11th International Conference on Electrical, Electronics and Computer Engineering (IcETRAN), Nis, 3-6 June 2024

SegFormer Model in Mammography Lesion Segmentation: A Study on the Impact of GLAM Saliency Maps

Jovana Kljajić
Faculty of Technical Sciences
University of Novi Sad
Novi Sad, Serbia
jovanakljajic18@uns.ac.rs
ORCID: 0000-0003-2360-9743

Nataša Đukić
Faculty of Technical Sciences
University of Novi Sad
Novi Sad, Serbia
djukic.e163.2023@uns.ac.rs
ORCID: 0009-0000-6834-6858

Ivan Lazić
Faculty of Technical Sciences
University of Novi Sad
Novi Sad, Serbia
ivan.lazic@uns.ac.rs
ORCID: 0000-0001-8613-0049

Tatjana Lončar-Turukalo
Faculty of Technical Sciences
University of Novi Sad
Novi Sad, Serbia
turukalo@uns.ac.rs
ORCID: 0000-0002-3582-8073

Jasmina Boban
Faculty of Medicine
University of Novi Sad
Novi Sad, Serbia
jasmina.boban@mf.uns.ac.rs
ORCID: 0000-0001-9701-5484

Milan Rapaić
Faculty of Technical Sciences
University of Novi Sad
Novi Sad, Serbia
rapaja@uns.ac.rs
ORCID: 0000-0003-0598-0979

Nikša Jakovljević
Faculty of Technical Sciences
University of Novi Sad
Novi Sad, Serbia
jakovnik@uns.ac.rs
ORCID: 0000-0002-7283-3939

***Abstract*— This paper explores the potential improvement in the SegFormer model's performance for lesion segmentation in mammograms (MGs) by incorporating saliency maps from the GLAM model. The GLAM model was trained on a million MGs, thus it is reasonable that the model demonstrates a broader scope of generalization compared to the models trained on just a few hundred or thousand images (which are available in the open datasets). Consequently, the GLAM model outputs can be considered as robust input features. The study was conducted by comparing the performance of the SegFormer model trained on *i*) exclusively on MGs (referred to as "only MG"), *ii*) a combination of saliency maps and MGs (referred to as "combined") and *iii*) exclusively on saliency maps (referred to as "only saliency"). The findings suggest that despite the GLAM model being pretrained on a significant number of MGs, the saliency maps it generated did not enhance the segmentation task. Instead, they introduced uncertainty for both the saliency-only and combined models. This led to an average F1 score of 25.65% and 49.91%, respectively, in comparison to the only MG model, which achieved a higher score of 52.95%.**

Keywords— GLAM, SegFormer, segmentation, mammographic images

I. Introduction

There is a clear need for an automated high-quality interpretation of digital mammography for timely support to combat breast cancer, the most frequent cancer type in women [1]. The problem of lesion detection in breast mammography (MG) images has received new incentives with advances in performance in classification on natural images using deep convolutional neural networks. This has been facilitated by the increasing availability of collected cancer medical imaging data, providing an opportunity to develop and advance timely and affordable cancer detection [2], [3] enhancing the efficiency in clinical workflow.

Breast cancer lesion segmentation models can be trained using state-of-the-art semantic segmentation frameworks developed based on powerful natural image segmentation architectures [4]. Reviews of models for breast segmentation tasks in [3] confirm the most frequent usage of U-Net architecture [5] and individual open datasets, modest in size and quality. Some recent works report results on the NYU Breast Cancer Screening Dataset (NYUBCS) [6], which contains about 1 million MG images but remains closed to the public. The NYUBCS database offers enough data to train deep MG segmentation models, without resorting to transfer learning from models trained on natural images. There are multiple aspects that differentiate natural from medical images, and in the context of MG, it is their high resolution and small regions of interest, critical for early cancer detection. Thus, downsizing MG to capture the global image features has to be complemented with high-capacity networks working on high resolution image patches, as in [7]–[9] where the ResNet architecture has been deployed. The globally aware multiple instance classifier (GMIC) [10] for breast cancer screening achieves an area under the curve of 0.93 on NYUBCS. This work is further extended into Global-Local Activation Maps

This research received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no 952179 (INCISIVE). It also has been supported by the Ministry of Science, Technological Development and Innovation (Contract No. 451-03-65/2024-03/200156) and the Faculty of Technical Sciences, University of Novi Sad through project "Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad" (No. 01-3394/1).

