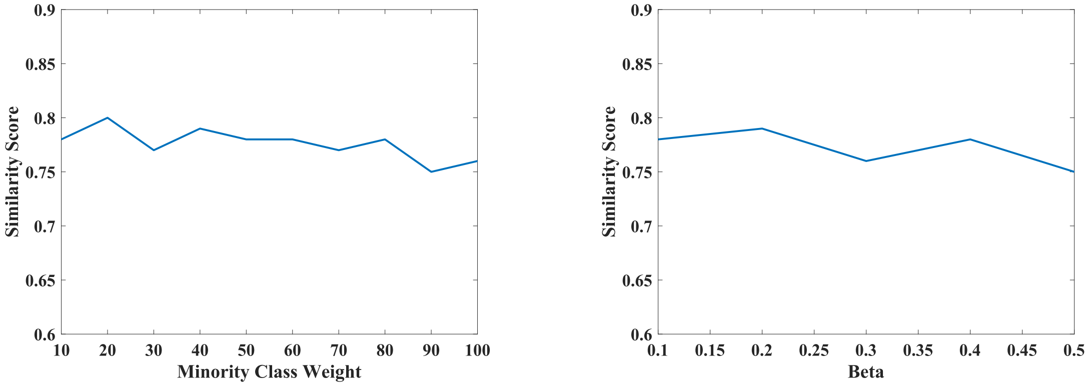
1) PIXEL-GRAINED PERFORMANCE EVALUATION METRICS

As mentioned earlier regarding the semantic segmentation tasks, the quality of a segmentation model is typically determined by the pixel-grained similarity between the segmented image and the reference image [58]. Two well-known performance metrics for evaluating medical volume segmentation tasks are the Jaccard similarity [59]and Dice similarity [60], which estimates the fraction of correctly matched pixels given all the pixels and the ratio of correctly and inaccurately matched pixels, respectively. We calculate both similarity scores separately per class to comprehensively capture the performance of the model in terms of identifying the lesion class, as well as the non-lesion class. Researchers in [61], however, indicated that including both metrics as validation measures does not offer additional insights, as they have overlapping evaluation measures, thereby presenting similar rankings. Since Dice similarity is often preferred for validating segmentation tasks in the medical domain and it is comparable to F1-score for binary segmentation tasks [61], we select this as our preferred pixel-grained performance measure.

Next, we consider an additional performance measure, namely the validation accuracy due to its adoption as a standard performance measure for verifying and analyzing the performance of CAD systems. The validation accuracy considered in our work can be defined as the fraction of correctly classified pixels within the pixel space of the original (input) image. By adopting this metric, we aim to present performance evaluation with regard to comparable studies, e.g., INBreast dataset-based mass segmentation. Note that the validation accuracy could, however, suffer from the class imbalance in our chosen dataset, that could result in overestimated performance toward the non-lesion samples.

2) LESION-LEVEL EVALUATION

We discovered that evaluating measures at the level of mass lesions provides highly intuitive indicators of prediction quality, which makes sense given that the goal here is tumor segmentation. Two metrics, namely the miss-detection rate and the false-positive rate, are proposed for evaluating segmentation models made for this type of imaging data, in accordance with the suggestion made in [24]. The percentage of reference masses that the algorithm fails



(a) Weighted cross-entropy loss.

(b) Tversky loss.

FIGURE 7. **Plot of the five-fold validation sensitivity against the loss function hyper-parameters. (a) shows how the performance of the WCE loss-based model changes when varying the minority class (lesion) wight from 10 to 100. (b) shows the performance of the Tversky loss-based network when tuning the Beta parameter in the range [0.1,0.5].**

to detect is referred to as the miss-detection rate, while the percentage of automatically detected masses that do not match the real masses is known as the false-positive rate. These two metrics offer simple yet effective measures of model performance that radiologists, as well as other caregivers, may easily interpret.

We first ascertain if a detected mass is accurately categorized before computing the two lesion-level metrics. This is achieved by calculating the ratio of overlap between the manually annotated mass and the detected mass. When the overlap (i.e., the intersection over union) between the detected and real lesions exceeds 0.5, the detection is deemed accurate. The number of reference masses that were missed is then used to calculate the count of miss-detected lesions, while the total automatically detected masses are subtracted from the correctly categorized masses to obtain the count of false-positive lesions. We normalize both numbers by dividing them by the total number of masses in the validation set. This yields the lesion-level rates for false positive and miss-detection, respectively. The reason behind our choice for these performance measures can be attributed to their wide adoption in the literature as object detection algorithms in medical imaging and CAD [24].

B. RESULTS AND DISCUSSION

This section assesses how well our envisioned techniques for tumor segmentation in mammography screenings perform. First, we report on the experiments carried out with regard to the baseline model to address the class imbalance at the pixel level. Next, we demonstrate and discuss the effectiveness of our proposed dilation module for convolutional. Then, we describe the efficacy of our chosen XAI (explainable) methods and demonstrate the produced heatmaps for a few arbitrarily chosen mammography segmentation results. The average of the five validation sets produced by the conventional k-fold validation procedure is employed to report all results.

