Comparison of multiple deep models on semantic segmentation for breast tumor detection

Sajid Ullah Khan[1, 2], Sharif. Muhammad Nawaz[1], Mussaab Ibrahim. Niass[1], Mehtab Afzal [2], Muhammad Shoaib[3]

[1]National Center for International Joint Research of Electronic Materials and System, School of Information Engineering, Zhengzhou University, Zhengzhou, Henan 450001, P. R. China

[2]Department of Computer Science & IT, The University of Lahore, Lahore, Pakistan

[3] Cecos University Peshawar, Pakistan

Corresponding authors:

itssukhan@gmail.com, sajid.ullah@cs.uol.edu.pk, shoaib1646@gmail.com, mehtab.afzal@cs.uol.edu.pk

***ABSTRACT**-- The early diagnosis of breast tumor detection is the most significant research issue in mammography. Computer-aided diagnosis (CAD) is one of the highly essential methods to prevent breast cancer. This research work explored the effectiveness of deep-based pixel-wise segmentation models for low energy X-rays (mammographic imagery) to detect tumors in the breast region. For this purpose, various semantic segmentation models were incorporated into the experimental procedure. All the models were analyzed using the medical images dataset, which was gathered and annotated from one of the largest teaching hospitals in the Khyber Pakhtunkhwa province, known as Lady reading hospital. It is coordinated in cooperation with local health specialists, radiologists, and technologists. The comparative analysis of the incorporated segmentation techniques' performance was observed, selecting the most appropriate model for detecting tumors and normal breast regions. The experimental evaluation of the proposed models performs efficient detection of tumor and non-tumor areas in breast mammograms using traditional evaluation metrics such as mean IoU and Pixel accuracy. The performance of the semantic segmentation techniques was evaluated on two datasets (Cityscapes and mammogram). Dilation 10 (global) performed the best among the four semantic segmentation models by achieving a higher pixel accuracy of 93.69%. It reflects the effectiveness of the pixel-wise segmentation techniques by outperforming other state-of-the-art automatic image segmentation models.*

Keywords: Semantic segmentation, Breast tumor detection, Mammography.

I. INTRODUCTION

Breast cancer is the most common cancer disease for women across the globe. According to statistical analysis, a large number of women each year are diagnosed with breast cancer. Many women die from breast cancer, and a trend of increasing cases has been observed in recent decades. It is observed that breast cancer screening has reduced mortality by around 30% [1].]. If cancer is discovered between screenings, it is called interval cancer (IC). They mostly originate when a woman discovers a lump herself. Interval cancers are more aggressive and result in higher mortality than screen-detected tumors; therefore, we must find ways to improve the screening process and identify which women are at risk. Unfortunately, increased demand for screening resources comes when the supply of qualified radiologists is low, and their duties are over-stretched.

To start with, success in developing a competent model for semantic segmentation on mammographies could help improve the process of generating and revising mammographies in hospitals. For example, it could be used for quality control, automatically detecting mammographies that need to be repeated due to low quality or artifacts. This research work helps recognize early signs of breast cancer perform semantic segmentation in mammographies. It focuses on the region of interest (tumor) and a non-region of interest (background) in the grayscale images. It discovers a reliable model to perform semantic segmentation on mammographic images after statistical analysis of the results. Success in this task could improve posterior detection and risk assessment networks' quality, ultimately increasing cancer control quality via mammographic images. The task of performing semantic segmentation on mammographic imagery is motivated by several factors.

Secondly, semantic segmentation provides spatial information that could be potentially useful for cancer risk prediction networks. Cancer risk prediction is a relatively new path of investigation. It aims at predicting the risk of developing cancer in the future, given some medical data. Thus, a model attempting to predict the risk of developing cancer from mammographies could extract useful insight. It can be achieved by providing the semantic segmentation of mammography. In addition to the mammography itself, it focuses only on some areas of interest while ignoring the areas with no relevant information (like the image's background). Furthermore, a pre-trained model on semantic segmentation could boost the performance of current tumor classification/detection approaches, as proposed in [2].

Finally, it is a research question that, for the moment, remains unexplored. Up to the best of my knowledge, this research area is less focused on the literature reporting results on this specific task. Contrarily, most of the current research related to mammographies attempts to locate tumors directly. Most of them doesn't show the comparison of extensive experimental analysis to identify the outperformance of specific segmentation technique [3] [4] [5]. However, this work's scope of this study is limited to

determining the viability of adopting such models rather than their implementation as a market solution that is optimized and practical. Furthermore, during training, testing more models was emphasized over a thorough search of the hyper-parameter space. As a result, a more thorough hyper-parameter tuning could result in a marginal improvement in the performance measures described in this paper for any given model.

The risk of cancer prediction in mammographies could provide a significant improvement in the early diagnosis of cancer. It can be achieved using semantic segmentation techniques to focus on the region of interest while ignoring the areas with no relevant information (like the image background). A pre-trained model may be utilized to diagnose current tumor classification using semantic segmentation with ground truth labeling. In [17], a deep convolutional neural network strategy is implemented to detect micro-calcification in mammograms obtained from three different manufacturers. The two CNNs were trained on the training set to detect the micro-classification candidates. The deep CNN was compared to the state-of-the-art cascade classifier, where the CNN outperforms the cascade classifier. Xiaobo Lai et al. [18] presented the automatic segmentation method using a U-Net architecture to work on digital breast tomosynthesis (DBT) images. It achieved the automatic segmentation accuracy of breast masses. In [3, 24], a multi-scale convolutional neural network (CNN) is proposed and trained for mammogram classification. It is based on feature learning with a curriculum learning strategy to provide a labeled mammogram image as an outcome to facilitate early diagnosis. According to Deszo et al. [4], A CAD system is proposed, which outperforms the INbreast database classification. It is capable to automatically detect benign and malignant lesions in mammography on the object detection frameworks, Faster R-CNN. Li Shen [5] introduced an end-to-end training algorithm using a convolutional design that outperformed the previous methods. It is utilized for mammogram images to detect early breast cancer based on lesion annotations for training. Image level labels are applied for image classification. A DDSM is used as a benchmark to prove its good performance.

