Nova Information Management School

Master Program in Data Science and Advanced Analytics

Report of Deep Learning's project

CLASSIFICATION OF SKIN LESIONS IMAGES

Group 2: Biazi Bayer, Breno Lehmann, Carolina Bellani and Luiz Samuel Corradi

INTRODUCTION AND TASK DEFINITION

A skin lesion is an abnormal lump, bump, ulcer, score or colored area on the skin. Skin cancer is one of the most common cancers in the world and more people are diagnosed with skin cancer each year in the U.S. than all other cancers combined (Cancer Facts and Figures 2019, 2019).

The skin cancer is firstly inspected visually in the initial clinical screening, being followed by dermatoscopic analysis, histopathological examination, follow-up examination, expert consensus and confirmation by in-vivo microscopy.

A correct diagnosis of the skin lesions are extremely important to start the correct treatment and prevent skin cancer.

A previous application (M. Binder, 1994) used dermatoscopic images successfully to train an artificial neural network to differentiate melanomas, the deadliest type of skin cancer, from melanocytic nevi. Although the results were promising, the study, like most earlier studies, suffered from a small sample size and the lack of dermatoscopic images other than melanoma or nevi.

Building a classifier for multiple diseases is more challenging than binary classification and currently, reliable multi-class predictions are only available for clinical images of skin diseases but not for dermatoscopic images.

The 10,015 dermatoscopic images of this project were extracted from the Harvard Dataverse (Tschandl, 2018) and they include 7 lesion's types (Figure 1).

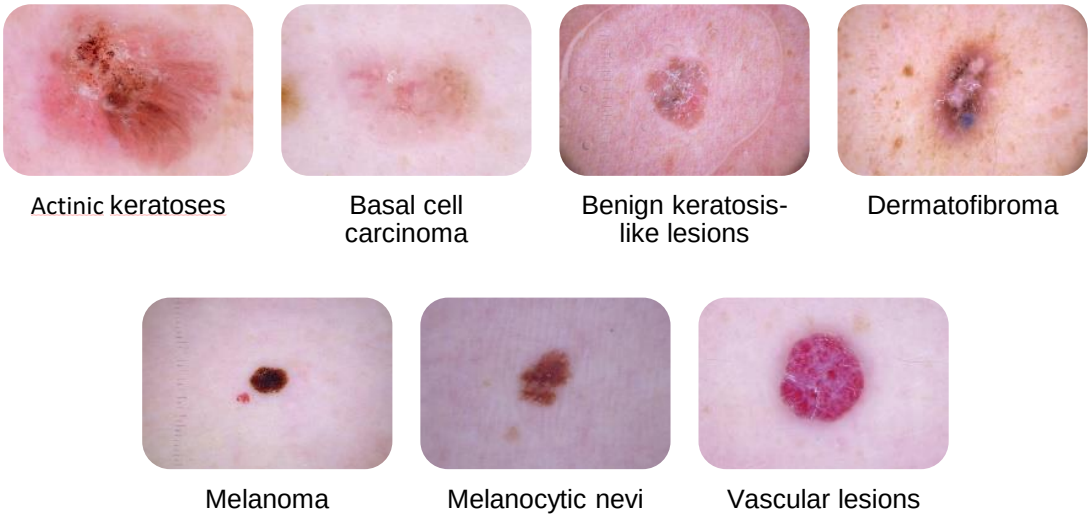

Figure 1. The seven considered lesion types.

We would like to proposed a multi-classification model which is able to identify the type of skin lesion, considering the 7 available types and using the 10,015 dermatoscopic images.

APPROACH'S DESCRIPTION

In the following sections, our processes are detailed explained. After an exploration of the metadata information and the collection of the images, we explored the common pre-processing image's techniques. We took into account different architectures of Convolutional Neural Networks.

▪ Metadata's exploration

It is possible to notice that the target distribution is unbalanced (Table 1).

Lesion's Type	Percentage	Count
Melanocytic nevi	66.950%	6705
Melanoma	11.113%	1113
Benign keratosis-like lesions	10.974%	1099
Basal cell carcinoma	5.132%	514
Actinic keratoses	3.265%	327
Vascular lesions	1.418%	142
Dermatofibroma	1.148%	115

Table 1 Classes' distribution

There are other patients' information: the method through the classification is done (histopathology, follow-up examination, expert consensus or confirmation by in-vivo confocal microscopy), the age, the sex and the localization of the lesion. There are more images for the same lesion; the unique lesions are 7,470 and the duplicates are not due to the different method of classification.