TABLE 2. **Outcomes of a 5-fold validation of the baseline model with various loss functions and adjusted parameters to address the issue of class imbalance at the pixel level. Cross-Entropy (CE), Weighted Cross-Entropy (WCE), Dice (D), and Tversky (T) are the loss functions that are taken into consideration. Evaluation metrics at the pixel and lesion levels are both reported.**

	CE	WCE [1, 20]	Dice	Tversky [0.2, 0.8]
Miss Detection Rate	0.17	**0.08**	0.33	**0.08**
False Positive Rate	0.29	0.29	0.23	**0.20**
Similarity	0.74	**0.79**	0.75	**0.78**
Accuracy	0.98	0.99	0.99	0.99

1) PIXEL-LEVEL CLASS BALANCING

First, we demonstrate how altering the hyper-parameters utilized in loss functions affects the baseline model's performance. We present the five-fold validation sensitivity plotted against various parameters of both functions in Figure 7, since the WCE and T loss functions are the ones that involve parameter tuning. It is evident from Figure 7(a) that performance is not always enhanced by raising the weight of the minority (lesion) class. For the non-lesion and lesion classes, the model in our instance achieved the best similarity at class weights = [1,20]. This elucidates how important it is to adjust the class weights when considering the use of the WCE loss to fine-tune the hyper-parameters for the given scenario. On the other hand, as illustrated in the results of Figure 7(b), altering the beta parameter in the Tversky loss formula to naturally give the less-frequent class greater weight during model training produced a varying validation similarity. This further suggests that when using the Tversky loss, parameter searching is necessary because increasing the β value does not always translate into better performance in highly skewed datasets. When α and β were set to 0.2 and 0.8, respectively, the baseline network configured with Tversky loss achieved the best sensitivity coefficient given our task and data.

TABLE 3. Results of the 5-fold validation of the original Deeplabv3+ model and the modified model with the double-dilated module. Both lesion-level and pixel-level evaluation metrics are reported.

	Deeplabv3+	Double-Dilated Deeplabv3+
Miss Detection Rate	0.08	**0.04**
False Positive Rate	0.20	0.20
Similarity	0.79	**0.81**
Accuracy	0.99	0.99

Next, we look at the outcomes that the original Deeplabv3+ model produced when it was trained with all of the various loss functions and the adjusted parameters. The results are displayed in Table 2. We averaged the outcomes across all validation folds and reported them. It can be observed that while the WCE loss produced a higher overall similarity coefficient and a lower number of miss-detected masses than the standard CE loss, both types of losses exhibited similar False Positive Rates. However, at the cost of missing a noticeably higher number of real lesions, the Dice loss function was able to marginally lower the number of unmatched detected lesions when compared to the Entropy losses. Thus, it is evident that in terms of the pixel-level similarity coefficient and the miss detection rate, the Weighted Cross Entropy loss and the Tversky loss produced nearly identical performances. However, in contrast with the WCE loss, employing the Tversky loss function produced a 9% lower False Positive Rate. This suggests that in segmentation tasks involving highly imbalanced data, the Tversky loss function is able to enhance the precision-recall trade-off. Similar results were found in other studies where the best segmentation results were obtained for detecting brain tumors [62], skin lesions [63], and sclerosis tumors [17], utilizing variants of the Tversky loss function. Thus, in the subsequent experiments, we employ the Tversky loss with the adjusted parameters as the network's loss function.

2) DOUBLE-DILATED CONVOLUTION

The results of the updated double-dilated model and the original Deeplabv3+ model are compared in Table 3. It is evident that the model performed better in terms of both the Miss Detection rate and the Dice similarity after the new convolution module with double dilation rates is introduced. This enhancement can be associated with the proposed double-dilated module's raising of the inner kernels' resolution in order to encode multi-scale context information present in the input image. Early diagnosis could thus be more viable using our adopted approach minimizing the amount of miss-detected masses. For 22 test mamography imaging sets, the snapshot of the radiologist-annotated mammogram images, as well as the automatically-segmented images that correspond to the images in the snapshot obtained by the improved model, are shown in Figure 8.

On the contrary, the number of erroneous (unmatched) lesions predicted by the suggested model remained significant, at 20% of the actual number of tumors, the same rate as the single-dilated network. This is consistent with findings from earlier research that suggested that CAD systems produced comparatively large false positive rates [64], [65], [66]. More integration of 3D mammograms in CAD systems could also improve the performance as suggested by researchers in [67], since the high rates of false positives were found to be triggered by the superimposition of tissues in 2D digital mammography.

Figure 9 provides a closer look at the areas where the modified model outperforms the original one in terms of miss-detection rate. It displays four mammograms from a single validation set that allow us to examine this behavior. It is evident that in these screenings, the masses that the original network failed to identify are comparatively small in size. By using a denser kernel at the convoluion window's core, the double-dilated network, on the other hand, was able to partially segment many of these small masses while maintaining the input image's local resolution across the CNN network. This pattern held true across several screenings with various validation sets. However, some masses, like the bottom lesion that radiologists segmented during screening, were too small to be picked up by either network (i.e., in Figure 9(c)). Additionally, we may observe that in screenings (in Figures 9(a) and 9(d)), the number of falsely detected masses increased while the number of missed masses decreased. This indicates that the modified network misclassified certain other anatomical structures in the screenings as suspicious lesions. Though the overall average false positive rate was the same for both networks, as shown in Table 3, the double-dilated network still has an advantage because of its lower miss-detection rate.

3) EXPLAINABLITY VIA VISUALIZATION

In this subsection, we show the results of integrating different XAI methods in our segmentation experiments. First, we provide some qualitative results by showing examples of the localization maps obtained using different interpretability approaches. Then, we display the results of the quantitative analysis performed as explained the Section IV to evaluate different methods in terms of randomness and faithfulness.

a: QUALITATIVE RESULTS

In Figure 10, we present a selection of heatmaps generated for a single correctly-segmented mammogram screening using the three explanation techniques that were considered: Activation Visualization, Grad-CAM, and Occlusion Sensitivity. This is because it is challenging to visualize the outcomes of applying different explanation models on all images in the validation set. But when other segmented mammography explanation results were visualized, similar comments were made. The network arrived at a final segmentation decision that matches the extracted features, which can be seen by firstly visualizing the activation maps from six channels at the final feature extraction layer. Even though it may

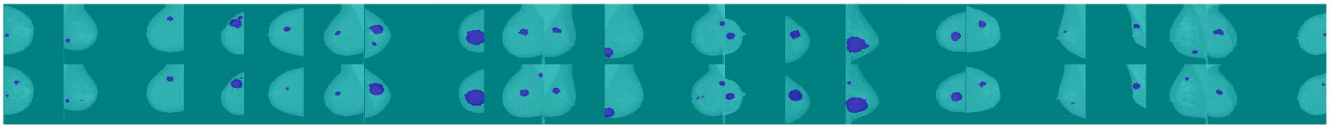


FIGURE 8. A snapshot of segmentation results of 22 validation mammograms generated by the proposed network after plugging the double-dilated convolution module. Images include CC or MLO views from left or right breasts. The top images show the mass annotations made by radiologists while the bottom images display the segmentation done by our CAD model.