In recent years, a few of the approaches utilizing semantic segmentation for specific tasks or subtasks, such as segmenting region-of-interest (ROI) [17] and mass segmentation [19, 20]. Some of the deep learning techniques in Mammography are highlighted in [21, 22, 23]. The other attempts show the high significance of designing approaches for detecting cancer status based on mammogram images to classify for possible cancer detection**.**

Table 1 highlights the state-of-the-art literature about the common deep learning models, benchmarks, and their noted accuracy levels to highlight their performances. It is significant because of the relevance to the work in this paper.

Table 1. Common deep learning benchmarks

Reference	**Dataset**	**Model**	**Accuracy**	**mIoU**
(Cordts et al., 2016)[6]	Cityscape 2016	FCN32, FCN16, FCN8	67.1 %	77.9%
(Long et al., 2015)[7]	PASCAL VOC 2012	FCN32, FCN16 FCN8	94.3%	62.2%
(Ronneberger et al.,	ISBI 2015	U-Net (2015)	92%	77.5%
(Zhou et al., 2018)[9]	Liver Dataset	U-Net ++	90.4%	82.90%
(Yu & Koltun,	VOC 2012	Dilation 10	71.3%	69.8%
(Hamaguchi, 2017)[11]	Cityscapes	LFE+ Dilation	63.6%	50%
(Chen et al., 2017)[12]	Cityscapes 2015	DeepLabv3	79.30 %	81.3%
(Simon et al., 2017)[13]	CamVid 2015, Gatech	FCN-8 Dense-Net	91.5%	66.9%
(Nedra et al., 2018)[14]	Drive 2012	FCN32	91%	64%
(Lin et al., 2016)[15]	PASCAL VOC	FCN	71.5%	53.9%
(Badrinarayanan et al., 2015)[16]	Cam-Vid road scenes	SegNet, FCN	90.40 %	60.10%

II. PROPOSED SEGMENTATION MODEL

Many different neural network architectures have been proposed in recent years to tackle semantic segmentation within the scope of deep learning. It is intended to use some of the most relevant ones to be trained in segmenting mammographic imagery. In this work, semantic segmentation is performed to detect the tumor and non-tumor regions in breast mammograms. The positive feature of the Cityscape benchmark is that it can perform semantically and object segmentation as well. It returns pixels with class labeling, and the objects (anatomical regions) are separately segmented. At first, the Cityscapes dataset is used to train the model. It is used for the semantic segmentation of the breast Mammogram dataset based on the ground truth annotations obtained from radiologists and clinical experts. The detail of the architecture of the models that have been considered are following;

2.1 Fully Convolutional Network

The Fully Convolutional Network (FCN) is a semantic segmentation model used to execute any images without using fully connected layers. At first, an encoder is used to obtain contextual information about an image. The encoder architecture with alternating sequences of 2 or 3 convolutional layers with max-pooling layers is deployed. It is almost similar to VGG16 [7], except in the original VGG16 architecture, the convolutional layers are used instead of fully connected layers. Each of the convolutional operations is followed by the activation function as a rectified linear unit (ReLU). The details of the encoder architecture are described in table 2. Each layer of the network includes convolution strides, feature map dimensions, kernel sizes, and receptive fields. T input and output layers' sizes are based on calculations concerning the input patches of 256 x 256 pixels. It is the size initially used for FCN without any change.

Table 2. FCN encoder architecture [7]

Name:	Layer	Kernel Size	Stride	Padding	Input Size	Output Size	Input Feature Maps	Output Feature Maps	Receptive Field
conv1_1	1	3	1	SAME	256	256	3	64	3
conv1_2	2	3	1	SAME	256	256	64	64	5
max_pool1	3	2	2	SAME	256	128	64	64	6
conv2_1	4	3	1	SAME	128	128	64	128	10
conv2_2	5	3	1	SAME	128	128	128	128	14
max_pool2	6	2	2	SAME	128	64	128	128	16
conv3_1	7	3	1	SAME	64	64	128	256	24
conv3_2	8	3	1	SAME	64	64	256	256	32
conv3_3	9	3	1	SAME	64	64	256	256	40
max_pool3	10	2	2	SAME	64	32	256	256	44
conv4_1	11	3	1	SAME	32	32	256	512	60
conv4_2	12	3	1	SAME	32	32	512	512	76
conv4_3	13	3	1	SAME	32	32	512	512	92
max_pool4	14	2	2	SAME	32	16	512	512	100
conv5_1	15	3	1	SAME	16	16	512	512	132
conv5_2	16	3	1	SAME	16	16	512	512	164
conv5_3	17	3	1	SAME	16	16	512	512	196
max_pool5	18	2	2	SAME	16	8	512	512	212
conv6_1	19	7	1	SAME	8	8	512	4096	292
conv6_2	20	1	1	SAME	8	8	4096	4096	292
scores	21	1	1	SAME	8	8	4096	num_classes	292

