Quantifying the effects of segmentation in image classification for melanoma recognition

Rafael Luz Araújo[1,2], Daniel de S. Luz[1,2], Bruno Vicente de Lima[4]
Júlio V. M. Marques[2], Rodrigo de M. S. Veras[3], Antônio O. de C. Filho[2]
Flávio H. D. Araújo[2], Romuere Rodrigues Veloso e Silva[2]

[1]Instituto Federal do Piauí (IFPI) – Picos – PI – Brasil.

[2]Universidade Federal do Piauí (UFPI) – Picos – PI – Brasil.

[3]Universidade Federal do Piauí (UFPI) – Teresina – PI – Brasil.

[4]Instituto Federal do Maranhão (IFMA) – Coelho Neto – MA – Brasil.

{rafaluzaraujo,daniel.luz}@ifpi.edu.br, brunovicente.lima@ifma.edu.br,

{juliomonteiro, rveras, antoniooseas, flavio86, romuere}@ufpi.edu.br

Abstract. *Melanoma remains the leading cause of skin cancer-related deaths worldwide, emphasizing the critical need for early detection to enhance survival rates. Computational methods are pivotal in aiding its diagnosis through medical imaging, necessitating accurate lesion segmentation to facilitate effective interpretation. Our study investigates the comparative efficacy of skin lesion classification with and without segmentation, leveraging pre-trained convolutional neural networks (CNNs) and CapsNet architectures. Findings underscore CNNs' superiority, highlighting segmentation's beneficial impact on their classification performance, while CapsNet exhibits a degree of independence from segmentation.*

1. Introduction

The number of people with cancer is increasing worldwide. The International Agency for Research on Cancer (IARC) reported that by 2020 the estimated global cancer burden has increased to 19.3 million new cases and 10 million deaths [IARC 2020]. This research addresses cancer known as malignant melanoma, the leading cause of death from skin cancer. Several studies indicate that the risk of malignant melanoma is correlated with genetic and personal features and with the behavior of exposure to ultraviolet radiation.

Melanoma is a serious form of cancer that starts in cells known as melanocytes and is dangerous because of its ability to spread to other organs quickly if not treated early [SCF 2021]. The medical examination of the skin performed by the specialist can result in an inaccurate diagnosis due to the similarity between the skin lesions and malignant tissues. Dermatologists have a 65% to 80% accuracy rate when making a diagnosis without additional technical support, such as a special high-resolution camera and magnifying glass [Argenziano and Soyer 2001].

Researchers are developing several Computer-Aided Diagnostic (CAD) tools that use medical imaging to assist professionals and provide additional insight. Two essential steps in the automatic diagnosis of melanoma are segmentation and classification of

the skin lesion. Segmentation involves isolating the region of interest to prevent interference from external elements during image analysis. According to [Tang et al. 2019], segmentation is a complex task due to images' overlapping elements, noise, shadows, and extraneous body parts. Classification entails determining whether an image exhibits the disease. Developing CAD systems for skin lesions in the last decade encountered challenges due to insufficient dataset sizes, hindering learning performance and feature extraction, thus complicating classification.

In this sense, this work has as main contributions: 1) Evaluation of the impact that segmentation causes in melanoma classification; 2) Comparison between pre-trained Convolutional Neural Networks and Capsule Networks architectures applied in melanoma classification.

2. Related Works

We've compiled various melanoma segmentation and classification studies, detailed in Table 1. Commonly used segmentation techniques include traditional methods like Otsu threshold and K-means. However, the most successful approaches in the literature involve deep learning-based methods such as U-net [Ronneberger et al. 2015]. Skin lesion classification employs both traditional methods and deep learning approaches.

Table 1. Related work on melanoma segmentation and classification.

Work	Datasets	Segmentation	Method
[Barata et al. 2013]	PH^2	x	Thresholding, color and texture descriptor
[Giotis et al. 2015]	MED-NODE	x	K-means, color and texture descriptor
[Jafari et al. 2016]	MED-NODE	x	Color and border descriptors
[Karabulut and Ibrikci 2016]	DERMIS	-	Texture descriptor
[Namozov and Im Cho 2018]	ISIC 2018	-	LeNet + APL Units
[Pal et al. 2018]	ISIC 2018	-	DCNN
[Reddy 2018]	ISIC 2018	-	Pre-trained CNNs
[Hekler et al. 2019]	ISIC 2018	-	Human + CNN
[Alom et al. 2019]	ISIC 2018	x	RRCNN
[Kassani and Kassani 2019]	ISIC 2018	-	DCNN
[Khan et al. 2019]	DERMIS	x	Clustering (K-means), color and texture descriptor
[Saba et al. 2019]	PH^2, ISBI 2016 e ISBI 2017	x	DCNN
[Sarkar et al. 2019]	DERMIS, PH^2, ISIC, MED-NODE	-	CNN
[Hosny et al. 2019]	DERMQUEST, MED-NODE, ISIC	-	Pre-trained CNNs
[Milton 2019]	ISIC 2018	-	Pre-trained CNNs
[Hosny et al. 2020]	DERMIS, DERMQUEST, MED-NODE, ISIC 2017	-	Pre-trained CNNs
[Amin et al. 2020]	ISIC 2018, PH^2, ISBI 2016, ISBI 2017	x	Thresholding and Pre-trained CNNs
[Al Nazi and Abir 2020]	PH^2, ISIC 2018	x	U-net and Pre-trained CNNs
[Rodrigues et al. 2020]	PH^2, ISIC	-	Pre-trained CNNs
[Moura et al. 2019]	DERMIS	-	Texture features and Deep features

