# Distinguish Melanoma from Benign Melanocytic and Non-Melanocytic Lesions

Mónica Andreia Teixeira do Amaral
Technical University of Lisbon
Instituto Superior Técnico
Electrical and Computer Engineering
Lisbon, Portugal
Email: monica.amaral@tecnico.ulisboa.pt

Professor Jorge dos Santos S. Marques
Technical University of Lisbon
Instituto Superior Técnico
Lisbon, Portugal
Email: jsm@isr.ist.utl.pt

Dr. Ana Catarina Fidalgo Barata
Technical University of Lisbon
Instituto Superior Técnico
Lisbon, Portugal
Email: ana.c.fidalgo.barata@ist.utl.pt

***Abstract*—Melanoma is a very aggressive type of cancer. Consequently, the detection of melanoma at early stages is crucial to prevent it from metastasizing, increasing survival rate. However, distinguishing between melanoma and other types of lesions, especially non-melanocytic ones, remains a difficult task. Dermoscopy is a procedure used by physicians for skin cancer diagnosis, that has shown to be more accurate than naked-eye analysis, since it can magnify the size of the inspected lesions by up to 100x. Several medical procedures are used to differentiate lesions, mostly for melanoma diagnosis, such as ABCD rule or Menzies scoring method. Several Computer-Aided Diagnosis systems have been developed in the interest of helping physicians better detect melanomas. Nonetheless, most of CAD systems aren't able to distinguish between melanocytic and non-melanocytic lesions, which is still an open problem and poorly addressed in literature. This thesis proposes a deep neural network to automatically discriminate melanoma from melanocytic (benign nevi) and non-melanocytic (seborrheic keratosis) lesions. To achieve this goal several regularization techniques were investigated in order to improve the performance of the developed model. The best evaluated configuration achieved a Balanced Accuracy ($BACC$) of $59,9\%$ for the validation dataset and $53,1\%$ for the test dataset. Furthermore, the best two models at correctly detecting melanomas, reached a same Sensitivity ($SE$) of $56,7\%$ in validation dataset and $SE = 32,5\%$ and $SE = 49,6\%$ for test dataset. The dataset used in this thesis was collected from the 2017 ISIC Challenge database.**

***Index Terms*—Skin Lesion Diagnosis, Dermoscopy, Melanoma, Non-melanocytic Lesions, Deep Neural Networks, Regularization Techniques.**

## I. Introduction

Skin cancer is the most common type of cancer in the world [4]. The incidence rates of skin cancer have been increasing for the last decades and so have the death rates associated with it. Skin cancer can be divided into two different categories: melanoma and non-melanoma. The latter is the most common type of skin cancer and has a higher chance of cure [17]. Despite melanoma accounting for about 2% of all skin cancers, it is one of the deadliest forms of cancer. According to the cancer research UK, melanoma ranks the ninth position as the most common cancer in Europe and the nineteenth place in worldwide [5]. The American Cancer Society estimates that for 2019, in the United States alone, more than 96,000 new cases of melanoma will be diagnosed and more than 7200 people are expected to die of this form of cancer [4].

There has been increasingly challenges to improve the classification of skin lesions raised in the research community, and, consequently, diagnosis of melanoma, as well as an attempt to a more accurate and early detection [13], preventing it to metastasize.

Furthermore, several Computer-Aided Diagnosis (CAD) systems have been developed in order to help physicians with the diagnosis. However, existing CAD systems have several limitations. Most of the CAD systems are not able to distinguish between melanocytic or non-melanocytic skin lesions. Most of these systems only allow dermatologists to classify the lesion as benign melanocytic or malignant melanoma. Both of them have been the main focus of most of the research community, which is trying to develop new methods and algorithms in order to perform the distinction between these lesions. As result, non-melanocytic lesions have often been disregarded. In spite of the efforts that have been done, automatic detection of melanoma considering both melanocytic and non-melanocytic lesions remains barely addressed in the literature and is a less explored field, where there are still many questions and challenges. Therefore, it is important to develop strategies to deal with the distinction of melanoma from melanocytic and non-melanocytic lesions.

By developing a deep neural network, and by using deep learning techniques, this thesis aims to provide a new strategy for the discrimination of melanoma from melanocytic and non-melanocytic lesions.