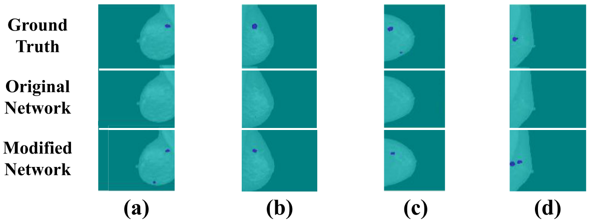


FIGURE 9. Compared to the ground truth segmentation, a selection of segmentation results for four validation mammograms produced by the original and modified networks. For analysis, the four samples are designated as a, b, c, and d in the subfigures.

not be a very convincing explanation for radiologists, this can reassure us as researchers that the network is not overfitting to unimportant attributes of the image while classifying a particular region as a mass. When we contrast the visual outcomes of the two XAI approaches that were used, we find that Occlusion Sensitivity typically produces explanation maps that are hotter than the corresponding heatmap produced by GradCAM. This suggests that more areas of the image are given high positive relevance values for the purpose of segmenting existing masses. The significant distinction in the fundamental methodology between the two approaches could be the cause of this. To determine how much each region contributed to the output decision, several regions are eliminated in Occlusion. Because the removal of those pixels affected the segmentation results, the model may find it easier to assign higher weights to pixels outside of the tumor area. By weighting the gradients of the activation maps at the final layer of our network, which are strongly correlated with the segmentation output itself, the GradCAM maps' red areas are considerably more restricted in the mass region.

We present experimental results of the heatmaps generated by both Grad-CAM and Occlusion Sensitivity methods for a mammogram screening where the model failed to detect an existing lesion in Figure 11 to further investigate the validity of the adopted explainability models in interpreting the segmentation results and highlighting the model's errors. The lack of any heated areas in the Grad-CAM map and the irrelevant/random red patches in the Occlusion map, as depicted in the figure, can both be signs that the mass was not segmented in the model's output because the network did not consider the pertinent features when classifying pixels in this specific screening. Even though it may not be a very convincing explanation for radiologists, this can reassure us as researchers that the network is not overfitting to unimportant attributes of the image while classifying a particular region as a mass. When we contrast the visual outcomes of the two XAI approaches that were used, we find that Occlusion Sensitivity typically produces explanation maps that are hotter than the corresponding heatmap produced by GradCAM. This suggests that more areas of the image are given high positive relevance values for the purpose of segmenting existing masses.

TABLE 4. The original and double-dilated segmentation networks were used to create explanation maps with varying levels of entropy. The validation set's average results for each image are displayed in the table.

	Activation	Grad-CAM	Occlusion
Original	3.154	**0.119**	2.526
Double-Dilated	3.505	**0.139**	1.445

b: QUANTITATIVE RESULTS

Initially, we computed the image entropy of each explanation map created for the mammography screenings in the validation set in order to assess the complexity of various explanation techniques quantitatively. We took into account the activation neurons of the final channel in our computations for the activation visualization method. The average entropy results for various XAI techniques applied to images segmented by the double-dilated network and the original Deeplab V3+ network are displayed in Table 4. The Activation Visualization had the highest average entropy value, according to the results, indicating a very high degree of explanational randomness. When using the modified network, the Occlusion Sensitivity had values distributed over a wide range (e.g., [0.113, 2.089]), with a lower but still relatively high average entropy. On the other hand, the Grad-CAM approach demonstrated the least amount of randomness in explanation mapping, achieving the lowest complexity with an average value of approximately 0.12 in both networks. Additionally, we can see that when both Grad-CAM and Activation maps were used, the suggested network produced slightly more complex explanations than the original network. This might be as a result of the double-dilated kernels' added complexity, which is reflected in the feature maps used in these two approaches' explanation mapping. With the Occlusion maps, on the other hand, this was not the case because the algorithm used is independent of the activation values.

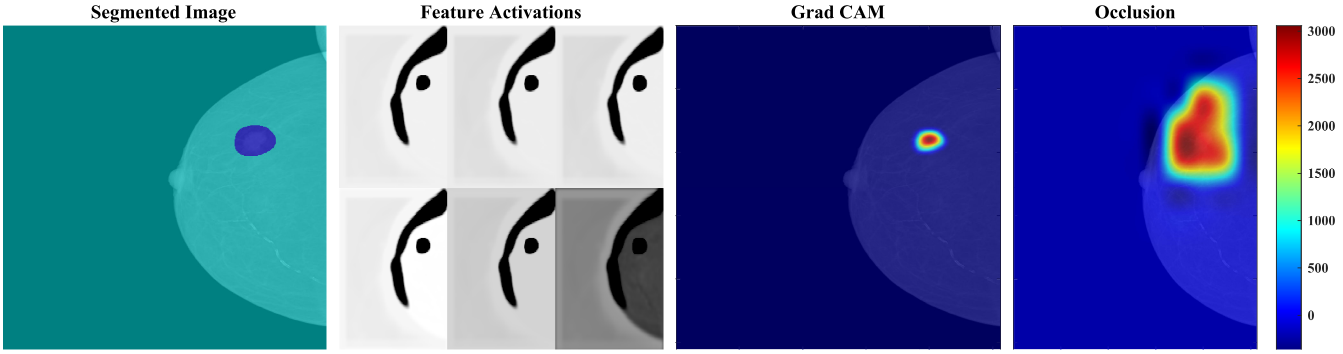


FIGURE 10. Exhibits of accurately segmented mammogram images produced by the altered network are displayed in this illustration, accompanied by descriptions of the output segmentation. Note that GradCAM, Occulsion Sensitivity, and Activation Visualization are employed to display various explanation maps.