(GLAM) [9] which further exploits weekly supervised learning, and guided by class labels, the model additionally outputs high-resolution saliency maps identifying malignant and benign regions in the breast tissue, further improving the GMIC Dice-score for 20 % on NYUBCS.

Advances in natural language processing were a motivation to introduce transformers in computer vision tasks, starting from the Vision Transformer (ViT) [11] for classification, while its potential in semantic segmentations was shown in [12]. SegFormer architecture proposed in [4] is a ViT-based semantic segmentation framework introducing changes in both the encoder and decoder levels. SegFormer offers a positional-encoding-free and hierarchical transformer encoder and an all-multi-layer perceptron (All-MLP) lightweight decoder, simultaneously achieving accuracy, efficiency, and robustness [4].

In this work we try to build on the previous efforts regarding breast lesion segmentation [7]–[9] and use saliency maps outputs of the GLAM model, trained on about 1 million MG images, combined with the original image as the 3-channel input of the SegFormer MiT-B3 model [4]. We rigorously validate this performance for different datasets to monitor improvements in the segmentation task using the saliency maps, as an additional localization assistance. Moreover, the performance on different datasets has been observed to stress the change in model performance for heterogeneous datasets and the importance of model robustness to image quality.

The research is limited only to the SegFormer architecture [4], since in the previous study on the INBrest database [13] it obtained an F1 score (85.6%) [14] which was in the range of the best models (UNet 69.3%, FusionNet 73.2%, FCDenseNet103 76.1% and AUNet 79.1%)[15].

II. Databases

The databases that were used in this study are three public datasets CBIS-DDSM [16], INbreast [13], and CSAW-S [17], and a new MG dataset collected within the ongoing Horizon 2020 INCISIVE project [18].

CBIS-DDSM – Curated Breast Imaging Subset [16] is an updated and standardized version of the Digital Database for Screening Mammography (DDSM). The CBIS-DDSM collection includes a subset of the DDSM data selected and curated by a trained radiologist. All pathologies are confirmed histologically. Original scanned images stored in LJPEG format have been decompressed into 16-bit TIFF files and then converted into DICOM format. The image resolution varies in the range of 42-50 microns. The annotations include labels for calcifications and masses. Since the task in this study was lesion segmentation, only images containing masses (1514 of them) were included.

INbreast mammography dataset [13] is one of the most frequently exploited datasets because it contains highly accurate annotations proved by histological information. The MGs are genuinely digital, where the size of pixels is 70 microns and pixel contrast resolution is 14 bits. INbreast contains annotations for four different classes: masses, calcifications, asymmetries, and distortions. As in the previous dataset, only MGs with masses (107) were included.

CSAW-S – Cohort of Screen-Aged Women – Segmentation [17] is an open curated digital MG dataset where trained radiologists annotated cancer, calcifications, and lymph nodes, and a non-expert annotated other parts such as thick vessels, foreign objects, skin, nipple, text, non-mammary tissue, pectoral muscle, mammary gland, and background. Public data were available as 8-bit PNG images obtained from original DICOM files. Although the CSAW-S dataset contains 349 MGs, only 305 containing lesion labels were included in this study. Information about image resolution is not available.

INCISIVE-MG-L is a small subset of the INCISIVE dataset [18] containing 1563 annotated MGs stored in DICOM format. Experienced radiologists annotated image regions corresponding to lesions (benign, malignant, or suspicious), calcifications, and clips. Although there are only 3 data providers (AUTH, HCS, and UNS), data was collected from more than three different institutions, as HCS contained the data from several medical clinics. MGs obtained from AUTH and UNS are originally digital, while the HCS subset contains scanned MGs as well.

Since the lesions in MGs are usually small compared to the size of the whole image, in order to reduce class imbalance, the dataset is reduced to MGs containing lesions. The number of images per dataset and its share of the total number of images is illustrated in Fig. 1.