In our models, we didn't take in account these additional information but only the dermatoscopic images.

▪ Extraction and Data Partition

To automatize the extraction of the images from the zip file, we unzipped them with for loop and store each image name.

We automatized the unzip-phase and we organized the image ID with the correspondent target class.

We randomly shuffle the images to not preserve any particular order.

To measure the generalization ability of the machine learning algorithms we partitioned the image in training, validation and test sets. In order to do so, we used two different methods.

In the first phase, the method of partitioning took in consideration the target's distribution and the training, validation and test proportions are respectively 60%,20% and 20%.

In the second phase, the method took in consideration the lesion ID: in fact, because there are more images for the same lesion (as said, there are 7,470 unique ID and 10,015 images), we considered to not have the same lesion in the training and in the validation or in the test.

We identified the images that have duplicated lesion ID and the ones that don't have (Table 2): 5,514 images have no duplicates and 4,501 have duplicates or more than duplicates.

	Number of lesions
Unique	5514
Duplicated	1423
Triplicated	490
Quadrupled	34
Pentaplicated	5
Sextuplicated	4
	7470

Table 2 Unique and duplicated lesions

We used in the validation and in the test all the lesion ID that are unique and in the training the remaining unique ones and the duplicates or more than duplicates and we obtained the training, validation and test proportions as 78%,11% and 11% considering the target distribution of the unique lesion ID (smoothly different from the original target distribution but acceptable).

- **Images Pre-Processing**

Using "ImageDataGenerator", we considered rescale, rotation, zoom, shifts and slips and to better improve the memory's performance we generated batches of augmented data for the sets.

We reduced the original size of the images 600x450: considering 1/3 or 1/6 to faster experiment the architectures. Also, we considered 224x224, because the square size and the dimension multiple of two could help in the performances of the models.

- **Models**

Convolutional neural networks (CNN) are often used in images' classification because it is possible to give as input a n-dimensional space and they are able to received images without a detailed images' pre-processing.

We experiments different architectures considering the two different methods of partitioning: the first phase of architectures considers the partition 60%-20%-20% and the second phase considers the partition 78%-11%-11%.

We considered the most common CNN in images' classification prediction at the end, we used a pre-trained CNN with some additional trained layers.

The challenges were the running time, the vastity of parameters to define together and the unbalanced dataset.

In all the models, we used Adam optimizer, categorical cross-entropy as loss functions and we took into considerations various measures such as mean of the classes' precision and recall, f-score and categorical accuracy. We also declared to save the best (in terms of categorical accuracy or in terms of f-score) model, to reduce the learning rate and to early stop considering the change of categorical accuracy or f-score with callbacks.

1. Architectures – First phase

- **Using the pre-trained InceptionV3**

We reused a previously trained model architecture and then we use a standard training for the last layers, with non-trained parameters: we froze the weights and biases of all the pre-trained model. The additional model is composed by a layer of 128 neurons, a layer of 52 neurons and the 7 output neurons (code: PreTrainedInceptionV3.ipynb).