After passing through the encoder, which includes applying five max-pooling layers, the feature maps' spatial dimensions are 32 times smaller than the input image patch. Since semantic segmentation requires generating predictions with the same spatial dimensions as the input, it is necessary to up-sample the encoded logits somehow. The solution proposed by Long et al. [7] involves defining three different sub-models of FCN. These sub-models are trained sequentially, using the weights obtained from training the previous model as initial weights for the following one. The first sub-model, called FCN32, takes the feature maps of size 8 x 8 obtained after *max_pool*5 and applies a bilinear up-sampling step to resize them back to 256 x 256 pixels (the input size). The second sub-model, referred to in the original paper as FCN16, uses a transposed convolutional layer to learn the best strategy to upsample *max_pool*5 outputs from size 8 x 8 to 16 x 16. The 16x16 upsampled version of *max_pool*5 and the 16 x 16 output of *max_pool*4 are summed up and finally resized to 256 x 256 pixels with a bilinear upsampling layer. The last sub-model on the decoding strategy is called FCN8 and goes a step further than FCN16. The result of summing the 16 x 16 version of *max_pool*5 to the 16 x 16 output of *max_pool*4 is up-sampled again utilizing another transposed convolution to size 32x32. This 32 x 32 feature map is summed up with the 32x32 output of *max_pool*3 and finally up-sampled to 256 x 256 pixels with a bilinear up-sampling layer first 2 cases. Transposed convolutions are followed by ReLU activation functions, just like any other convolution on the model. The exact details of the decoder are shown in table 3.

Table 3. FCN decoder architecture

Name	Layer	Kernel Size	Stride	Padding	Input Size	Output Size	Input Feature Maps	Output Feature Maps	Receptive Fields
t_conv_1	1	4	2	SAME	8	16	Num Classes	Num Classes	-
t_conv_2	1	4	2	SAME	16	32	Num Classes	Num Classes	-

During training, FCN32 will be initialized using VGG16 ImageNet pre-trained weights. Following FCN16 will be trained using FCN32 weights as initial weights. Finally, FCN8 will be initialized with FCN16 weights, trained, and used for inference.

2.2 U-Net

The U-Net [8] is a Semantic segmentation model that follows an encoder-decoder architecture. The name U-Net comes from the fact that the encoder and the decoder are symmetric. It is helpful to identify the region of interest. The encoding path, which aims to capture the image's context information, resembles most image classification models, alternating between convolutions and max-pooling operations. The encoder's specific details are defined in table 4, where input and output layer sizes are calculated using 572 x 572 pixels. The decoding path tries to retrieve localization information lost during pooling operations. It follows an architecture symmetric to the encoding path but replacing max-pooling steps with transposed convolutions. Besides, at several points along the decoding path, the feature maps are concatenated to the ones coming from the same stage at the encoder before

proceeding.

Table 4. U-Net encoder architecture

Name:	Layer	Kernel Size	Stride	Padding	Input Size	Output Size	Input Feature Maps	Output Feature Maps	Receptive Field
convl_1	1	3	1	VALID	572	570	3	64	3
convl_2	2	3	1	VALID	570	568	64	64	5
max_pool1	3	2	2	VALID	568	284	64	64	5
conv2_1	4	3	1	VALID	284	282	64	128	10
conv2_2	5	3	1	VALID	282	64	128	128	14
max_pool2	6	2	2	VALID	280	140	128	128	16
conv3_1	7	3	1	VALID	140	138	128	256	24
conv3_2	8	3	1	VALID	138	136	256	256	32
max_pool3	9	2	2	VALID	136	68	256	512	36
conv4_1	10	3	1	VALID	68	66	512	512	52
conv4_2	11	3	1	VALID	66	64	512	512	68
max_pool4	12	2	2	VALID	64	32	512	1024	76
conv5_1	13	3	1	VALID	32	30	1024	1024	108
Conv5_2	14	3	1	VALID	30	28	1024	1024	140

It is important to note that U-net convolutions are not padded. This is why the decoder's output size is 388 x 388, corresponding to the 388x388 central patch of the input image. Furthermore, each convolution or transposed convolution along the network is followed by a rectified linear unit (ReLU) activation function.

2.3 Dilation 10

Unlike the preceding models, the dilation10 [10] model does not use an encoder-decoder architecture. It can be separated into two distinct components, however: the front-end module and the context module. Like an encoder, the front-end module is designed to extract the "what" information from the images. Encoder architectures, like image classifiers, typically alternated convolutional layers with max-pooling operations to provide a field of view that covered the entire image or a substantial chunk of it—however, the spatial dimensions after the design are much less than the input size. The Dilation10 model eliminates a number of the maximum pooling operations. For compensation, the dilation of posterior convolutions is doubled by two after each eliminated max-pooling. This method maintains the same field of view as traditional encoders while decreasing the amount of detailed information lost during pooling processes. Table 5 shows the dilation10 front module, which is based on VGG16. The crop sizes utilized in the various training phases were used to calculate the input and output feature map sizes. For the first, second, and third training stages, the crop sizes are 632 × 632 pixels, 1024 x 1400 pixels, and 1400 x 1400 pixels, respectively.

Because spatial dimensions are still very big after the front module due to the suppression of some pooling steps, an alternative technique to an up-sampling decoder can be used. Dilation10, in particular, employs a module that employs a series of dilated convolutions with increasing dilation factors. This design allows for the systematic aggregation of multi-scale contextual data without sacrificing resolution. Table 6 shows the architecture of the context module in more detail. The sizes of the input and output feature maps were calculated based on the size obtained at the front module's end. As previously stated, the dilation10 network is trained in stages. Network training, in particular, is divided into three stages. Only the front module is trained in the first step. The entire network (front module + context module) is formed in the second stage of the training process. The weights in the front section, on the other hand, are frozen, and only the ones in the context module are modified. The training process concludes with a final stage in which the entire network is trained without any frozen variables. A rectified linear unit (ReLU) activation function follows all convolutions in the network.