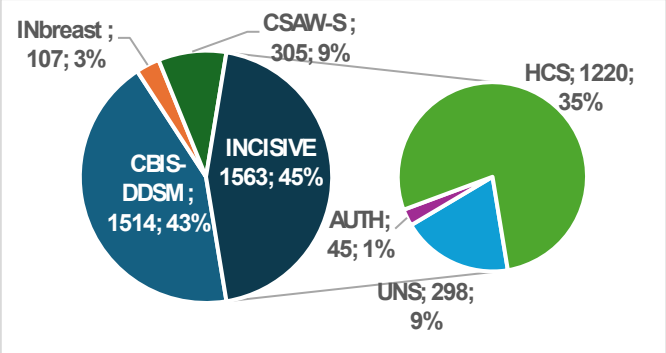

Fig. 1. Distribution of images per dataset

III. Methodology

Three experimental setups were evaluated using the same SegFormer [4] architecture to produce a lesion segmentation mask, but differ in the input. In the first setup, the input is a 3-channel grayscale image constructed from the preprocessed original MG (referred as “only MG”). In the second setup, the input image, as separate channels, contains the preprocessed original MG, and benign and malignant saliency maps obtained from the GLAM network outputs for given MG (referred to as “combined”). In the third setup, malignant saliency maps were employed to create a three-channel grayscale image, serving as the input. The last setup, referred as "only saliency", was not part of our initial study plan. It was prompted by the unsatisfactory F1 score of the combined model, aiming to get deeper insight of the impact of saliency maps.

All experimental setups have a preprocessing step to adapt

MGs to GLAM input, as the first step. In the combined and only saliency setups, pretrained GLAM model [9] was used to create saliency maps. After that, depending on the setup, preprocessed MG and saliency maps were used to create inputs for a SegFormer model. From this point, SegFormer models in all setups were trained and tested in the same manner.

A. MG Preprocessing

To adapt the used MGs to the GLAM input, several preprocessing steps have been applied, the results of which are shown in Fig. 2. The first step is to create uniform MGs with dark backgrounds (Fig. 2B). After that, images are constrained by removing constant edge pixels. The next step is applying linear interpolation based on their pixel spacing attribute, where this information is known, to a fixed spacing of 0.1. The images are oriented towards the right (Fig. 2C). Based on intensity levels, breast segmentation masks are obtained which are used to remove non-breast artifacts such as image label information (breast view position or laterality) by cropping (Fig. 2D). Using the same mask, based on histogram information of the covered pixels, a piecewise linear transformation is applied to the intensities to clip 2 % of the darkest and 1 % of the lightest intensities (Fig. 2E). Additionally, to adapt to image size requirements of the GLAM input the images are zero-padded to a fixed size of 2944×1920 pixels.

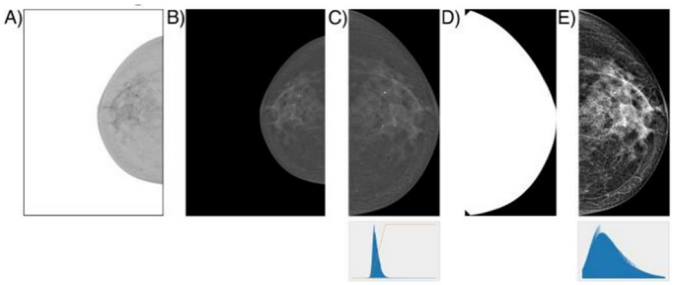

Fig. 2. Example of MMG transformation in preprocessing A) original image; B) negation; C) orientation toward right with histogram of the breast tissue; D) segmentation mask; E) preprocessed result with the histogram of the breast tissue.

B. Saliency Maps Creation

To create saliency maps, the pretrained GLAM model [9] is used, which was trained on 186816 examinations of NYU Breast Cancer Screening Dataset v1.0. The GLAM model consists of 2 modules (global and local) that jointly produce high-resolution saliency maps over the initial image. The global network, a ResNet-like architecture, combines 3 saliency maps at different scales to produce the global saliency map. The resulting global saliency map is used for the selection of patches that represent inputs for the local network. The local network is based on the ResNet-34 architecture and creates fine-grained saliency maps for each patch. Finally, the model aggregates the results from the local and global outputs to create the two final maps containing scores for each pixel belonging to a malignant or benign lesion. These scores for malignant and benign lesions will be further referred to in the text as malignant and benign saliency maps (see Fig. 3.). Additionally, as our model does not need to differentiate between benign and malignant lesions, these saliency maps were merged into a single saliency map encompassing all possible lesions, referred to in the text as "joint".