We have identified some of the main features and their limitations in analyzing the segmentation and classification works. Some works use only an image dataset or concatenate some datasets to increase data and are not concerned with cross-evaluating

the datasets to investigate whether the model is learning features of the dataset or the disease. Others have few evaluation metrics, such as accuracy as a classification metric, which alone is not enough for this type of operation.

Most works pre-process and resize the images, not worrying about the distortion of skin lesions. We also found great use of data augmentation to solve challenges such as bases having few samples, unbalanced classes, and overfitting since deep learning models require many images for adequate training [Al-Masni et al. 2018]. Furthermore, among the studies that perform segmentation, only [Barata et al. 2013, Giotis et al. 2015, Jafari et al. 2016, Alom et al. 2019, Khan et al. 2019, Saba et al. 2019, Amin et al. 2020, Al Nazi and Abir 2020] perform the classification step, but they do not assess the impact that segmentation causes on melanoma classification.

We noticed that certain papers [Hosny et al. 2019, Hosny et al. 2020] applied data augmentation to the entire image set before dividing it into training and testing sets, resulting in identical images appearing in both sets with minor adjustments, leading to inflated results. Many studies either rely on physician-provided masks or forego segmentation altogether. While some studies generate segmentations, expert intervention is often required to rectify imperfect segmentations. Consequently, there needs to be more analysis regarding the impact of segmentation on classifying skin lesion images, which is the primary focus of this research.

3. Materials and Methods

The proposed methodology follows the steps presented in the Figure 1, starting with image acquisition; soon after, the segmentation obtains the regions of interest; in the classification, we define which class each image belongs to; and finally, we apply evaluation metrics to monitor the performance of the methods.

3.1. Image Acquisition

In this research, we gathered the most used public image datasets in the literature. All datasets have the classes Melanoma and non-melanoma, with the exception of ISIC 2018, which has seven classes.The details of each one are present in the Table 2.

Table 2. Information about datasets used.

Datasets	Classes	Images		Mask
		By class	All	
PH^2 [Mendonca et al. 2015]	Non-melanoma	160	200	yes
	Melanoma	40		
DermIS [DermIS 2020]	Non-melanoma	87	206	yes
	Melanoma	119		
ISIC 2018 [Codella et al. 2019, Tschandl et al. 2018]	Melanoma	1113	10015	no
	Nevus	6705		
	Basal cell carcinoma	514		
	Actinic keratosis	327		
	Benign keratosis	1099		
	Dermatofibroma	115		
	Vascular lesion	142		
MED-NODE [Giotis et al. 2015]	Non-melanoma	100	170	no
	Melanoma	70		

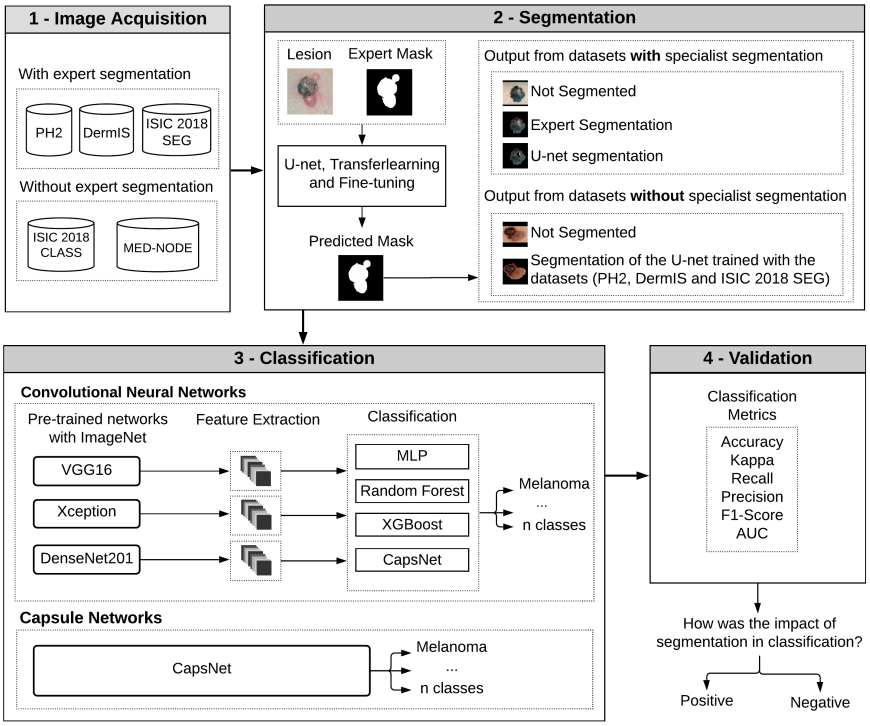

Figure 1. The flow of steps of the proposed methodology.