Table 5. Dilation10 front module architecture

Name:	Layer	Kernel Size	Stride	Padding	Dilation	Input Size	Output Size	Input Feature Maps	Output Feature Maps	Receptive Field
conv1_1	1	3	1	SAME	1	632/1024/1400	632/1024/1400	3	64	3
conv1_2	2	3	1	SAME	1	632/1024/1400	632/1024/1400	64	64	5
max_pool1	3	2	2	SAME	1	632/1024/1400	316/512/700	64	64	6
conv2_1	4	3	1	SAME	1	316/512/700	316/512/700	64	128	10
conv2_2	5	3	1	SAME	1	316/512/700	316/512/700	128	128	14
max_pool2	6	2	2	SAME	1	316/512/700	158/256/350	128	128	16
conv3_1	7	3	1	SAME	1	158/256/350	158/256/350	128	256	24
conv3_2	8	3	1	SAME	1	158/256/350	158/256/350	256	256	32
conv3_3	9	3	1	SAME	1	158/256/350	158/256/350	256	256	40
max_pool3	10	2	2	SAME	1	158/256/350	79/128/175	256	256	44
conv4_1	11	3	1	SAME	1	79/128/175	79/128/175	256	512	60
conv4_2	12	3	1	SAME	1	79/128/175	79/128/175	512	512	76
conv4_3	13	3	1	SAME	1	79/128/175	79/128/175	512	512	92

dil_conv5_1	14	3	1	SAME	2	79/128/175	79/128/175	512	512	124
dil_conv5_2	15	3	1	SAME	2	79/128/175	79/128/175	512	512	156
dil_conv5_3	16	3	1	SAME	2	79/128/175	79/128/175	512	512	188
dil_conv6_1	17	7	1	SAME	4	79/128/175	79/128/175	512	4096	380
conv6_2	18	1	1	SAME	1	79/128/175	79/128/175	4096	4096	380
scores	19	1	1	SAME	1	79/128/175	79/128/175	4096	num_classes	380

Table 6. Dilation10 context module architecture

Name:	Layer	Kernel Size	Stride	Padding	Dilation	Input Size	Output Size	Input Feature Maps	Output Feature Maps	Receptive Field
context1	20	3	1	SAME	1	79/128/175	79/128/175	num_classes	num_classes	3
context2	21	3	1	SAME	1	79/128/175	79/128/175	num_classes	num_classes	5
context3	22	3	1	SAME	2	79/128/175	79/128/175	num_classes	num_classes	9
context4	23	3	1	SAME	4	79/128/175	79/128/175	num_classes	num_classes	17
context5	24	3	1	SAME	8	79/128/175	79/128/175	num_classes	num_classes	33
context6	25	3	1	SAME	16	79/128/175	79/128/175	num_classes	num_classes	65
context7	26	3	1	SAME	16	79/128/175	79/128/175	num_classes	num_classes	97
context8	27	3	1	SAME	32	79/128/175	79/128/175	num_classes	num_classes	161
context9	28	3	1	SAME	1	79/128/175	79/128/175	num_classes	num_classes	161
context10	29	3	1	SAME	1	79/128/175	79/128/175	num_classes	num_classes	161

2.4 Deep Lab v3

The Deep Lab v3 semantic segmentation system [12] is an approach that, just like dilation10, moves away from the Encoder-Decoder architecture by limiting the number of downsampling operations used encoding path. Similar to dilation10, it achieves a field of view comparable to models with a higher amount of max-pooling layers using atrous convolutions with increasing dilation factors. The feature extraction path is also based on a reference image classification model pre-trained on the ImageNet dataset. However, instead of using VGG like the previously described models, the authors of deep lab v3 use a ResNet-based architecture. In [12], the author experiments with residual networks with a different number of layers. In this research work, the network implemented is the one with 18 layers, the architecture of which can be found in figure 1.

Layer Name	Output Size	18-Layer	34-Layer	50-Layer	101-Layer	152-Layer
conv1_x	112 x 112	7 x 7, 64, stride 2				
conv2_x	56 x 56	3 x 3 max pool, stride 2				
		[3 x 3, 3 x 3,	[3 x 3, 3 x 3,	[1 x 1, 3 x 3, 1 x 1,	[1 x 1, 3 x 3, 1 x 1,	[1 x 1, 3 x 3, 1 x 1,
conv3_x	28 x 28	[3 x 3, 3 x 3,	[3 x 3, 3 x 3,	[1 x 1, 3 x 3, 1 x 1,	[1 x 1, 3 x 3, 1 x 1,	[1 x 1, 3 x 3, 1 x 1,
conv4_x	14 x 14	[3 x 3, 3 x 3,	[3 x 3, 3 x 3,	[1 x 1, 3 x 3, 1 x 1,	[1 x 1, 3 x 3, 1 x 1,	[1 x 1, 3 x 3, 1 x 1,
conv5_x	7 x 7	[3 x 3, 3 x 3,	[3 x 3, 3 x 3,	[1 x 1, 3 x 3, 1 x 1,	[1 x 1, 3 x 3, 1 x 1,	[1x1, 3x3, 1x1,
	1 x 1	Average pool, 1000-d fc, SoftMax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Fg. 1. ResNet architecture [12]

The only difference to the original ResNet implementation is that the convolutions found in the last two blocks (block3 and block4) are replaced by atrous convolution with dilation rates 2 and 4, respectively. With these changes, the output size is only eight times smaller than the input size, and the output is 16 times smaller, and therefore only the final block contains dilated convolutions.

One of the main differences for dilation10 is the module's architecture added at the end of the encoding path. This model employs spatial pyramid pooling to capture context information. With some similarities with the inception module [11], this module combines several parallel atrous convolution layers with different rates to capture multi-scale information. Figure 1 describes the architectural details of the spatial pyramid module. The input size used by the paper authors and also for the experiments in this research work is 769 x 769 pixels. Thus, input and output sizes in figure 1 are computed for 96, the feature map size at the beginning of the

spatial pyramid pooling module.
A rectified linear unit (ReLU) activation function follows all convolutions in the network. Finally, the original model includes batch normalization layers [12] along with the network. However, due to memory constraints, the mini-batch used in this work during training was of size 2. Therefore, batch normalization layers were replaced by group normalization layers [25], whose performance is not affected by the small batch size.