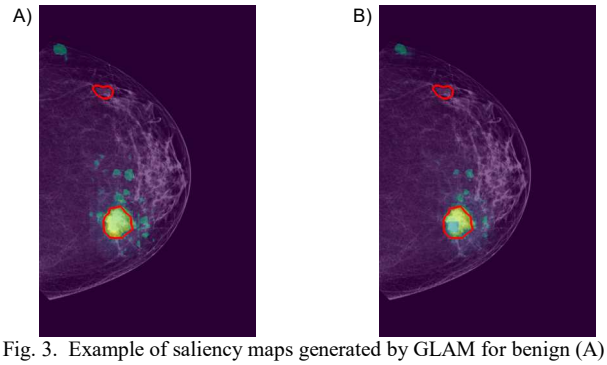

Fig. 3. Example of saliency maps generated by GLAM for benign (A) and malignant (B) lesions. Saliency maps are coded with viridis colormap where blue is 0 and yellow is 1. Red contours mark true lesions.

C. SegFormer Architecture

SegFormer [4] architecture consists of 3 main parts as illustrated in Fig. 4. The first part is the transformer encoder which produces hierarchical feature maps at various resolutions (backbone in Fig. 4). The second part is the neck which uses multi-layer perceptron (MLP) to up-sample these features to the resolution equal to one-fourth of the input image resolution. The last part is an MLP which produces a lesion segmentation mask.

1) Semantic Segmentation Backbone

The backbone extracts features with different resolutions in 4 stages ($128\times128\times C_1$, $64\times64\times C_2$, $32\times32\times C_3$, and $16\times16\times C_4$) from an input image whose size is 512×512×3 (Fig. 3). The basic building block is the Mix Transformer encoder (MiT) consisting of Overlap Patch Embedding, SegFormer Block, and Overlap Patch Merging. Authors in [4] proposed 6 different architectures MiT0–MiT5 which differ in the number of MiT in each stage, expansion ratio, and number of channels in the Overlap Patch Merging block. For the MiT-B3 encoder, which is used in this study, the number of channels in each stage follows the ResNet principle thus $C_1 = 64$, $C_2 = 128$, $C_3 = 320$, $C_4 = 512$, and the numbers of MiTs in each stage are $N_1 = 3$, $N_2 = 3$, $N_3 = 18$ and $N_4 = 3$.

Overlap Patch Embedding is used to transform input features into patches of predefined sizes, as in [11]. The applied stride is smaller than kernel size to ensure sharing of information between patches and it is implemented as a convolutional layer. The output of Overlap Patch Embedding goes into the SegFormer block.

SegFormer block normalizes input data, which goes through an efficient self-attention module whose output is added with input data. After that, these new features are normalized again, and go into the Mix feed-forward network whose mapping can be described as follows:

$$\mathbf{x}_{out} = \mathrm{MLP}\left(\mathrm{GELU}\left(\mathrm{DWC}_{3\times3}\left(\mathrm{MLP}(\mathbf{x}_{in})\right)\right)\right) + \mathbf{x}_{in} \qquad (1)$$

where $\mathbf{x}_{in}$ and $\mathbf{x}_{out}$ are input and output features, DWC – Depth-Wise Convolution [19], and GELU – Gaussian Error Linear Unit [20].

The major bottleneck of the SegFormer architecture is the self-attention module, thus it is implemented as an efficient