All datasets used have images in the RGB color system; most of them have different resolutions. Therefore, all the tests we performed used a proportional resizing algorithm that does not distort skin lesions. This resizing varied depending on the input shape of the architectures used for the classification, as detailed in Section 7. The resizing algorithm used bilinear interpolation, which fills the interpolated point with the weighted average of the four nearest pixels providing the image with smoother transitions in high-frequency locations. After resizing, the image is inserted in the center of a matrix with the chosen dimensions, with the pixels of each RGB channel having a value of 0.

4. Segmentation

We used a methodology based on U-net, Transfer learning, and Fine-tuning (UTF) that we developed in our work [Araújo et al. 2021] for the segmentation step. We performed experiments with the three datasets with specialist segmentation masks and obtained promising results, with Dice coefficient (DSC) = 0.923 in the PH2 dataset, DSC = 0.879 in the DermIS dataset, and DSC = 0.893 in the ISIC 2018 dataset. Figure 2 presents some examples of segmentations obtained with the proposed methodology.

5. Classification

In two classification approaches, we compare Convolutional Neural Networks (CNNs) with Capsule Networks (CapsNet).

Convolutional Neural Networks: We utilized pre-trained VGG16 [Simonyan and Zisserman 2014], Xception [Chollet 2017], and DenseNet201

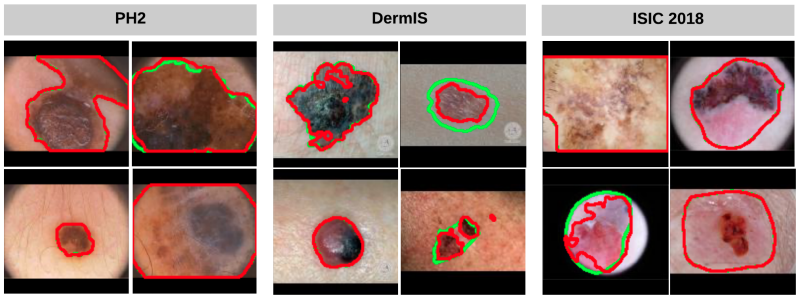

Figure 2. Examples of segmentations. In red, the segmentation with the proposed methodology, and in green, the specialist's marking.

[Huang et al. 2017] networks, pre-trained on ImageNet, selected for their performance and prevalence in the literature. These networks extracted features from images for analysis by four classifiers: Multi-Layer Perceptron (MLP), Random Forest, XGBoost, and CapsNet. CapsNet was used to evaluate its efficiency in recognizing spatial patterns within high-level features extracted by CNN compared to traditional classifiers. We removed the fully connected layer from the networks and did not fine-tune the convolutional layers. Additionally, empirical tests were conducted to determine the best hyperparameters for the classifiers: 1) MLP with 256 neurons, dropout of 0.5, and 30 epochs; 2) Random Forest with random state = 50 and 100 estimators; 3) XGBoost with 100 estimators, maximum depth of 3, and learning rate of 1.0.

Capsule Networks: we use CapsNet with the same settings as [Sabour et al. 2017]. Figure 3 illustrates the used CapsNet architecture. The image with a 64×64 size in the RGB model enters the encoder that starts with a convolutional layer to detect the features that the capsules will analyze. These extracted features serve as input to the primary capsules, where 32 different capsules apply eight $9 \times 9 \times 256$ convolutional kernels producing a 4D vector output that, through dynamic routing, are routed to the higher-level capsules (Melanoma capsule). Melanoma capsules will produce a 16D vector containing all the instantiation parameters needed to reconstruct the lesions. Then, a decoder receives the output vector and learns how to decode the lesion instantiation parameters. The decoder used is a neural network composed of two dense layers of 512 and 1024 neurons and a final layer with a softmax activation function.

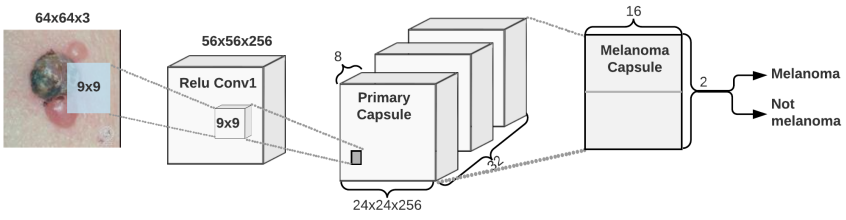

Figure 3. CapsNet architecture used in classification.

6. Validation

The validation step consists of measuring the results of the techniques. For this, we use the classification metrics most present in the literature. The classification metrics used were: Accuracy (ACC), Recall, Precision (PRE), F1-Score (F1), Area under the curve receiver operating characteristic (AUC) and Coefficient kappa (Kappa) [Cohen 1960].