III. EXPERIMENT

The experiment is performed in this work on the two datasets, which are the Cityscapes (benchmark) and a breast mammogram dataset. The Cityscape benchmark is chosen for the experiment because it can achieve class labeled pixels for semantic segmentation of the anatomical regions separately and object segmentation. The Lady reading hospital is one of the largest hospitals in the province in Pakistan. A large corpus of mammography data access was granted to train the models obtained from the local population, recorded between 2011 and 2019. However, none of the screenings had pixel-wise annotations. Consequently, the annotations were created under authorized expert radiologists, technologists, and other clinical experts.

3.1 Dataset

3.1.1 Cityscapes Dataset

A prevalent semantic segmentation benchmark called Cityscapes [6] dataset was used to train and test the model. The images are about the same size as those of the Mammography dataset in the Cityscapes dataset. Moreover, papers on semantic segmentation in the literature also report the performance of the Cityscape dataset, which has proved very helpful in validating the model's implementation.

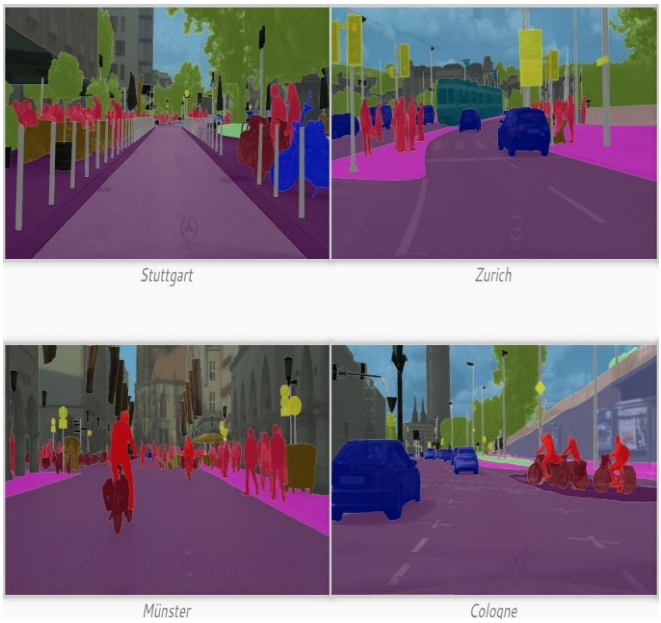

Fig. 2. Cityscapes sample images. [6]

According to the authors, classes were selected based on their frequency, relevance from an application point of view, practical considerations regarding the effort to annotate, and compatibility with existing datasets.

3.1.2 Mammography dataset

The Mammography dataset is a medical dataset composed of 100 subjects (high-resolution grayscale images) of mammography screenings. Each subject is an image and is replicated ten times using different data augmentation techniques. So the total image amount is 1000. Annotations were generated by clinical experts of the local medical teaching institute as binary pixel-wise maps for different anatomical regions. It was mainly categorized into two regions the region of interest "tumor and non-tumor region. QuPath, a software application specifically designed for bioimage analysis, was used for the task [26].
The ground truth, labeling, and segmented comparisons are made based on two classes' tumor and non-tumor regions. The detailed descriptions of all the anatomical areas of mammography are shown in table 7;

Table 7. Classes and anatomical regions in

	Tumor region	Non-Tumor region
Breast Anatomical Regions	Nipple	
	Areola	
	Calcifications	
	Skin	
	Thick vessels	
	Pectoral muscle	
	Auxiliary lymph nodes	
	Calcified vessels	
	Text	
	Submammory tissue	
	Foreign object	
	Mammary gland	
	Background	

The ranking above indicates the different anatomical regions in the breast mammogram. In extension to this work, for the semantic segmentation of all the anatomical regions, the pixels should superpose other pixels when merging the binary maps into single pixel-wise annotations. All binary maps except calcifications, text, nipple and auxiliary lymph nodes can be smoothed using a Gaussian filter at merging time. It is usually performed to avoid sudden transitions in the annotations.
One of the main challenges with this dataset is that it is very imbalanced. Some anatomical regions, such as the background or mammary gland, have a much higher presence than others. It is necessary to create a dataset composed of crops obtained from the original images sampled to mitigate the imbalance problem to prevent model training hurdles. Several crop datasets with different crop sizes may be generated from the original high-resolution images to train different models that require different input sizes. Crop sizes used to generate the datasets are 256 x 256, 700 x 700, 900 x 900 and 1500 x 1500 pixels.
Thus, to create a (more) balanced dataset containing crops of a given size x, take each image from the original dataset, and a list of the generated unique labels. Then, categories mammary gland and background are to be removed from the unique label list because one of the two will always be present in any

crop generated. From the remaining labels, one is to be selected randomly using a uniform distribution. A pixel of the selected category is to choose randomly with a uniform probability to use as the new crop center.

3.2 Data Augmentation

The augmentation is achieved by applying image processing practices, e.g., rotation, smoothing, mirror effect, sharpening, noise addition, contrast enhancement, etc. The total augmented dataset consists of 1000 images. The replicated samples are displayed in figure 4 below.