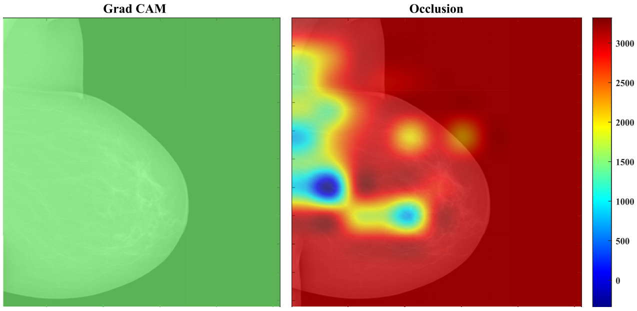


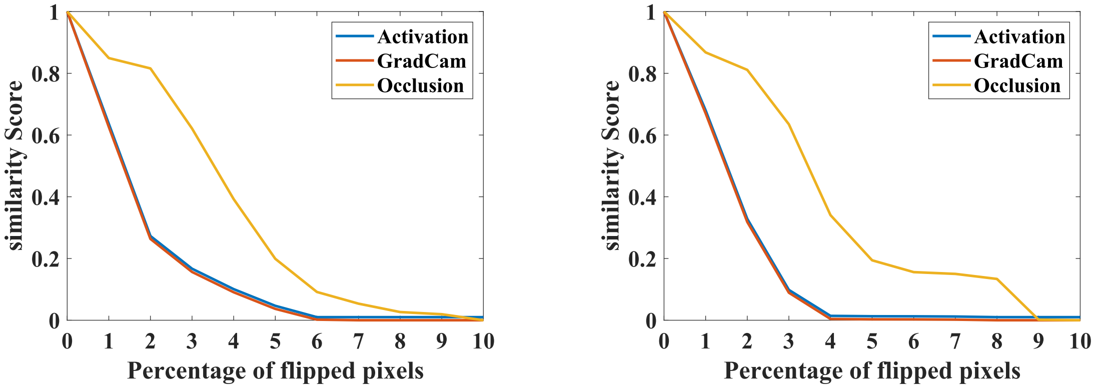
FIGURE 11. Heatmaps produced for an improperly segmented mammography screening in which the model was unable to identify an existing mass by Grad-CAM and Occlusion Sensitivity explainable AI techniques. The generated maps' irregularities and randomness may be signs of inaccurate segmentation outcomes.

Next, we show the pixel-flipping graphs for various explanation techniques using the original and modified networks in Figure 12. The averaged graphs for the entire validation set are represented by the plotted graphs. Given that both techniques primarily rely on the activations of the network's final convolution layer, it makes sense that the Grad-CAM and the visualized activation methods produced the same average similarity scores for varying percentages of pixel flipping, as shown in the figure. Furthermore, the Grad-CAM approach consistently outperformed the Occlusion Sensitivity method in terms of decay rate, indicating that Grad-CAM explanations are more accurate than those provided by the other approach. The Grad-CAM curve decayed more quickly with the double-dilated Deeplab V3+ network than with the original network, despite the fact that this trend is the same in both the original and modified networks. This suggests that the Grad-CAM explanations of the suggested segmentation network are more truthful. This demonstrates how the performance of explanation techniques, as covered in [68], can be impacted by the network structure. Overall, our visualized analysis and the pixel-flipping results concur with the entropy results, indicating that the Grad-CAM explainability technique can offer accurate and understandable explanations for mammogram mass segmentation results. In order to increase their transparency and foster their integration into clinical practice, this encourages the use of this straightforward but effective tool in medical image segmentation networks.

4) COMPARISON WITH PREVIOUS WORK

Additionally, we conducted a survey on the previous work done on the INBreast dataset to compare our methods and results to the existing ones. We found that our selected FFDM dataset was utilized in a wide range of use cases including density estimation [73], image classification [74], tumor segmentation [71] and breast region extraction [75]. In order to provide a relevant comparison with work of the same scope as ours, we only examined the recent papers on the INBreast data that was done specifically for mass segmentation and presented a summary of their approaches as well as ours in Table 5. The table shows the methods used for segmentation and validation split and indicates if any methods were used for augmentation and explainability in each work. Although each work has its separate train and test environments, we still show the test Dice Similarity scores stated in each work, as it was the common metric in all works, in order to analyze the similarities and differences between these works and our work.

It is clear from the table that the results achieved in different works significantly varied due to several factors other than the architecture of the underlying segmentation network. First, it is noticeable how applying augmentation methods in [70], [71], and [72] resulted in higher scores of similarity (0.95-0.96) when compared to using raw images alone. For example, Abdelhafiz et al. first combined the INBreast dataset with another two datasets forming a huge set then used rotation and translation techniques to additionally generate synthetic data, which resulted in an augmented dataset that was ten times larger than the combined one. In contrast, in our work, no augmentation is performed on the existing data as it is beyond the scope of this paper which mainly focuses on evaluating the proposed convolution



(a) Original network.

(b) Modified network.

FIGURE 12. Comparison of explanation techniques employing the pixel-flipping metric. The original DeeplabV3+ Segmentation Network is demonstrated in (a) to illustrate how various explainable models perform. On the other hand, the results associated with the double-dilated network are displayed in (b).