The document is organized as follows. Section II presents state of the art. Additionally, the limitations of deep neural networks are briefly discussed. Section III describes the strategy adopted in this thesis to automatically discriminate melanoma from benign melanocytic and non-melanocytic lesions. In particular, it presents the steps of the system architecture for this thesis' classification problem and showcases the developed deep learning network and regularization techniques used considering the main goal to automatically discriminate melanoma from benign melanocytic and non-melanocytic lesions. This thesis was developed using an available public database. The experimental results are discussed in Section IV and conclusions are demonstrated in Section V.

## II. State of the Art

Melanomas are not easily discriminated from other types of skin lesions. Dermatologists may be able to see the differences of these lesions based on a physical examination but sometimes it is necessary to perform a histological exam, which is expensive and may lead to psycho-emotional consequences for patients [2], especially due to a permanent scar. In order to avoid the histological exam, dermatologists use a noninvasive technique as a second opinion tool. Dermoscopy is a procedure to diagnose skin lesions and detect melanomas, that allows the in-vivo observation and evaluation of skin lesions [6].

Furthermore, there are several medical procedures that can be used, such as Pattern Analysis, ABCD rule, Menzies scoring method, and 7-point checklist in order to differentiate lesions. Nonetheless, even when using these medical algorithms, the detection of melanoma is not trivial, even for experienced dermatologists [3] [11]. The use of automatic systems has the potential to improve dermatologists' performance in discriminating between benign and malignant lesions and it can be used as a second opinion, either by both experienced and non-experienced dermatologists.

Several predictive models have been proposed for the diagnostic of skin cancer, differentiating between benign melanocytic and melanoma lesions. However, there are few models that try to distinguish between melanocytic and non-melanocytic lesions in an automatic way. Furthermore, in case of doubt, the latter are commonly generalized and treated as melanocytic. Since malignant lesions in both classes have different visual features, they cannot be addressed in the same way. The models for melanoma detection are built based on machine learning methods.

Traditionally, in CAD systems, the features of dermoscopic images are extracted prior to the classification algorithm. Alternatively, the end-to-end learning approach makes direct use of images as a classifier input, skipping feature extraction step, since the learning algorithm can learn the features by itself. End-to-end learning allows training the system as a whole, so there is no separation between feature extraction and classification. The difference between both approaches can be seen in Figure 1.

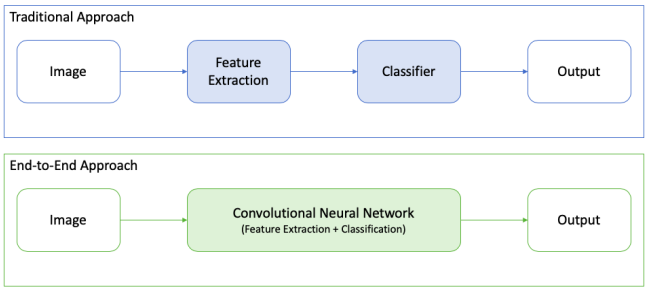


Fig. 1. Difference between traditional and end-to-end approach.

Deep learning performs an "end-to-end" approach and its techniques have obtained excellent performance in different classification tasks, achieving human-level accuracy. Thus, there have been efforts to also take advantage of these algorithms and apply them to medical imaging tasks [9]. One example is the usage of deep learning techniques to automatically detect melanoma in dermoscopy analysis [7] [10] [14] [16].

### A. Limitations of Deep Neural Networks

*1) Reducing Overfitting:* When a model is tuned to a small set of training patterns, it may not be able to perform well for the validation and test sets. Thus, the model has a good fit on the training set but a poor performance in new examples, which means that the model is overfitting the training set. By using label-preserving transformations, it is possible to reduce overfitting [1]. The easiest and most common method is called dataset augmentation, which aims at increasing the size and diversity of training data (sometimes also the validation data).

Dropout is a different technique applied during the training phase of the model to reduce overfitting. It refers to removing units from the network for a limited time, simultaneously with all its incoming and outgoing connections. As in each training iteration there are different sets of units dropped and a another set of units considered, it results in a different set of outputs, forcing a more powerful and robust learning. [1]

*2) Exploding Gradients:* Another important issue that should be addressed is that large learning rates may result in the exploding gradients or vanishing gradients. Batch normalization is a regularization technique which helps to address these problems and reduce overfitting, due to its regularization effect. Batch normalization allows the model to converge much faster in training since that the higher learning rates may be used [18], and by reducing the quantity of required training epochs it consequently allows the model to better generalize.