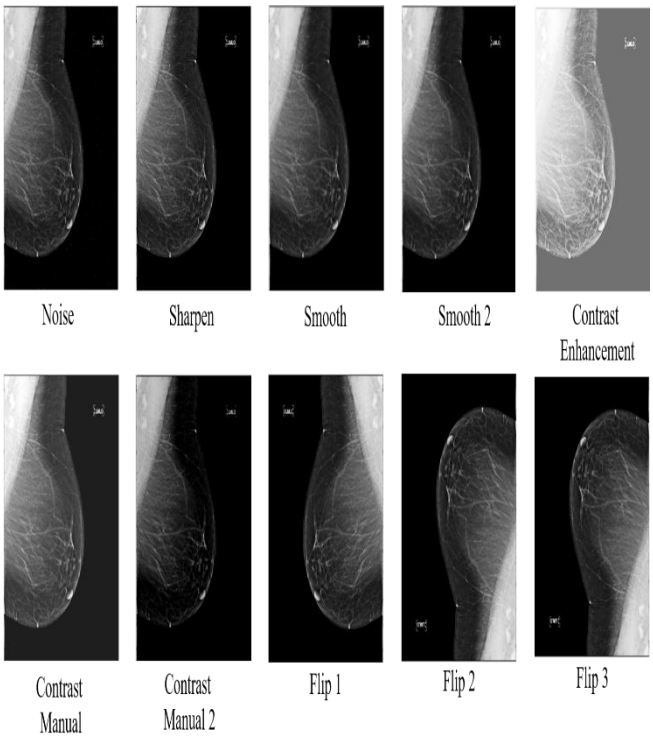

Fig. 4. Image Augmentation replicated samples

3.3 Input Pipeline

For all the experiments performed in this work, a unified input pipeline was employed for each of the datasets. For the Cityscapes dataset, which contains RGB images, a random horizontal flipping operation was first applied with a probability of 0.5. Following, with probabilities of 0.5, random distortions of the saturation, the brightness, and the contrast of the images were employed. Finally, a random crop of the desired size (depending on the network to be trained) was generated and normalized to have values between 0.5 and 0.5. For the Mammography dataset, the crop dataset of sufficient size to cover the input size of the network was trained with the crop dataset of sufficient size to cover the network's input size. The network was trained with a crop dataset of adequate size to cover the network's input size. The grayscale images were transformed into RGB images with a Tensorflow function. Data augmentation was limited in this case to random horizontal flipping and random contrast distortion; both applied with probabilities of 0.5.

3.4 Training Procedure

Stochastic Gradient Descent (SGD) optimizer with momentum has been used to reduce the cross-entropy loss function. The loss function is used to measure each label's contribution, which is the weighted value in the percentage of a class label from a dataset. More specifically, if a ' 'label's class supposes x% of the labels in the dataset, its contribution to the loss function will be weighted by $w = (1 — x/100)$. Also, early stopping have been used during training, with patience factor number of epoch without improvement before stopping) ranging from 20 to 50 depending on the time needed to train an epoch. Moreover, parameter tuning is adapted wherever required to reduce memory size, computational time, and maximum accuracy.

The data split is beneficial, depending on the kind of classifier used. In our case, the image data are cross-validated using the hold out technique with a percentage of 70/30, where 70 % of the data is selected for model training randomly. While the remaining 30, which will be by default randomized used for model testing.

3.5 Hyper-Parameters and options for models training

The learning rate is adjusted to learn quickly, adapting to a high initial rate. It follows a schedule piecewise and reduces by a factor of 3 at every ten epochs. It gives a solution nearer to the local optimum with a dropout of learning rate. The validation data parameter is set to test the network's validation data at every epoch, and the 'ValidationPatience' is set as 4 for the early stop of data training with the convergence of validation accuracy to avoid training dataset overfitting issues. Batch size is kept as a mini with the specific value depending on the parameter tuning to minimize memory use during the training phase. The batch size depends on the power capacity of the available system.

Table 8. Experimental Parameters for Dilation 10

Option	Parameters
Optimizer	SGDM
Learn Rate Schedule	Piecewise
Learn Rate Drop Period	10
Learn Rate Drop Factor	0.3
Momentum	0.9
Initial Learn Rate	0.0001
L2 Regularization	0.005
Validation Data	Yes
Max Epochs	100
Mini Batch Size	64
Shuffle	Every-epoch
Verbose Frequency	10
Validation Patience	4

IV. RESULTS EVALUATION

4.1 Cityscapes results

All the intentional models have been implemented from scratch, including the current model. Thus, it is an important step to check the validity against a standard benchmark dataset. The following table shows the results obtained on the Cityscapes dataset.

Table 9. Results on the validation dataset

Model	Pixel accuracy	Mean IoU accuracy	Mean per class accuracy
CN32	91.84	57.35	66.59
FCN16	92.88	59.25	67.86
FCN8	92.57	59.01	67.96

U-Net	91.81	58.87	68.86
Dilation10 - Front	92.98	62.94	72.78
Dilation10 – Context	95.37	63.29	75.16
Dilation10 – Global	96.09	67.87	79.05
DeepLab v3 + ResNet	93.97	63.75	74.53

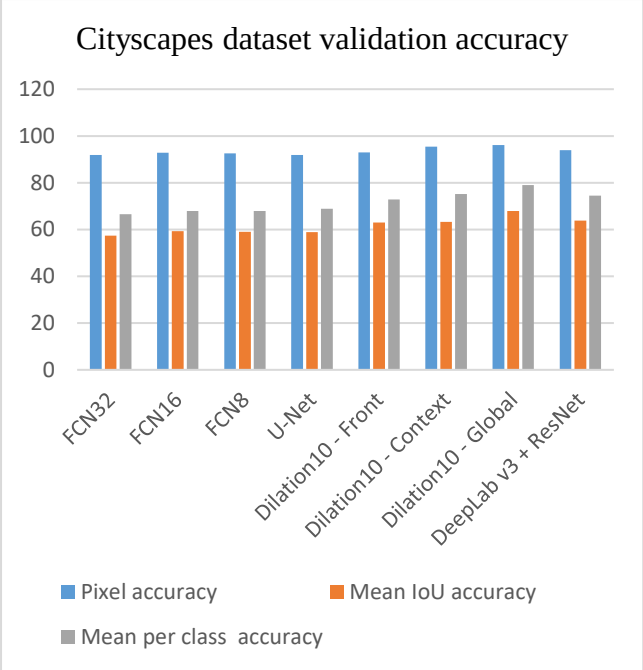

Fig. 5. Cityscapes Dataset Validation Accuracy.