TABLE 5. Analysis of the various approaches, including ours, to mass segmentation using the INBreast dataset. *Network* Papers are compared in terms of the architecture used for segmentation, the usage of augmentation techniques, the method used for splitting the data, the explainability of the model, and the obtained results. According to the Dice Similarity evaluation metric, all the results are reported as they were in their respective papers. (*Train-Val-Test: Train-(Validation)-Test split ratio, AM-MSP-cGAN: Attention Mechanism and Multi-Scale Pooling Conditional Generative Adversarial Network, CV: Cross Validation.*

Work	Network	Augmentation	Train-Val-Test Split	Explainability	Results
Wang et al. (2020) [69]	AM-MSP-cGAN	No	75%-25%	No	0.83
Abdelhafiz et al. (2020) [70]	Vanilla U-Net	Yes (combined x10)	83%-10%-7%	No	0.95 (mean)
Baccouche et al. (2021) [71]	Connected-UNets	Yes (x6)	70%-20%-10%	No	0.95
Alkhaleefah et al. (2022) [72]	Connected-SegNets	Yes (x8)	70%-15%-15%	No	0.96
Our Work	Double-Dilated Deeplabv3+ with Tversky Loss	No	5-Fold CV	Yes	0.86 (best) 0.81 (mean)

architecture compared to the standard dilated convolution adopted by the state-of-art segmentation networks. This huge difference in the number of training mammograms can provide one important explanation for the 10% difference in the similarity score results. However, this encourages future work to investigate the usage of augmentation methods with double-dilated convolution by utilizing the necessary computational resources that are capable of processing huge datasets.

In addition to that, the validation splits used in previous work adopted the fixed train-test split method, such as the (3:1) ratio used in [69]. This means that they randomly divided the existing mammograms into train and test subsets and kept those subsets fixed throughout the entire model evaluation process. This lack of variability can result in overestimating the model's performance as the model may perform very well on the specific training and testing subsets, but its performance may significantly degrade when applied to new, unseen data. On the other hand, we incorporate the 5-Fold validation approach in our analysis, which involve training and testing the model on different combinations of data to provide more robust performance estimates. For example, the best-achieved similarity score by our proposed segmentation network is 86%, which is obtained by using the second validation fold. However, in Table 3, we report the mean score (81%) to give a reliable estimation of the model's segmentation capability. Finally, it is also noted from the comparison that this work is the first to integrate explainability algorithms on the mass segmentation problem using the INBreast mammography data.

VI. CONCLUSION

The use of computer-aided image segmentation in cancer detection is crucial for early diagnosis and improved outcomes. This study addresses three key challenges in medical image segmentation systems. First, a novel double-dilated convolution module is proposed to address the issue of declining local resolutions in medical images that often occurs in existing CNN-based segmentation methods. This module, implemented in the Deeplabv3+ network, employs two dilation factors in parallel, replacing the traditional dilated convolution layer. The evaluation is conducted on mammogram screenings from the INBreast dataset. Second, to tackle pixel-level class imbalance in the data, four commonly used loss functions are compared to identify the most effective approach. Third, we employed XAI methods for enhancing interpretability to yield segmentations that can be easily understood, and we quantitatively assessed their performance in relation to both complexity and truthfulness.

Our experimental results demonstrate that the double-dilated convolution method has the potential to improve similarity scores and reduce the miss-detection rate in CNN-based medical imaging segmentation networks. Additionally, the Tversky loss function stands out as the most suitable option based on validation results, highlighting the importance of selecting the right loss function with properly tuned hyper-parameters, particularly for underrepresented classes. Lastly, the study shows the validity of adopting explainability techniques, particularly the Gradient-weighted Class Activation Map (Grad-CAM), to provide interpretable segmentation results. The performance of Grad-CAM shows effectiveness in explaining CAD segmentation results, providing medical professionals with reliable and understandable insights for clinical decision-making.

VII. LIMITATIONS AND FUTURE DIRECTIONS

We provide a theoretical foundation for the proposed double-dilated convolution module and compared the effectiveness of the proposed double-dilated convolution with the standard dilated convolution module by integrating double-dilated convolution in the DeepLabV3+ instead of standard convolution. As mentioned earlier, there is more than one way to implement double-dilation convolution, each having its memory footprint and runtime requirements. We leave the task of implementing these different realizations of the double-dilated convolution module as future work. In addition, there is a need for a large-scale experimental comparison of different architectures proposed in our work and previous works [58], [70], [71], [72] under the same training and test conditions to be able to identify the best-performing network for the breast cancer segmentation problem. This requires access to the implementations of different networks, a large unified dataset and high computational resources.

Moving forward, our future endeavours involve testing the application of the suggested approaches on extensive datasets encompassing diverse medical image modalities to confirm the efficacy of our segmentation and interpretability strategies. Additionally, our research path includes assessing the potential of our demonstrated methodology from this study to be extended to non-medical dense prediction and object detection tasks. Lastly, we aim to delve into the issue of elevated false positive rates in CAD systems, aiming to mitigate the adverse psychological effects on patients and minimize the need for unnecessary biopsies.

REFERENCES

[1] H.-P. Chan, L. M. Hadjiiski, and R. K. Samala, "Computer-aided diagnosis in the era of deep learning," *Med. Phys.*, vol. 47, no. 5, pp. e218–e227, 2020.

[2] Nillmani, N. Sharma, L. Saba, N. Khanna, M. Kalra, M. Fouda, and J. Suri, "Segmentation-based classification deep learning model embedded with explainable AI for COVID-19 detection in chest X-ray scans," *Diagnostics*, vol. 12, no. 9, p. 2132, Sep. 2022.

[3] A. K. Dubey et al., "Ensemble deep learning derived from transfer learning for classification of COVID-19 patients on hybrid deep-learning-based lung segmentation: A data augmentation and balancing framework," *Diagnostics*, vol. 13, no. 11, p. 1954, Jun. 2023.

[4] B. Jena, S. Saxena, G. K. Nayak, A. Balestrieri, N. Gupta, N. N. Khanna, J. R. Laird, M. K. Kalra, M. M. Fouda, L. Saba, and J. S. Suri, "Brain tumor characterization using radiogenomics in artificial intelligence framework," *Cancers*, vol. 14, no. 16, p. 4052, Aug. 2022.