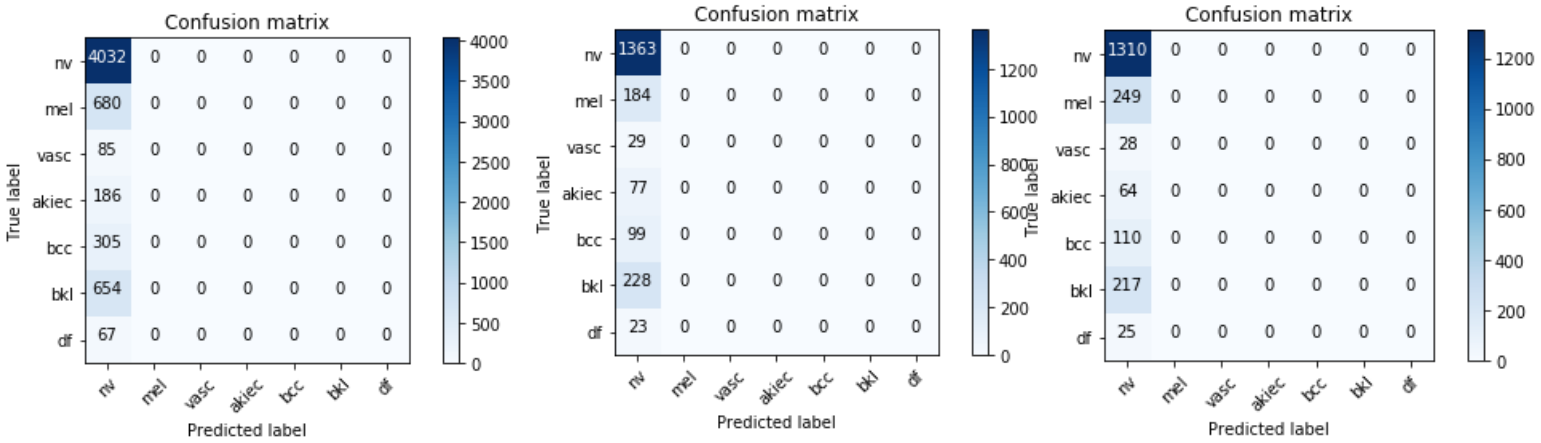

Figure 2 Training, Validation and Test results.

In Figure 2, the results are shown: also in the training, only the majority class is detected. It is possible to notice that the model has 100% precision and 65.40% recall for the majority class and 0% precision and recall for the remaining classes.

Because of this classification, from now on, we experimented the architectures sometimes (it will be specified when and in which sets) applying an oversampling technique: creating new images for the minority classes with data augmentation (random rotation, width and height shift, shear range, zoom, horizontal and vertical flip).

- **Using the pre-trained ResNet**

We used another type of pre-trained model architecture, the ResNet. We added a layer of 2048 neurons, one of 1024 neurons and the last one of 7 neurons. Because it is suggested, we also applied batch normalizations. We didn't freeze any layers and we trained all the architecture (code: PreTrainedResNet.ipynb).

In Figure 3 and in Table 3, there are the results.

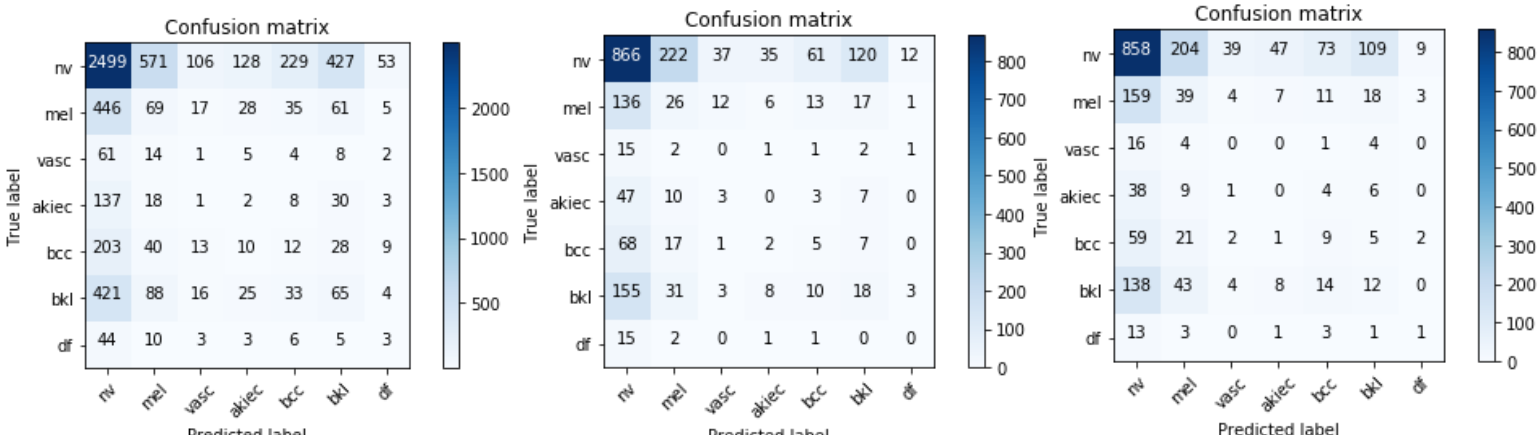

Figure 3 Training, Validation and Test results.

Skin Lesion Type	Class Percentage	Precision	Recall	F1-score
Melanocytic nevi (nv)	66.95%	64.08%	66.98%	65.50%
Melanoma (mel)	11.11%	16.18%	12.07%	13.83%
Vascular lesions (vasc)	10.97%	0.00%	0.00%	0.00%
Actinic keratoses (akiec)	5.13%	0.00%	0.00%	0.00%
Basal cell carcinoma (bcc)	3.27%	9.09%	7.83%	8.41%
Benign keratosis-like lesions (bkl)	1.42%	5.48%	7.74%	6.42%
Dermatofibroma (df)	1.15%	4.55%	6.67%	5.41%

Table 3 Summary of the model's results (test set)

The model is not able to classify two classes, but they are not the minor ones: dermatofibroma lesions, the minorities, are classified even if with low performance.