4.2 Breast Mammogram results

After the validity check for implementation, the next step is to implement it against the Mammography dataset. The results of the validation dataset are reported below in table 10 and the column charts.

Table 10. Results on the validation dataset

Model	**Pixel accuracy**	**Mean IoU accuracy**	**Mean per class accuracy**
FCN32	80.37	40.06	49.20
FCN16	78.90	41.75	53.13
FCN8	78.17	45.12	57.03
U-Net	80.49	43.47	53.78
Dilation10 - Front	93.03	47.65	60.74
Dilation10 - Context	93.46	50.26	64.39
Dilation10 - Global	93.69	57.91	67.97
DeepLab v3 + ResNet	92.17	51.34	62.01

The proposed model is first trained on the Cityscape dataset and is then applied to the mammogram. The ground truth is added from the clinical experts to generate semantic annotations of breast mammograms. Various classes have been defined for mammography labeling and classification.

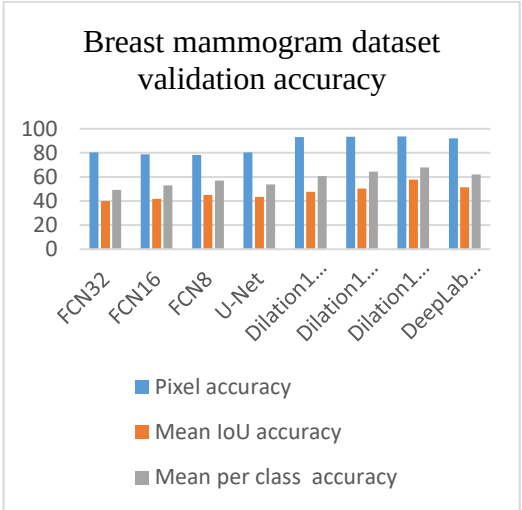

Fig. 6. Breast Mammogram Dataset Validation Accuracy

4.3 Visual Results

On a newly proposed medical dataset of mammography screenings, the performance of state-of-the-art semantic segmentation deep learning models is examined. All the reference models such as FCN with three variants (FCN 32, FCN 16, and FCN 8), U-Net, Deep Lab v3, Dilation 10 (context, front, global) are re-implemented and validated first on the benchmark dataset Cityscapes. The new medical image corpus for breast mammograms was collected and annotated to show that it is possible to boost segmentation performance by training the models in the classical training framework. The details of the visual results for the semantic segmentation techniques are shown in Figure 7 below shows that image (a) is the original image, consisting of a tumor region. Image (b) is an annotated image. The annotation/ground truth labeling is performed with radiologist collaboration. Image (c) is the binary mask of the ground truth labeled region. The image (d) consists of a segmented mask acquired by applying the dilation 10 (global) semantic segmentation model.

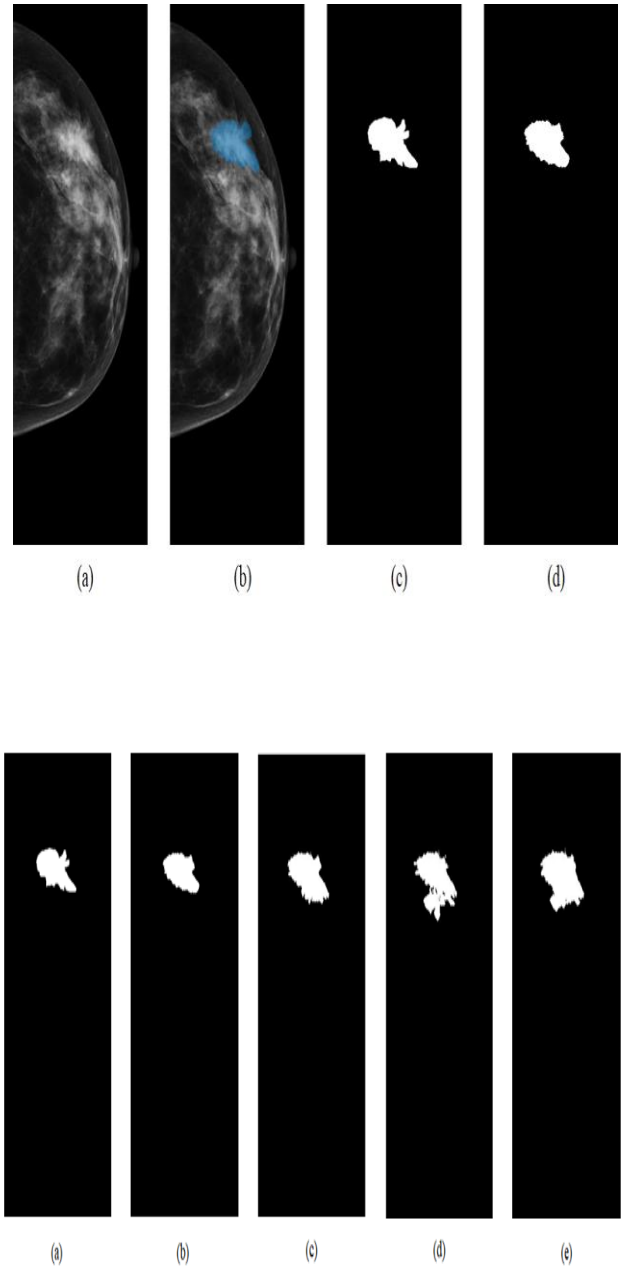

Fig. 8. (a) Ground Truth Mask, (b) Dilation Segmented Mask, (c) DeepLab-v3 Segmented Mask, (d) FCN Segmented mask, and (e) U-Net Segmented Mask.