7. Experimental Results and Discussion

In this section, we present the results of the two proposed classification approaches. The first uses pre-trained CNNs, and the second uses Capsule Networks.

7.1. Results of Convolutional Neural Networks

We divided this experiment into two steps: 1) identify the best pre-trained CNN to extract the features and which is the best classifier; 2) evaluate the impact of segmentation on ranking using the best combination of CNN and classifier obtained. During the experiments, we scaled the images proportionally to the size 224x224, as it obtained better results in preliminary tests as it is the standard size for most pre-trained networks. Finally, we randomly split the datasets into 60% training, 20% validation, and 20% testing.

The DenseNet201 architecture provides 94,080 features, Xception offers 100,352, and VGG16 yields 25,088. Reshaping the network's output was essential to integrate CapsNet as a classifier. We determined the optimal combination of pre-trained CNN and classifier through 60 tests detailed in Table 3. Based on these findings, we selected Xception for feature extraction and CapsNet for classification.

Several results fell below expectations, suggesting potential issues like imperfect segmentation, an imbalance or an insufficient number of images for effective training, or a necessity for fine-tuning to tailor the convolutional layers to skin lesion characteristics. Recall and precision metrics more accurately identified the negative class over the positive, except for the MED-NODE dataset. This discrepancy is problematic, particularly in melanoma detection, as false negatives can delay treatment for individuals with the disease, jeopardizing early intervention.

We employed the combination (Xception + CapsNet) to assess segmentation's impact on classification. We tested non-segmented images with specialist masks and images segmented using the UTF method [Araújo et al. 2021]. Since the MED-NODE and ISIC 2018 datasets lacked specialist masks, we segmented them using U-net trained on datasets with available masks. Table 4 presents the test results.

Upon examining the DermIS dataset, we observed that expert segmentation significantly improved classification, raising the ACC from 0.6904 to 0.8571. Additionally, U-net segmentation yielded gains, albeit less pronounced, elevating the ACC to 0.7142. This suggests that U-net achieved a DSC of 0.879 on DermIS, approximately 12.1% lower than the expert's (100%), leading to a decrease in positive impact from 16.67% to 2.38% (roughly -14.29%), yet still surpassing unsegmented images.

In the PH^2 dataset, specialist and U-net segmentation achieved the same positive performance of +2.5%. U-net's DSC in this base is 0.932, higher than the previous base, but the performance gain in the classification was smaller. The U-net trained in the DermIS, PH^2 and ISIC 2018 SEG datasets obtained DSC = 0.893 and presented high generalization capacity when segmenting the MED-NODE dataset with no specialist segmentation. The segmentation made the ACC of the MED-NODE rating rise from 0.8823 to 0.9411 (+5.88%).

The impact was negative in the 2018 ISIC dataset (-3.5%), as the ACC decreased from 0.7615 to 0.7265. Finally, we performed a test with the combination of all datasets to solve the problem of having too few sample images. The results indicated that there is

Table 3. Test results to choose the best CNN and classifier.