[5] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: Achievements and challenges," *J. Digit. Imag.*, vol. 32, no. 4, pp. 582–596, Aug. 2019.

[6] J. S. Suri, M. Bhagawati, S. Agarwal, S. Paul, A. Pandey, S. K. Gupta, L. Saba, K. I. Paraskevas, N. N. Khanna, J. R. Laird, A. M. Johri, M. K. Kalra, M. M. Fouda, M. Fatemi, and S. Naidu, "UNet deep learning architecture for segmentation of vascular and non-vascular images: A microscopic look at UNet components buffered with pruning, explainable artificial intelligence, and bias," *IEEE Access*, vol. 11, pp. 595–645, 2023.

[7] S. Saxena, B. Jena, N. Gupta, S. Das, D. Sarmah, P. Bhattacharya, T. Nath, S. Paul, M. M. Fouda, M. Kalra, L. Saba, G. Pareek, and J. S. Suri, "Role of artificial intelligence in radiogenomics for cancers in the era of precision medicine," *Cancers*, vol. 14, no. 12, p. 2860, Jun. 2022.

[8] R. Hussein, S. Lee, R. Ward, and M. J. McKeown, "Semi-dilated convolutional neural networks for epileptic seizure prediction," *Neural Netw.*, vol. 139, pp. 212–222, Jul. 2021.

[9] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.

[10] 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.*, 2015, pp. 234–241.

[11] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[12] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.

[13] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian, "A real-time algorithm for signal analysis with the help of the wavelet transform," in *Wavelets*, 1990, pp. 286–297.

[14] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1451–1460.

[15] Y. Ho and S. Wookey, "The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling," *IEEE Access*, vol. 8, pp. 4806–4813, 2020.

[16] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[17] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3D fully convolutional deep networks," 2017, *arXiv:1706.05721*.

[18] G. Baselli, M. Codari, and F. Sardanelli, "Opening the black box of machine learning in radiology: Can the proximity of annotated cases be a way?" *Eur. Radiol. Exp.*, vol. 4, no. 1, p. 30, May 2020.

[19] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *J. Imag.*, vol. 6, no. 6, p. 52, Jun. 2020.

[20] P. Dardouillet, A. Benoit, E. Amri, P. Bolon, D. Dubucq, and A. Crédoz, "Explainability of image semantic segmentation through SHAP values," in *Proc. 26th Int. Conf. Pattern Recognit., 2nd Workshop Explainable Ethical AI (ICPR-XAIE)*, Montreal, QC, Canada, Aug. 2022, pp. 188–202.

[21] M. Abukmeil, A. Genovese, V. Piuri, F. Rundo, and F. Scotti, "Towards explainable semantic segmentation for autonomous driving systems by multi-scale variational attention," in *Proc. IEEE Int. Conf. Auton. Syst. (ICAS)*, Aug. 2021, pp. 1–5.

[22] *Canadian Cancer Statistics Advisory Committee in Collaboration With the Canadian Cancer Society, Statistics Canada and the Public Health Agency of Canada. Canadian Cancer Statistics 2021*, Can. Cancer Soc., Toronto, ON, Canada, 2021. Accessed: Jul. 6, 2023. [Online]. Available: http://cancer.ca/Canadian-Cancer-Statistics-2021-EN

[23] J. Mendes, J. Domingues, H. Aidos, N. Garcia, and N. Matela, "AI in breast cancer imaging: A survey of different applications," *J. Imag.*, vol. 8, no. 9, p. 228, Aug. 2022.

[24] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "INbreast: Toward a full-field digital mammographic database," *Acad. Radiol.*, vol. 19, no. 2, pp. 236–248, 2012.

[25] X. Yu, Q. Zhou, S. Wang, and Y. Zhang, "A systematic survey of deep learning in breast cancer," *Int. J. Intell. Syst.*, vol. 37, no. 1, pp. 152–216, Jan. 2022.

[26] 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.

[27] X. He, R. S. Zemel, and M. A. Carreira-Perpinan, "Multiscale conditional random fields for image labeling," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun./Jul. 2004, p. 2.

[28] J. Yao, S. Fidler, and R. Urtasun, "Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 702–709.

[29] L. Pei, L. Vidyaratne, M. M. Rahman, and K. M. Iftekharuddin, "Context aware deep learning for brain tumor segmentation, subtype classification, and survival prediction using radiology images," *Sci. Rep.*, vol. 10, no. 1, Nov. 2020, Art. no. 19726.

[30] E. Tappeiner, M. Welk, and R. Schubert, "Tackling the class imbalance problem of deep learning-based head and neck organ segmentation," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 17, no. 11, pp. 2103–2111, May 2022.

[31] P. F. Christ, F. Ettlinger, F. Grün, M. Ezzeldin A. Elshaera, J. Lipkova, S. Schlecht, F. Ahmaddy, S. Tatavarty, M. Bickel, P. Bilic, M. Rempfler, F. Hofmann, M. D Anastasi, S.-A. Ahmadi, G. Kaissis, J. Holch, W. Sommer, R. Braren, V. Heinemann, and B. Menze, "Automatic liver and tumor segmentation of CT and MRI volumes using cascaded fully convolutional neural networks," 2017, *arXiv:1702.05970*.

[32] M. B. Naceur, M. Akil, R. Saouli, and R. Kachouri, "Fully automatic brain tumor segmentation with deep learning-based selective attention using overlapping patches and multi-class weighted cross-entropy," *Med. Image Anal.*, vol. 63, Jul. 2020, Art. no. 101692.

[33] S. Lu, F. Gao, C. Piao, and Y. Ma, "Dynamic weighted cross entropy for semantic segmentation with extremely imbalanced data," in *Proc. Int. Conf. Artif. Intell. Adv. Manuf. (AIAM)*, Oct. 2019, pp. 230–233.