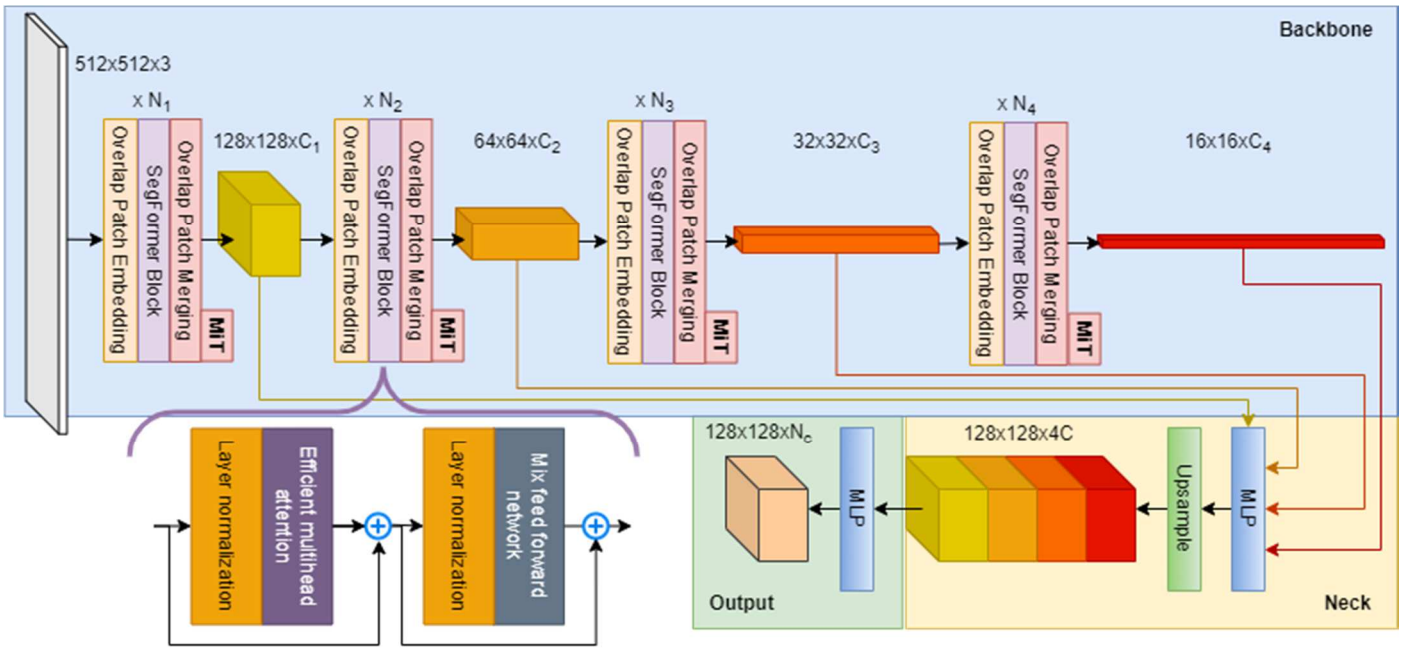

Fig.4 SegFormer architecture

realization described in [12]. The main idea is to reshape the matrix of key vectors and linear transform it into a new space and after that multiply it with the query matrix.

Overlap Patch Merging combines features generated by the SegFormer block and reorganizes them into required setting required space dimensions.

2) MLP decoder

By creating hierarchical feature maps in the backbone, that provide sufficient details, SegFormer can use a lightweight decoder consisting only of MLP layers. The first MLP layer transforms feature maps form different stages (in Fig. 4. represented as rectangular cuboids) such that they have the same number of features (in our case C = 768), and second layer up-sample them to have the same space dimension (in our case 128, see Fig. 4). Note that features from first stage do not require up-sampling. After that, additional MLP transforms the concatenated features into lesion segmentation mask ($N_C = 2$).

D.Experiment Setup

In all conducted experiments, we used the same setup for training SegFormer models. Instead of random initialization, the models' weights were initialized with the weights of the model [4] pretrained on several image datasets (ImageNet-1k, ADE20K, Cityscapes, and COCO-stuff). Since semantic segmentation is a classification task at the pixel level, cross-entropy was selected as a loss function. The model parameters were estimated by Adam optimizer with the learning rate 0.00006, decay rate for momentum 0.9, and decay rate for squared gradients 0.99. In order to avoid overfitting early stopping is applied if in 10 consecutive epochs F1 score on the validation set does not increase. Although the maximum number of epochs was set to 1000, training procedures stopped after 16 to 33 epochs. Due to memory limitations on GPU, the mini-batch size was set to 4.