Dataset	CNN	Classifier	ACC	Kappa	Recall	PRE	F1
DermIS	DenseNet201	MLP	0.6666	0.2794	0.6319	0.6781	0.6250
		Random Forest	0.5952	0.1904	0.5972	0.5952	0.5931
		XGBoost	0.6904	0.3546	0.6736	0.6851	0.6755
		CapsNet	**0.6904**	**0.3809**	**0.6944**	**0.6904**	**0.6888**
	Xception	MLP	0.7142	0.3913	0.6875	0.725	0.6888
		Random Forest	**0.8095**	**0.5942**	**0.7847**	**0.8416**	**0.7925**
		XGBoost	0.5714	0.1000	0.5486	0.5535	0.5456
		CapsNet	0.7142	0.4084	0.7013	0.7091	0.7035
	VGG16	MLP	0.6666	0.2575	0.6180	0.7361	0.5916
		Random Forest	0.6428	0.2222	0.6041	0.6515	0.5906
		XGBoost	0.5952	0.1438	0.5694	0.5795	0.5654
		CapsNet	**0.7142**	**0.3823**	**0.6805**	**0.7437**	**0.6785**
PH^2	DenseNet201	*MLP*	0.9250	0.7272	0.8125	0.9571	0.8622
		Random Forest	0.9000	0.6551	0.7968	0.8725	0.8268
		XGBoost	0.8250	0.5205	0.7968	0.7382	0.7584
		CapsNet	**0.9500**	**0.8275**	**0.8750**	**0.9705**	**0.9134**
	Xception	**MLP**	**0.9500**	**0.8275**	**0.8750**	**0.9705**	**0.9134**
		Random Forest	**0.9500**	**0.8275**	**0.8750**	**0.9705**	**0.9134**
		XGBoost	0.9000	0.6875	0.8437	0.8437	0.8437
		CapsNet	**0.9500**	**0.8275**	**0.8750**	**0.9705**	**0.9134**
	VGG16	MLP	0.9000	0.6551	0.7968	0.8725	0.8268
		Random Forest	0.9000	0.6551	0.7968	0.8725	0.8268
		XGBoost	0.8000	0.4285	0.7343	0.7000	0.7132
		CapsNet	**0.9250**	**0.7540**	**0.8593**	**0.8982**	**0.8769**
MED-NODE	DenseNet201	**MLP**	**0.8529**	**0.7058**	**0.8642**	**0.8529**	**0.8517**
		Random Forest	0.7058	0.3511	0.6642	0.7211	0.6640
		XGBoost	0.7352	0.4813	0.7535	0.7491	0.7350
		CapsNet	0.7647	0.4925	0.7357	0.7750	0.7424
	Xception	MLP	0.9117	0.8197	0.9142	0.9070	0.9098
		Random Forest	0.7941	0.5509	0.7607	0.8244	0.7700
		XGBoost	0.7647	0.5244	0.7678	0.7604	0.7614
		CapsNet	**0.9411**	**0.8811**	**0.9500**	**0.9375**	**0,9403**
	VGG16	**MLP**	**0.7647**	**0.5142**	**0.7571**	**0.7571**	**0.7571**
		Random Forest	0.6176	0.1264	0.5571	0.6103	0.5252
		XGBoost	0.6470	0.2714	0.6357	0.6357	0.6357
		CapsNet	0.7352	0.4476	0.7214	0.7271	0.7235
ISIC 2018	**DenseNet201**	MLP	0.6735	0.0348	0.1901	0.2656	0.1874
		Random Forest	0.6325	0.2714	0.2479	0.1330	0.1492
		XGBoost	0.6493	0.2652	0.2423	0.2533	0.2415
		CapsNet	**0.757**	**0.5040**	**0.4593**	**0.5910**	**0.4759**
	Xception	MLP	0.6670	0.0380	0.1480	0.1677	0.1264
		Random Forest	0.6145	0.2614	0.2441	0.1327	0.1458
		XGBoost	0.6495	0.2847	0.2533	0.2765	0.2615
		CapsNet	**0.7265**	**0.4430**	**0.3669**	**0.4199**	**0.3813**
	VGG16	MLP	0.6775	0.0712	0.1603	0.2391	0.1460
		Random Forest	0.6170	0.2195	0.2279	0.1264	0.1429
		MLP	0.6620	0.3279	0.3014	0.4519	0.3209
		XGBoost	0.6150	0.2182	0.2221	0.2310	0.2255
		CapsNet	**0.7135**	**0.3602**	**0.3093**	**0.4452**	**0.3407**

The results in bold represent the best performance obtained in this comparison.

still a need to solve the class imbalance problem. The impact of segmentation in this test was also negative, but we conclude that it is due to the presence of the ISIC 2018 dataset.

We performed additional analysis to understand why segmentation impaired the classification of the 2018 ISIC dataset. Of the 2000 images in the test set, segmentation impaired the classification of 207 images and improved 147. Looking at these images, we found that the segmentations are accurate and the difficulty in distinguishing the seven classes that cause the decrease in performance. With that, we concluded that, for all datasets with only two classes, segmentation brought great performance gains. But for ISIC 2018, which has seven classes with similar lesions, segmentation diminishes the model's distinguishing ability.

7.2. Capsule Network Results

CapsNet's experiments aim to compare their performance with traditional CNNs. In preliminary tests, we identified that the 64×64 size achieved the best results. The images

Table 4. Segmentation impact results on classification using (Xception + CapsNet). Where (+) positive and (-) negative.

Datasets		ACC	Kappa	Recall	PRE	F1	AUC	Impact
DermIS (2 classes)	No Segmentation	0.6904	0.3809	0.6944	0.6904	0.6888	0.7893	
	Expert Segmentation	**0.8571**	**0.7042**	**0.8472**	**0.8605**	**0.8517**	**0.9212**	+
	U-net Segmentation	0.7142	0.4084	0.7013	0.7091	0.7035	0.7476	+
PH^2 (2 classes)	No Segmentation	0.9250	0.7540	0.8593	0.8982	0.8769	0.9160	
	Expert Segmentation	**0.9500**	**0.8275**	**0.8750**	**0.9705**	**0.9134**	**0.9863**	+
	U-net Segmentation	**0.9500**	**0.8275**	**0.8750**	**0.9705**	**0.9134**	**0.9863**	+
MED-NODE (2 classes)	No Segmentation	0.8823	0.7571	0.8785	0.8785	0.8785	0.9053	
	U-net Segmentation	**0.9411**	**0.8811**	**0.9500**	**0.9375**	**0.9403**	**0.9857**	+
ISIC 2018 CLASS (7 classes)	No Segmentation	**0.7615**	**0.4910**	**0.4507**	**0.5775**	**0.4953**	**0.8626**	
	U-net Segmentation	0.7265	0.4430	0.3669	0.4199	0.3813	0.8089	-
All* (2 classes)	No Segmentation	**0.8947**	**0.3803**	**0.6461**	**0.7953**	**0.6858**	**0.8254**	
	U-net Segmentation	0.8871	0.2917	0.6051	0.7768	0.6381	0.7335	-

(*) Combination of all datasets that are divided in two classes: melanoma and non-melanoma (all other classes).
The results in bold represent the best performance obtained in this comparison.

used were in their original RGB model, and we randomly divided the datasets between training, validation, and testing, with a proportion of 60%, 20%, and 20%, respectively. Table 5 presents the results of the execution of CapsNet for 100 epochs in the four datasets, and the combination of all datasets.