[34] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Proc. Int. Workshop Deep Learn. Med. Image Anal.*, vol. 10553, 2017, pp. 240–248.

[35] M. Tulio Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?': Explaining the predictions of any classifier," 2016, *arXiv:1602.04938*.

[36] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep Taylor decomposition," *Pattern Recognit.*, vol. 65, pp. 211–222, May 2017.

[37] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, Jul. 2015, Art. no. e0130140.

[38] V. Pitroda, M. M. Fouda, and Z. M. Fadlullah, "An explainable AI model for interpretable lung disease classification," in *Proc. IEEE Int. Conf. Internet Things Intell. Syst. (IoTaIS)*, Nov. 2021, pp. 98–103.

[39] R. R. Kontham, A. K. Kondoju, M. M. Fouda, and Z. M. Fadlullah, "An end-to-end explainable AI system for analyzing breast cancer prediction models," in *Proc. IEEE Int. Conf. Internet Things Intell. Syst. (IoTaIS)*, Nov. 2022, pp. 402–407.

[40] V. Couteaux, O. Nempont, G. Pizaine, and I. Bloch, "Towards interpretability of segmentation networks by analyzing deepdreams," in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, vol. 11797. Cham, Switzerland: Springer, 2019, pp. 56–63.

[41] P. Lakhani, "The importance of image resolution in building deep learning models for medical imaging," *Radiol., Artif. Intell.*, vol. 2, no. 1, Jan. 2020, Art. no. e190177.

[42] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.

[43] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3640–3649.

[44] Z. Wu, C. Shen, and A. van den Hengel, "Bridging category-level and instance-level semantic image segmentation," 2016, *arXiv:1605.06885*.

[45] 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," 2017, *arXiv:1606.00915*.

[46] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[47] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," 2018, *arXiv:1802.02611*.

[48] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[49] L. W. Bassett, K. Conner, and I. Ms, "The abnormal mammogram," in *Holland-Frei Cancer Medicine*, 6th ed. Hamilton, ON, Canada: BC Decker, 2003.

[50] 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.

[51] I. Good, "Rational decisions," *J. Roy. Stat. Soc. B, Methodol.*, vol. 14, pp. 107–114, Jan. 1952.

[52] Y.-D. Ma, Q. Liu, and Z.-B. Quan, "Automated image segmentation using improved PCNN model based on cross-entropy," in *Proc. Int. Symp. Intell. Multimedia, Video Speech Process.*, Oct. 2004, pp. 743–746.

[53] M. Masood, T. Nazir, M. Nawaz, A. Mehmood, J. Rashid, H.-Y. Kwon, T. Mahmood, and A. Hussain, "A novel deep learning method for recognition and classification of brain tumors from MRI images," *Diagnostics*, vol. 11, no. 5, p. 744, Apr. 2021.

[54] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020.

[55] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision—ECCV 2014*. Cham, Switzerland: Springer, 2014, pp. 818–833.

[56] J. Kauffmann, K.-R. Müller, and G. Montavon, "Towards explaining anomalies: A deep Taylor decomposition of one-class models," *Pattern Recognit.*, vol. 101, May 2020, Art. no. 107198.

[57] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, "Explaining deep neural networks and beyond: A review of methods and applications," *Proc. IEEE*, vol. 109, no. 3, pp. 247–278, Mar. 2021.

[58] Z. Wang, E. Wang, and Y. Zhu, "Image segmentation evaluation: A survey of methods," *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5637–5674, Dec. 2020.

[59] P. Jaccard, "The distribution of the flora in the alpine zone.1," *New Phytologist*, vol. 11, no. 2, pp. 37–50, Feb. 1912.

[60] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, Jul. 1945.

[61] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool," *BMC Med. Imag.*, vol. 15, no. 1, p. 29, Dec. 2015.

[62] N. Abraham and N. M. Khan, "A novel focal Tversky loss function with improved attention U-Net for lesion segmentation," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 683–687.

[63] S. Jadon, O. P. Leary, I. Pan, T. J. Harder, D. W. Wright, L. H. Merck, and D. Merck, "A comparative study of 2D image segmentation algorithms for traumatic brain lesions using CT data from the ProTECTIII multicenter clinical trial," in *Proc. Med. Imag., Imag. Informat. Healthcare, Res., Appl.*, Mar. 2020, p. 48.

[64] N. Dhungel, G. Carneiro, and A. P. Bradley, "Automated mass detection in mammograms using cascaded deep learning and random forests," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl. (DICTA)*, Nov. 2015, pp. 1–8.

[65] R. M. Nishikawa, M. Kallergi, and C. G. Orton, "Computer-aided detection, in its present form, is not an effective aid for screening mammography," *Med. Phys.*, vol. 33, no. 4, pp. 811–814, Apr. 2006.

[66] K. Loizidou, R. Elia, and C. Pitris, "Computer-aided breast cancer detection and classification in mammography: A comprehensive review," *Comput. Biol. Med.*, vol. 153, Feb. 2023, Art. no. 106554.

[67] S. Ciatto, N. Houssami, D. Bernardi, F. Caumo, M. Pellegrini, S. Brunelli, P. Tuttobene, P. Bricolo, C. Fantò, M. Valentini, S. Montemezzi, and P. Macaskill, "Integration of 3D digital mammography with tomosynthesis for population breast-cancer screening (STORM): A prospective comparison study," *Lancet Oncol.*, vol. 14, no. 7, pp. 583–589, Jun. 2013.

[68] L. Rieger, P. Chormai, G. Montavon, L. K. Hansen, and K.-R. Müller, "Structuring neural networks for more explainable predictions," in *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Cham, Switzerland: Springer, 2018, pp. 115–131.

[69] Y. Wang, S. Wang, J. Chen, and C. Wu, "Whole mammographic mass segmentation using attention mechanism and multiscale pooling adversarial network," *J. Med. Imag.*, vol. 7, no. 5, Oct. 2020, Art. no. 054503.