In the above figure 8, the image (a) is the ground truth masked by the clinical expert, the image (b) is obtained using Dilation 10 (global) segmentation where it can be observed that the tumor is little under-segmented, the image (c) is the segmented image of DeepLab-v3 which seems to be bit over-segmented, the image (d) is achieved by applying FCN, and in the last image (e) the segmentation is performed using U-Net semantic segmentation model. The (d) and (e) are too much over-segmented, so from the image above, it can be analyzed that Dilation 10 (global) and DeepLab-v3 performed the best in tumor regions segmentation.

V. CONCLUSION

The two annotated datasets, such as Cityscapes a benchmark and a local medical imaging dataset Breast mammogram, train semantic segmentation algorithms. The FCN semantic segmentation model with its three variants FCN 32, FCN 16, FCN 8, U-Net, Dilation 10 (front, context, global) and DeepLab-v3 (network-based on ResNet) are implemented. A competent segmentation model for mammographies is highlighted after the detailed experimental analysis. The Dilation 10 (global) outperform compared to other segmentation models with a higher pixel accuracy of 93.69 %. In this work, the goal is limited to recognize the region of interest (tumor) and a non-region of interest (background) in the grayscale images. It may be extended in the future to segment the other anatomical regions associated with breast mammograms accurately.

REFERENCES

[1] Wong, K., Syeda-Mahmood, T., & Moradi, M. (2018). Building medical image classifiers with very limited data using segmentation networks. *Medical image analysis*, *49*, 105–116. https://doi.org/10.1016/j.media.2018.07.010

[2] Wong, K., Syeda-Mahmood, T., & Moradi, M. (2018). Building medical image classifiers with very limited data using segmentation networks. *Medical image analysis*, *49*, 105–116. https://doi.org/10.1016/j.media.2018.07.010

[3] Lotter, W., Sorensen, G., & Cox, D. (2017). A multi-scale CNN and curriculum learning strategy for mammogram classification. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (pp. 169-177). Springer, Cham.

[4] Ribli, D., Horváth, A., Unger, Z., Pollner, P., & Csabai, I. (2018). Detecting and classifying lesions in mammograms with Deep Learning. *Scientific reports*, *8*(1), 4165. https://doi.org/10.1038/s41598-018-22437-z.

[5] Shen, L. (2017). End-to-end training for whole image breast cancer diagnosis using an all convolutional design. *arXiv preprint arXiv:1711.05775*.

[6] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3213-3223).

[7] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431-3440).

[8] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Springer, Cham.

[9] Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, *39*(12), 2481-2495.

[10] Yu, F., & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

[11] Hamaguchi, R., Fujita, A., Nemoto, K., Imaizumi, T., & Hikosaka, S. (2018, March). Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery. In *2018 IEEE winter conference on applications of computer vision (WACV)* (pp. 1442-1450). IEEE.

[12] Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

[13] Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., &

Bengio, Y. (2017). The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 11-19).

[14] Nedra, A., Shoaib, M., & Gattoufi, S. (2018, March). Detection and classification of the breast abnormalities in Digital Mammograms via Linear Support Vector Machine. In *2018 IEEE 4th Middle East Conference on Biomedical Engineering (MECBME)* (pp. 141-146). IEEE.

[15] Lin, G., Shen, C., Van Den Hengel, A., & Reid, I. (2016). Efficient piecewise training of deep structured models for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3194-3203).

[16] Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, *39*(12), 2481-2495.

[17] Mordang, J. J., Janssen, T., Bria, A., Kooi, T., Gubern-Mérida, A., & Karssemeijer, N. (2016, June). Automatic microcalcification detection in multi-vendor mammography using convolutional neural networks. In International Workshop on Breast Imaging (pp. 35-42). Springer, Cham.

[18] Lai, X., Yang, W., & Li, R. (2020). DBT Masses Automatic Segmentation Using U-Net Neural Networks. *Computational and Mathematical Methods in Medicine*, *2020*.

[19] Boot, T., & Irshad, H. (2020, October). Diagnostic Assessment of Deep Learning Algorithms for Detection and Segmentation of Lesion in Mammographic Images. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 56-65). Springer, Cham.

[20] Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., McBride, R., & Sieh, W. (2019). Deep Learning to Improve Breast Cancer Detection on Screening Mammography. Scientific reports, 9(1), 12495. https://doi.org/10.1038/s41598-019-48995-4.

[21] Arevalo, J., González, F. A., Ramos-Pollán, R., Oliveira, J. L., & Guevara Lopez, M. A. (2016). Representation learning for mammography mass lesion classification with convolutional neural networks. Computer methods and programs in biomedicine, 127, 248–257. https://doi.org/10.1016/j.cmpb.2015.12.014.

[22] Yi, D., Sawyer, R. L., Cohn III, D., Dunnmon, J., Lam, C., Xiao, X., & Rubin, D. (2017). Optimizing and visualizing deep learning for benign/malignant classification in breast tumors. arXiv preprint arXiv:1705.06362.

[23] Geras, K. J., Wolfson, S., Shen, Y., Wu, N., Kim, S., Kim, E., & Cho, K. (2017). High-resolution breast cancer screening with multi-view deep convolutional neural networks. arXiv preprint arXiv:1703.07047.

[24] Agarwal, R., Diaz, O., Lladó, X., & Martí, R. (2018, July). Mass detection in mammograms using pre-trained deep learning models. In 14th International Workshop on Breast Imaging (IWBI 2018) (Vol. 10718, p. 107181F). International Society for Optics and Photonics.

[25] Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.

[26] Bankhead, P., Loughrey, M.B., Fernández, J.A. et al. QuPath: Open source software for digital pathology image analysis. Sci Rep 7, 16878 (2017). https://doi.org/10.1038/s41598-017-17204-5