Table 5. Segmentation impact results on classification using CapsNet. Where (+) positive and (-) negative.

Datasets		ACC	Kappa	Recall	PRE	F1	AUC	Impact
DermIS (2 classes)	No Segmentation	**0.7500**	**0.4681**	**0.7292**	**0.7418**	**0.7333**	0.5560	
	Expert Segmentation	0.7000	0.3750	0.6875	0.6875	0.6875	0.7070	-
	U-net Segmentation	0.7000	0.3617	0.6771	0.6868	0.6800	**0.7266**	-
PH^2 (2 classes)	No Segmentation	**0.9500**	**0.8276**	**0.875**	**0.9706**	**0.9134**	**0.8828**	
	Expert Segmentation	0.8750	0.4898	0.6875	0.9324	0.7365	0.8223	-
	U-net Segmentation	0.8750	0.5455	0.7344	0.8429	0.7704	0.8242	-
MED-NODE (2 classes)	No Segmentation	0.8333	0.6471	0.8143	0.8438	0.8222	0.7893	
	U-net Segmentation	**0.8750**	**0.7391**	**0.8643**	**0.8778**	**0.8693**	**0.8893**	+
ISIC 2018 (7 classes)	No Segmentation	**0.7455**	**0.4615**	**0.4295**	**0.5209**	**0.4534**	**0.8817**	
	U-net Segmentation	0.6978	0.3154	0.2640	0.3708	0.2877	0.8417	-
All* (2 classes)	No Segmentation	**0.8804**	**0.2318**	**0.5794**	**0.7772**	**0.6039**	**0.8013**	
	U-net Segmentation	0.8726	0.0982	0.5305	0.7484	0.5263	0.7909	-

(*) Combination of all datasets that are divided in two classes: melanoma and non-melanoma (all other classes).
The results in bold represent the best performance obtained in this comparison.

Regarding CapsNet, the segmentation had a positive impact only on the MED-NODE dataset and had a better performance than CNN for non-segmented images, indicating that CapsNet needs information around skin lesions to identify them better. We believe that it considers information about the spatial relationships between healthy skin and lesions important, and this information is lost after segmentation.

7.3. Comparison with related works

The Table 6 presents a comparison with related studies, showing promising results for our proposed methods. In the PH^2 dataset, [Sarkar et al. 2019] and [Saba et al. 2019]

achieved higher accuracy (ACC) than our CapsNet-based approach, with 0.9677 and 0.9840, respectively, compared to our 0.9500 ACC. For the DermIS dataset, [Sarkar et al. 2019], [Hosny et al. 2019], and [Hosny et al. 2020] achieved recall rates of 1.0000, 0.9690, and 0.9892, respectively. Notably, some studies applied data augmentation across the entire dataset, biasing performance evaluation, unlike our approach, which applies augmentation only to the training set. In the ISIC 2018 dataset, our method (Xception + CapsNet) achieved similar ACC results to [Milton 2019], [Pal et al. 2018], and [Hekler et al. 2019], with 0.7615 compared to 0.7600, 0.7750, and 0.8295, respectively.

Table 6. Comparison of the proposed methods with the related works. The results in bold represent the best performance obtained in this comparison.