IV. Results

Since the original GLAM implementation [9] does not provide an interpretation of saliency map values nor performs the analysis of an appropriate threshold for binarization, this work attempts to remap the outputs as accurately as possible to the expert annotations. Initial investigations were conducted on how well the obtained saliency maps match lesions within the dataset. As the probability values can be very low, we investigated a range of threshold values to create an initial binarized version of the maps to use for comparison. The best results were obtained for the threshold value 0.1 and F1 scores (observing only the positive, lesion class) are shown in Table 1. Raws UNS, AUTH, and HCS refer to different providers of mammography images in the INCISIVE dataset, while raw INCISIVE combines the results of all three providers observed jointly. The results indicate that there is a clear mismatch between the output maps even with a such generous threshold value, but considering the higher sensitivity result, the maps can exhibit the ability to detect something in areas of the lesions. However, these results are in line with the reported scores which have a high standard deviation in [9]. While our objective is to segment all types of lesions, the highest F1 scores were achieved with malignant ones. This is the reason why in the third experimental setup we selected malignant saliency maps as the input features.

During all conducted experimental setups MiT-B3 SegFormer architecture is used, because in previous investigations it achieved the best performances [14]. To obtain more confident results, a 5-fold cross-validation was used in all experiments. The obtained average F1 score plus-minus standard deviation across all folds are shown in Table 2. As mentioned before, in the first setup the model input was only MG (in Table 2. marked as Only MG) and in the second setup

TABLE 1
F1 SCORES IN PERCENTAGES BY DATASET FOR RESEMBLANCE OF BINARIZED MASKS PRODUCED BY GLAM MODEL AND GROUND TRUTH MASKS, FOR A THRESHOLD OF 0.1.

Dataset	Malignant	Benign	Joint
INbreast	**34.22**	30.86	30.14
CSAW-S	**22.14**	16.90	16.63
CBIS-DDSM	8.43	**8.53**	8.53
UNS	27.08	**27.71**	27.67
AUTH	**18.05**	18.47	18.47
HCS	**23.69**	22.32	22.09
INCISIVE	**24.18**	23.24	23.05
Total	**22.24**	19.88	19.34

TABLE 2
F1 SCORES IN PERCENTAGES OBTAINED FROM THE EXPERIMENTAL SETUPS TEST FOLDS, WHEN THE INPUT IS MGS (ONLY MG), COMBINATION OF MG AND SALIENCY MAPS (COMBINED), AND ONLY MALIGNANT SALIENCY MAPS.

Dataset	Only MG	Combined	Only Saliency
INbreast	**74.99 ± 5.49**	71.89 ± 5.35	56.17 ± 7.67
CSAW-S	34.91 ± 5.76	**39.54 ± 2.55**	30.33 ± 2.42
CBIS-DDSM	**56.78 ± 4.64**	51.22 ± 2.87	10.01 ± 1.10
UNS	**43.18 ± 9.43**	42.78 ± 7.98	25.47 ± 10.20
AUTH	**42.01 ± 13.78**	41.71 ± 15.38	16.9 ± 10.51
HCS	**46.12 ± 2.47**	42.35 ± 1.46	23.81 ± 1.14
INCISIVE	**45.04 ± 2.60**	41.83 ± 1.35	23.67 ± 2.62
Total	**52.95 ± 2.57**	49.91 ± 2.62	25.65 ± 2.77

model input was MG combined with the two GLAM saliency map outputs (in Table 2. marked as Combined). The SegFormer model trained only on MGs shows better performance overall, while the model with the combined input shows improvement in the case of CSAW-S dataset shows for the combined input. These results indicate that the saliency maps introduce only the noise. To prove this assumption, additional setup (termed "Only saliency") was introduced to train SegFormer only on malignant saliency maps (that produce the highest F1 scores in Table 1.) The "Only saliency" obtained results (Table 2) are the worst by the large margin. However, since they are better than those presented in Table 1, the saliency maps potentially contain information regarding true lesions besides the noise. This result eliminates another possible assumption that the SegFormer model pretrained on natural images cannot learn appropriate weights for given mixture of different pixel types (intensity and probability) per channel.

Regarding the images that were used for training the GLAM model, there is a resolution difference between them and the images in this study, thus potentially being a source of uncertainty from the obtained saliency maps. To this end, the input of the GLAM model was changed, so that the MGs were magnified in a way that the breast covered the most part of the image as in GLAM demo images. Unfortunately, there is no apparent difference between the obtained results.