[70] D. Abdelhafiz, J. Bi, R. Ammar, C. Yang, and S. Nabavi, "Convolutional neural network for automated mass segmentation in mammography," *BMC Bioinf.*, vol. 21, no. 1, p. 192, Dec. 2020, doi: 10.1186/s12859-020-3521-y.

[71] A. Baccouche, B. Garcia-Zapirain, C. Castillo Olea, and A. S. Elmaghraby, "Connected-UNets: A deep learning architecture for breast mass segmentation," *npj Breast Cancer*, vol. 7, no. 1, p. 151, Dec. 2021.

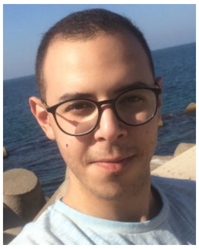
[72] M. Alkhaleefah, T.-H. Tan, C.-H. Chang, T.-C. Wang, S.-C. Ma, L. Chang, and Y.-L. Chang, "Connected-SegNets: A deep learning model for breast tumor segmentation from X-ray images," *Cancers*, vol. 14, no. 16, p. 4030, Aug. 2022.

[73] A. Larroza, F. J. Pérez-Benito, J.-C. Perez-Cortes, M. Román, M. Pollán, B. Pérez-Gómez, D. Salas-Trejo, M. Casals, and R. Llobet, "Breast dense tissue segmentation with noisy labels: A hybrid threshold-based and mask-based approach," *Diagnostics*, vol. 12, no. 8, p. 1822, Jul. 2022.

[74] N. Saffari, H. A. Rashwan, M. Abdel-Nasser, V. Kumar Singh, M. Arenas, E. Mangina, B. Herrera, and D. Puig, "Fully automated breast density segmentation and classification using deep learning," *Diagnostics*, vol. 10, no. 11, p. 988, Nov. 2020.

[75] K. Zhou, W. Li, and D. Zhao, "Deep learning-based breast region extraction of mammographic images combining pre-processing methods and semantic segmentation supported by deeplab v3+," *Technol. Health Care*, vol. 30, pp. 173–190, Feb. 2022. [Online]. Available: https://content.iospress.com/articles/technology-and-health-care/thc228017

AYA FARRAG received the B.Sc. degree in electrical engineering (communications and electronics) from Alexandria University, Egypt, in 2018. She is currently pursuing the master's degree in computer science with Lakehead University, ON, Canada. She is a Teaching Assistant with Lakehead University. Her current research interests include machine learning, health analytics, and natural language processing.



GAD GAD (Student Member, IEEE) received the bachelor's degree in computer engineering from Nile University, Egypt, in 2021, and the master's degree in computer science from Lakehead University, ON, Canada, in 2023. He is currently pursuing the Ph.D. degree in computer science with Western University, ON, Canada. His current research interests include deep learning, federated learning, and differential privacy. He is a Vector Institute AI Scholarship Recipient. He received the Best Poster Award at the NRSC 2020 Conference.

ZUBAIR MD. FADLULLAH (Senior Member, IEEE) is currently an Associate Professor with the Computer Science Department, Western University, London, ON, Canada. He was an Associate Professor with Lakehead University, ON, Canada, from 2019 to 2022. Prior to that, he was an Associate Professor with the Graduate School of Information Sciences (GSIS), Tohoku University, Japan, from 2017 to 2019, and an Assistant Professor, from 2011 to 2017. His current research interests include emerging communication systems, such as 5G new radio and beyond, deep learning applications for solving computer science and communication system problems, UAV-based systems, smart health technology, cyber security, game theory, smart grids, and emerging communication systems. He was a recipient of the Prestigious Dean's and President's Awards from Tohoku University, in March 2011; the IEEE Asia–Pacific Outstanding Researcher Award, in 2015; and the NEC Tokin Award for Research in, 2016, for his outstanding contributions. He received several best paper awards at conferences, including IEEE/ACM IWCMC, IEEE GLOBECOM, and IEEE IC-NIDC. He is currently an Editor of IEEE Transactions on Vehicular Technology, IEEE Transactions on Wireless Communications, IEEE Transactions on Green Communications and Networking, IEEE Open Journal of the Communications Society, and IEEE Access.

MOSTAFA M. FOUDA (Senior Member, IEEE) received the B.S. (as the valedictorian) and M.S. degrees in electrical engineering from Benha University, Egypt, in 2002 and 2007, respectively, and the Ph.D. degree in information sciences from Tohoku University, Japan, in 2011. He was an Assistant Professor with Tohoku University and a Postdoctoral Research Associate with Tennessee Technological University, TN, USA. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, Idaho State University, ID, USA. He is also a Full Professor with Benha University. He has (co)authored more than 210 technical publications. His current research interests include cybersecurity, communication networks, signal processing, wireless mobile communications, smart healthcare, smart grids, AI, and the IoT. He has received several research grants, including NSF Japan–U.S. Network Opportunity 3 (JUNO3). He has guest-edited several special issues covering various emerging topics in communications, networking, and health analytics. He currently serves on the editorial board of IEEE Transactions on Vehicular Technology, IEEE Internet of Things Journal, and IEEE Access.

MAAZEN ALSABAAN received the B.S. degree in electrical engineering from King Saud University (KSU), Saudi Arabia, in 2004, and the M.A.Sc. and Ph.D. degrees in electrical and computer engineering from the University of Waterloo, Canada, in 2007 and 2013, respectively. He is currently an Associate Professor with the Department of Computer Engineering, KSU. From 2015 to 2018, he was the Chairperson of the Department. He serves as a Consultant for different agencies and has been awarded many grants from KSU and King Abdulaziz City for Science and Technology (KACST). His current research interests include wireless communications and networking, surveillance systems, vehicular networks, green communications, intelligent transportation systems, and cybersecurity.