- **Data Augmentation's Oversampling and home-made GoogleNet**

In this case, we used the over-sampling technique in the training and in the validation sets.

The CNN is a version of the Google Net (code: Oversampling_GoogleNet.ipynb).

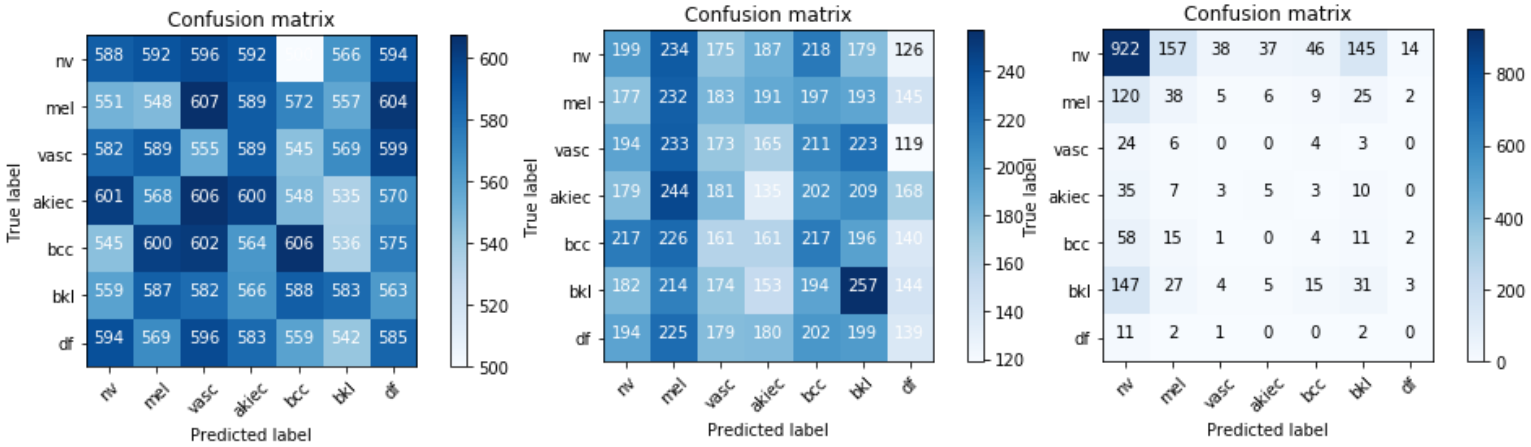

Figure 4 Training, Validation and Test results.

In Figure 4, contingency table of the classification in the training, validation and test sets are shown. In Table 4, precision, recall and F1-score are calculated. The model is not able to classify the two biggest minority classes.

Skin Lesion Type	Class Percentage	Precision	Recall	F1-score
Melanocytic nevi (nv)	66.95%	67.84%	70.01%	68.91%
Melanoma (mel)	11.11%	18.54%	15.08%	16.63%
Vascular lesions (vasc)	10.97%	0.00%	0.00%	0.00%
Actinic keratoses (akiec)	5.13%	7.94%	9.43%	8.62%
Basal cell carcinoma (bcc)	3.27%	4.40%	4.94%	4.65%
Benign keratosis-like lesions (bkl)	1.42%	13.36%	13.66%	13.51%
Dermatofibroma (df)	1.15%	0.00%	0.00%	0.00%

Table 4 Summary of the model's results (test set)

2. Architectures – Second phase

We used over-sampling technique only in the training during all the second phase (partition 78%-11%-11%).

- **Data Augmentation's Oversampling and using pre-trained MobileNet**

Using MobileNet, we canceled the last 5 layers, we froze the last 20 layers and we trained the first remaining 70 layers (code: 2phase_PreTrainedMobileNet.ipynb).