V. CONCLUSION

This study aimed to exploit the knowledge incorporated in the GLAM model, trained on approximately one million MGs, for lesion segmentation. Despite its extensive training data, the GLAM model exhibited insufficient performance on open and our datasets. One potential explanation could be variations in imaging protocols and medical devices used for image acquisition between the NYUBCS dataset and other datasets. Typically, a standard approach to leverage a pretrained model involves its fine-tuning, but this was unattainable due to hardware limitations. Consequently, we explored an alternative by utilizing the saliency maps generated by the GLAM model. Our experiments showed no apparent improvement when compared to the model that used only MGs as input of the SegFormer. We presume that saliency maps do not contain sufficient information to distinguish lesions from normal tissue, which we confirmed by training a model using only saliency maps as input.

COMPLIANCE WITH ETHICAL STANDARDS

This work uses four databases of mammography images, each resulting from studies with obtained ethical approvals. Datasets open for the public used in this study are CBIS-DDSM, INbreast, and CSAW-S. This research study was conducted retrospectively using mammography images from human subjects which were made available in open access by Lee et. al [16] for CBIS-DDSM, Moreira et al. [13] for INbreast, and Matsoukas et al. [17] for CSAW-S. Ethical approval was not required as confirmed by the license attached with the open access data. Within the INCISIVE project, the data used in this study has been collected in the retrospective study in several hospitals using the same protocol which has been evaluated and approved by the corresponding ethical councils during in the ethical board meetings: (1) No. 4/21/1-40/ (11.01.2021) of Vojvodina Institute of Oncology, Serbia (as data provider for University of Novi Sad, Serbia), (2) no. 3936 (25.02.2021) for Theageneio hospital, Thessaloniki, Greece and decision no. 114/17-2-2021 for the Radiology Laboratory of the General Hospital "Papageorgiou", Thessaloniki, Greece (for the Aristotle University of Thessaloniki, Greece), (3) No 589/13-10-2020 General Anti-Cancer - Oncological Hospital of Saint Savvas, Athens, Greece and No. 10/30-03-2021 for Theageneio hospital, Thessaloniki, Greece (for the Hellenic Cancer Society, Greece). It is worth noting that all INCISIVE images were de-identified using CTP DICOM anonymizer, prior to being uploaded by the data providers to the central repository. All image analysis has been done on the AI development platform on the central repository, without image download.

ACKNOWLEDGMENT

The authors would like to acknowledge the contribution of the INCISIVE partners in the collection of breast cancer MMG data used in this study: Aristotle University of Thessaloniki (Greece), Oncology Institute of Vojvodina (University of Novi Sad, Serbia), and Hellenic Cancer Society (Greece). In addition, the authors would also like to acknowledge the work of the rest of the INCISIVE partners.

REFERENCES

[1] C. P. Wild, E. Weiderpass, and B. W. Stewart, "World Cancer Report: Cancer Research for Cancer Prevention". Lyon, France: International Agency for Research on Cancer, 2020.

[2] K. Munir, H. Elahi, A. Ayub, F. Frezza, and A. Rizzi, "Cancer Diagnosis Using Deep Learning: A Bibliographic Review," *Cancers*, vol. 11, no. 9, p. 1235, Aug. 2019, doi: 10.3390/cancers11091235.

[3] E. Michael, H. Ma, H. Li, F. Kulwa, and J. Li, "Breast Cancer Segmentation Methods: Current Status and Future Potentials," *BioMed Res. Int.*, vol. 2021, pp. 1–29, Jul. 2021, doi: 10.1155/2021/9962109.

[4] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, Eds., Curran Associates, Inc., 2021, pp. 12077–12090.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, vol. 9351, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., in Lecture Notes in Computer Science, vol. 9351. , Cham: Springer International Publishing, 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.

[6] N. Wu, J. Phang, J. Park, Y. Shen, S. G. Kim, L. Heacock, L. Moy, K. Cho and K. J. Geras, "The NYU Breast Cancer Screening Dataset v1.0," Sep. 2019. Accessed: Nov. 22, 2023. [Online]. Available: https://cs.nyu.edu/~kgeras/reports/datav1.0.pdf

[7] N. Wu, J. Phang, J. Park, Y. Shen, Z. Huang, M. Zorin, S. Jastrzebski, T. Fevry, J. Katsnelson, E. Kim, S. Wolfson, "Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening," *IEEE Trans. Med. Imaging*, vol. 39, no. 4, pp. 1184–1194, Apr. 2020, doi: 10.1109/TMI.2019.2945514.