PH^2						
Work	**ACC**	**Kappa**	**Recall**	**PRE**	**F1**	**AUC**
[Barata et al. 2013]	—	—	0.9300	—	—	—
[Bi et al. 2016]	0.9200	—	0.8750	—	—	—
[Sarkar et al. 2019]	0.9677	—	**1.0000**	—	—	—
[Saba et al. 2019]	**0.9840**	—	0.9825	—	**0.9827**	**1.0000**
[Al Nazi and Abir 2020]	0.9200	—	—	—	—	—
[Rodrigues et al. 2020]	0.9316	—	0.9316	0.9325	0.9315	—
Proposed (Xception + CapsNet)	0.9500	0.8275	0.8750	0.9705	0.9134	0.9785
Proposed (CapsNet)	0.9500	**0.8276**	0.8750	**0.9706**	0.9134	0.9688
DermIS						
Work	**ACC**	**Kappa**	**Recall**	**PRE**	**F1**	**AUC**
[Karabulut and Ibrikci 2016]	0.7140	—	0.7080	—	—	—
[Khan et al. 2019]	0.9600	—	—	—	—	—
[Sarkar et al. 2019]	0.9444	—	**1,0000**	0.9166	—	—
[Hosny et al. 2019]	0.9686	—	0.9690	**0.9692**	—	—
[Hosny et al. 2020]	**0.9915**	—	0.9892	—	—	—
Proposed (Xception + CapsNet)	0.8571	0.7042	0.8472	0.8605	**0.8517**	**0.9212**
Proposed (CapsNet)	0.7500	0.4681	0.7292	0.7418	0.7333	0.5560
ISIC 2018						
Work	**ACC**	**Kappa**	**Recall**	**PRE**	**F1**	**AUC**
[Namozov and Im Cho 2018]	**0.9586**	—	—	—	—	—
[Pal et al. 2018]	0.7750	—	—	—	—	—
[Reddy 2018]	0.9100	—	—	—	—	—
[Kassani and Kassani 2019]	0.9208	—	—	—	—	—
[Hekler et al. 2019]	0.8295	—	—	—	—	—
[Alom et al. 2019]	—	—	**0.8700**	**0.8700**	**0.8600**	—
[Milton 2019]	0.7600	—	—	—	—	—
Proposed (Xception + CapsNet)	0.7615	**0.4910**	0.4507	0.5775	0.4953	0.8626
Proposed (CapsNet)	0.7455	0.4615	0.4295	0.5209	0.4534	**0.8817**
MED-NODE						
Work	**ACC**	**Kappa**	**Recall**	**PRE**	**F1**	**AUC**
[Giotis et al. 2015]	0.8100	—	0.8100	—	—	—
[Jafari et al. 2016]	0.7900	—	0.9000	—	—	—
[Han et al. 2018]	—	—	0.8763	—	—	—
[Sarkar et al. 2019]	0.9523	—	0.9233	**1,0000**	—	—
[Hosny et al. 2019]	0.9770	—	0.9922	—	—	—
[Hosny et al. 2020]	**0.9929**	—	**0.9922**	—	—	—
Proposed (Xception + CapsNet)	0.9411	0.8811	0.9500	0.9375	**0.9403**	**0.9857**
Proposed (CapsNet)	0.8750	0.7391	0.8643	0.8778	0.8693	0.8893

The results in bold represent the best performance obtained in this comparison.

In the MED-NODE dataset, recent studies, including [Han et al. 2018] and [Sarkar et al. 2019], achieved promising recall rates of 0.8763 and 0.9233, respectively. Our proposed method (Xception + CapsNet) surpassed these with a recall of 0.9500. Although [Hosny et al. 2020] reported a higher recall of 0.9922, this result may be attributed to the inappropriate use of data augmentation, as discussed earlier.

8. Conclusion and Future Works

Our study employed CNN (Xception + CapsNet) and Capsule networks (CapsNet) to evaluate segmentation's effect on classification. Our findings closely resembled existing literature, revealing that studies achieving near-perfect performance often utilized dataset-wide data augmentation, leading to test set bias. Segmentation positively impacted CNN tests, except for ISIC 2018, and only in $\frac{1}{7}$ of CapsNet tests. It demonstrated the potential to enhance CNN performance, contingent on its accuracy. However, segmentation could adversely affect non-segmented images in cases like ISIC 2018, where image classes are similar. Additionally, CapsNet did not benefit from segmentation, suggesting its reliance on spatial lesion-skin relationships, which detracts from segmentation performance.

For future work, we aim to refine segmentation masks from U-net using Autoencoders Networks, explore Capsule Networks in novel approaches to validate their segmentation benefits and devise a hybrid method for melanoma diagnosis. This approach will integrate CNN and CapsNet features without requiring the segmentation step.

References

Al-Masni, M. A., Al-Antari, M. A., Choi, M.-T., Han, S.-M., and Kim, T.-S. (2018). Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks. *Computer methods and programs in biomedicine*, 162:221–231.

Al Nazi, Z. and Abir, T. A. (2020). Automatic skin lesion segmentation and melanoma detection: Transfer learning approach with u-net and dcnn-svm. In *Proceedings of International Joint Conference on Computational Intelligence*, pages 371–381. Springer.

Alom, M. Z., Aspiras, T., Taha, T. M., and Asari, V. K. (2019). Skin cancer segmentation and classification with nabla-n and inception recurrent residual convolutional networks. *arXiv preprint arXiv:1904.11126*.

Amin, J., Sharif, A., Gul, N., Anjum, M. A., Nisar, M. W., Azam, F., and Bukhari, S. A. C. (2020). Integrated design of deep features fusion for localization and classification of skin cancer. *Pattern Recognition Letters*, 131:63–70.

Araújo, R. L., de Araújo, F. H., and Silva, R. R. (2021). Automatic segmentation of melanoma skin cancer using transfer learning and fine-tuning. *Multimedia Systems*, pages 1–12.

Argenziano, G. and Soyer, H. P. (2001). Dermoscopy of pigmented skin lesions–a valuable tool for early. *The lancet oncology*, 2(7):443–449.

Barata, C., Ruela, M., Francisco, M., Mendonça, T., and Marques, J. S. (2013). Two systems for the detection of melanomas in dermoscopy images using texture and color features. *IEEE Systems Journal*, 8(3):965–979.

Bi, L., Kim, J., Ahn, E., Feng, D., and Fulham, M. (2016). Automatic melanoma detection via multi-scale lesion-biased representation and joint reverse classification. In *2016 IEEE 13th international symposium on biomedical imaging (ISBI)*, pages 1055–1058. IEEE.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258.

Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al. (2019). Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

DermIS (2020). Dermatology information system. `https://www.dermis.net/dermisroot/en/home/index.htm`. Online; accessed 25 June 2020.

Giotis, I., Molders, N., Land, S., Biehl, M., Jonkman, M. F., and Petkov, N. (2015). Med-node: A computer-assisted melanoma diagnosis system using non-dermoscopic images. *Expert systems with applications*, 42(19):6578–6585.

Han, S. S., Kim, M. S., Lim, W., Park, G. H., Park, I., and Chang, S. E. (2018). Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *Journal of Investigative Dermatology*, 138(7):1529–1538.

Hekler, A., Utikal, J. S., Enk, A. H., Hauschild, A., Weichenthal, M., Maron, R. C., Berking, C., Haferkamp, S., Klode, J., Schadendorf, D., et al. (2019). Superior skin cancer classification by the combination of human and artificial intelligence. *European Journal of Cancer*, 120:114–121.

Hosny, K. M., Kassem, M. A., and Foaud, M. M. (2019). Classification of skin lesions using transfer learning and augmentation with alex-net. *PloS one*, 14(5):e0217293.

Hosny, K. M., Kassem, M. A., and Foaud, M. M. (2020). Skin melanoma classification using roi and data augmentation with deep convolutional neural networks. *Multimedia Tools and Applications*, 79(33):24029–24055.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

IARC (2020). Latest global cancer data: Cancer burden rises to 19.3 million new cases and 10.0 million cancer deaths in 2020. *International Agency for Research on Cancer*.

Jafari, M. H., Samavi, S., Karimi, N., Soroushmehr, S. M. R., Ward, K., and Najarian, K. (2016). Automatic detection of melanoma using broad extraction of features from digital images. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1357–1360. IEEE.

Karabulut, E. and Ibrikci, T. (2016). Texture analysis of melanoma images for computer-aided diagnosis. In *Int. Conference on Intelligent Computing, Computer Science & Information Systems (ICCSIS 16)*, volume 2, pages 26–29.

Kassani, S. H. and Kassani, P. H. (2019). A comparative study of deep learning architectures on melanoma detection. *Tissue and Cell*, 58:76–83.

Khan, M. Q., Hussain, A., Rehman, S. U., Khan, U., Maqsood, M., Mehmood, K., and Khan, M. A. (2019). Classification of melanoma and nevus in digital images for diagnosis of skin cancer. *IEEE Access*, 7:90132–90144.

Mendonca, T., Celebi, M., Mendonca, T., and Marques, J. (2015). Ph2: A public database for the analysis of dermoscopic images. In *Dermoscopy image analysis*. CRC Press.

Milton, M. A. A. (2019). Automated skin lesion classification using ensemble of deep neural networks in isic 2018: Skin lesion analysis towards melanoma detection challenge. *arXiv preprint arXiv:1901.10802*.

Moura, N., Veras, R., Aires, K., Machado, V., Silva, R., Araújo, F., and Claro, M. (2019). Abcd rule and pre-trained cnns for melanoma diagnosis. *Multimedia Tools and Applications*, 78(6):6869–6888.

Namozov, A. and Im Cho, Y. (2018). Convolutional neural network algorithm with parameterized activation function for melanoma classification. In *2018 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 417–419. IEEE.

Pal, A., Ray, S., and Garain, U. (2018). Skin disease identification from dermoscopy images using deep convolutional neural network. *arXiv preprint arXiv:1807.09163*.

Reddy, N. D. (2018). Classification of dermoscopy images using deep learning. *arXiv preprint arXiv:1808.01607*.

Rodrigues, D. d. A., Ivo, R. F., Satapathy, S. C., Wang, S., Hemanth, J., and Reboucas Filho, P. P. (2020). A new approach for classification skin lesion based on transfer learning, deep learning, and iot system. *Pattern Recognition Letters*, 136:8–15.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Saba, T., Khan, M. A., Rehman, A., and Marie-Sainte, S. L. (2019). Region extraction and classification of skin cancer: A heterogeneous framework of deep cnn features fusion and reduction. *Journal of medical systems*, 43(9):1–19.

Sabour, S., Frosst, N., and Hinton, G. E. (2017). Dynamic routing between capsules. *arXiv preprint arXiv:1710.09829*.

Sarkar, R., Chatterjee, C. C., and Hazra, A. (2019). Diagnosis of melanoma from dermoscopic images using a deep depthwise separable residual convolutional network. *IET Image Processing*, 13(12):2130–2142.

SCF (2021). Melanoma overview: A dangerous skin cancer. Disponível em: https://www.skincancer.org/skin-cancer-information/melanoma/. Accessed on: 15/02/2021.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tang, P., Liang, Q., Yan, X., Xiang, S., Sun, W., Zhang, D., and Coppola, G. (2019). Efficient skin lesion segmentation using separable-unet with stochastic weight averaging. *Computer methods and programs in biomedicine*, 178:289–301.

Tschandl, P., Rosendahl, C., and Kittler, H. (2018). The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5:180161.