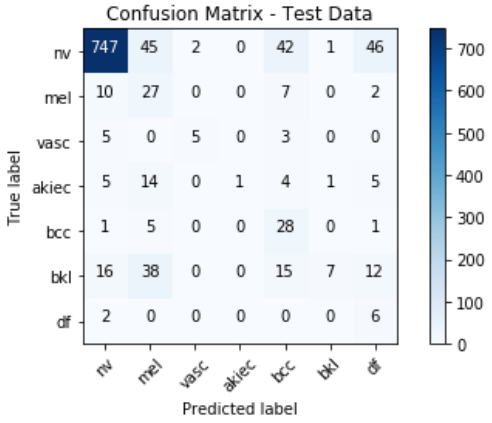

Figure 5 Test results

Skin Lesion Type	Original Class Percentage	Class Percentage of Unique Lesions	Precision	Recall	F1-score
Melanocytic nevi (nv)	66.95%	80.07%	95%	85%	90%
Melanoma (mel)	11.11%	4.17%	21%	59%	31%
Vascular lesions (vasc)	10.97%	7.98%	78%	8%	14%
Actinic keratoses (akiec)	5.13%	3.17%	28%	80%	42%
Basal cell carcinoma (bcc)	3.27%	2.74%	100%	3%	6%
Benign keratosis-like lesions (bkl)	1.42%	1.16%	71%	38%	50%
Dermatofibroma (df)	1.15%	0.71%	8%	75%	15%

Table 5 Summary of the model's results (test set)

Figure 5 and Table 5 show the results of the test set.

- **Data Augmentation's Oversampling and home-made GoogleNet**

Using the same home-made architecture Google-Net, we tried to compile again in this case (code: 2phase_GoogleNet).

Figure 6 and Table 6 show the results of the test set: comparing with Table 4, the results are better in the second phase. This can due for different reasons such as the different size of images considered in the two cases (150x200 and 224x224) and for different parameters in the callbacks (considering the value of categorical accuracy and the F-score).

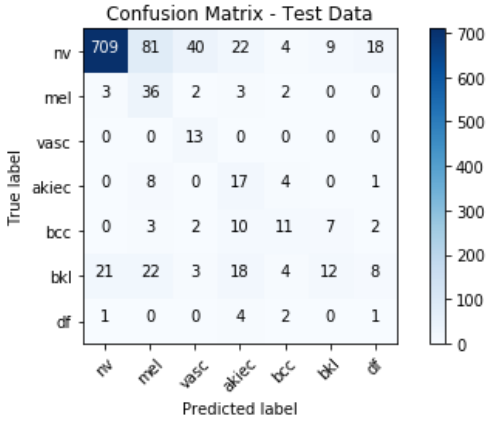

Figure 6 Test results

Skin Lesion Type	Original Class Percentage	Class Percentage of Unique Lesions	Precision	Recall	F1-score
Melanocytic nevi (nv)	66.95%	80.07%	97%	80%	88%
Melanoma (mel)	11.11%	4.17%	24%	78%	37%
Vascular lesions (vasc)	10.97%	7.98%	43%	14%	21%
Actinic keratoses (akiec)	5.13%	3.17%	41%	31%	35%
Basal cell carcinoma (bcc)	3.27%	2.74%	23%	57%	33%
Benign keratosis-like lesions (bkl)	1.42%	1.16%	22%	100%	36%
Dermatofibroma (df)	1.15%	0.71%	3%	12%	5%

Table 6 Summary of the model's results (test set)

CONCLUSIONS

Being able to identify the lesion from the first stage using dermascopic images can fasten the diagnosis process and consequently follow the appropriate treatment. In addition to that, it does not require a sample of the skin tissue and it is a cheaper diagnosis option. Although it is a challenging task due to the different appearances of the same skin lesion and therefore hard to predict.

We tried different CNN architectures but because of the complexity of the problem, the proposed predicted models are not able to achieve the classification task with acceptable performance. The best model in terms of the F1-score of the minority class is the MobileNet without the last 5 layers, with the last 20 layers frozen and the remaining 70 trained (Table 5) of 15% of F1-score of the dermatofibroma class.

The challenges were the running time and the difficulty to be able to classify correctly the minor classes.

Better image pre-processing, such using a background identification model or filtering the image's noises, could make the prediction improves. The tuning of the parameters could be explored, even if very slow.

Another partitioning method could be considered: in the second phase, instead of using the stratification considering the target of the unique lesions, create a data partition that consider the original classes' percentages.

Also, a lot of ways of using pre-trained models could be explored: where to freeze the layers and how many frozen layers to consider. Other pre-trained models could be tested.

REFERENCES

Cancer Facts and Figures 2019. (2019). *American Cancer Society*.

M. Binder, A. S. (1994). Application of an artificial neural network in epiluminescence microscopy pattern analysis of pigmented skin lesions: a pilot study. *British Journal of Dermatology*.

Tschandl, P. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Harvard Dataverse*.