[8] Y. Shen, N. Wu, J. Phang, J. Park, K. Liu, S. Tyagi, L. Heacock, S. G. Kim, L. Moy, K. Cho, K. J. Geras, "An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization," *Med. Image Anal.*, vol. 68, p. 101908, Feb. 2021, doi: 10.1016/j.media.2020.101908.

[9] K. Liu, Y. Shen, N. Wu, J. Chłędowski, C. Fernandez-Granda, and K. J. Geras, "Weakly-supervised High-resolution Segmentation of Mammography Images for Breast Cancer Diagnosis," *Proc. Mach. Learn. Res.*, vol. 143, pp. 268–285, Jul. 2021.

[10] Y. Shen, N. Wu, J. Phang, J. Park, G. Kim, L. Moy, K. Cho, K.J. Geras, "Globally-Aware Multiple Instance Classifier for Breast Cancer Screening," in *Machine Learning in Medical Imaging,* vol. 11861, H.-I. Suk, M. Liu, P. Yan, and C. Lian, Eds., in Lecture Notes in Computer Science, vol. 11861. , Cham: Springer International Publishing, 2019, pp. 18–26. doi: 10.1007/978-3-030-32692-0_3.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Yhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv, Jun. 03, 2021. Accessed: May 02, 2023. [Online]. Available: http://arxiv.org/abs/2010.11929.

[12] W. Wang, E. Xie, X. Li, DP. Fan, K. Song, D. Liang, T. Lu, P. Lu, L. Shao, "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions." arXiv, Aug. 11, 2021. Accessed: May 02, 2023. [Online]. Available: http://arxiv.org/abs/2102.12122.

[13] 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, Feb. 2012, doi: 10.1016/j.acra.2011.09.014.

[14] J. Kljajić, "Segmentacija lezija u mamografskim slikama korišćenjem SegFormer mrežne arhitekture," Zb. Rad. Fak. Teh. Nauka U Novom Sadu, vol. 38, no. 09, pp. 1204–1207, Sep. 2023, doi: 10.24867/24BE33Kljajic

[15] [1] H. Sun et al., "AUNet: attention-guided dense-upsampling networks for breast mass segmentation in whole mammograms," Phys. Med. Biol., vol. 65, no. 5, p. 055005, Feb. 2020, doi: 10.1088/1361-6560/ab5745.

[16] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Sci. Data*, vol. 4, no. 1, p. 170177, Dec. 2017, doi: 10.1038/sdata.2017.177.

[17] C. Matsoukas, A. B. Hernandez, Y. Liu, K. Dembrower, G. Miranda, E. Konuk, J. F. Haslum, A. Zouzos, P. Lindholm, F. Strand, K. Smith, "Adding Seemingly Uninformative Labels Helps in Low Data Regimes." arXiv, Aug. 11, 2020. Accessed: Nov. 21, 2023. [Online]. Available: http://arxiv.org/abs/2008.00807.

[18] I. Lazic, F. Agullo, S. Ausso, B. Alves, C. Barelle, J. L. Berral, P. Bizopoulos, O. Bunduc, I. Chouvarda, D. Dominguez, D. Filos, "The Holistic Perspective of the INCISIVE Project—Artificial Intelligence in Screening Mammography," *Appl. Sci.*, vol. 12, no. 17, p. 8755, Aug. 2022, doi: 10.3390/app12178755.

[19] Y. Guo, Y. Li, L. Wang, and T. Rosing, "Depthwise Convolution Is All You Need for Learning Multiple Visual Domains," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, pp. 8368–8375, Jul. 2019, doi: 10.1609/aaai.v33i01.33018368.

[20] D. Hendrycks and K. Gimpel, "Gaussian Error Linear Units (GELUs)." arXiv, Jun. 05, 2023. Accessed: Nov. 21, 2023. [Online]. Available: http://arxiv.org/abs/1606.08415.