*3) Unbalanced Data:* Unbalanced dataset occurs when most of the data of the training dataset belongs to a particular class. Class weighting intends to balance the weights of each class to ensure that all classes contribute equally to the cost function. This is achieved by associating lower weights to the training patterns of the majority class and higher weights to the minority class. The weights are usually defined as being inversely proportional to the class frequency in the dataset.

## III. Proposed Architecture

Melanocytic/non-melanocytic discrimination problem will be addressed in this document applying deep learning techniques that enable computers to learn from labelled data [12]. Deep learning has been evolving and showing promising results in various applications. It has also been introduced to the dermoscopy image analysis field, and progresses have been made using deep learning for melanoma detection, which it is extremely helpful to address the problem already stated [16] [20].

For the task of distinguishing melanoma from benign melanocytic and non-melanocytic lesions, a deep neural network was developed.

The dataset with skin lesions used in this thesis is the one adopted in the 2017 International Skin Imaging Collaboration (ISIC) Challenge (Phase 3 - lesion classification),

which contains 2750 images, where 2000 are for training purposes, 150 for validation and 600 for testing. The dataset includes medical annotations associated to each image that was considered as ground truth [13]. The labels adopted are: melanoma (malignant melanocytic lesion), benign nevi (benign melanocytic lesion) or seborrheic keratosis (benign non-melanocytic lesion). Although melanocytic lesions are well represented in the ISIC database, non-melanocytic lesions are not completely represented, as only seborrheic keratosis are considered.

Table I shows the number of images for each of the training, validation and test sets, split by the three classes of skin lesions (melanomas, benign nevi and seborrheic keratosis).

TABLE I
CONTAINING IMAGES IN 2017 ISIC DATASET.

| Dataset | Melanoma | Seborrheic Keratosis | Benign Nevi | Total |
|---|---|---|---|---|
| **Train** | 374 | 254 | 1372 | 2000 |
| **Validation** | 30 | 42 | 78 | 150 |
| **Test** | 117 | 90 | 393 | 600 |

### A. System Architecture

The system architecture considered in this thesis involves four main stages: image acquisition, lesion segmentation, pre-processing and classification model, as shown in Figure 2.

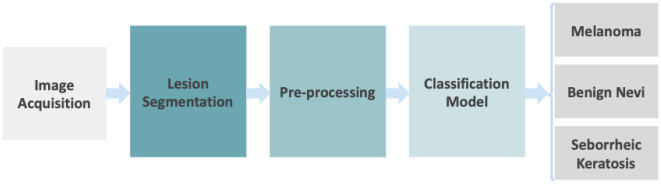


Fig. 2. An illustration of system architecture major steps.

The first stage refers to obtaining the necessary datasets of dermatoscopy images. These datasets are assumed to be available as well as the medical annotations: benign nevi, melanoma or seborrheic keratosis.

The next major phase is lesion segmentation. The objective is to separate the region of interest (lesion) from the healthy skin. In this thesis, segmentation masks are assumed available and created by experts [13]. Thus, segmentation masks were used to discover the coordinates of the bounding box containing the lesion. Once the coordinates were found, the original images were cropped in order to obtain images containing only the lesion (see Figure 3 for example). The latter were further resized to a common size of 256x256. Following this phase, the image was normalized in image pre-processing step (using $image = image \times 1.0/127.5 - 1.0$).

The classification model step is based on the best designed deep neural network architecture and allows the classification of the lesions as: benign nevi, seborreich keratosis, and melanoma.

Fig. 3. An example of an image and its respective segmentation mask [13], as well as the resulting image after "Lesion Segmentation" process has been applied in this thesis.

### B. Proposed Deep Neural Network

An end-to-end architecture of neural network was built, based on deep neural networks, which will be named as baseline architecture from now on, since it was the baseline for the experiences presented in this document. This model is based both on AlexNet and VGG16 architectures. The choice of the deep convolutional networks that support this work is, essentially, due to their outstanding performances in ImageNet Large Scale Visual Recognition Challenge (ILSVRC), namely the first was a breakthrough in image classification [1] and the second was a configuration improvement capable of achieving even better results [15].

The baseline architecture has five convolutional layers and two fully-connected layers, as shown in Figure 4. All the convolutional layers are followed by max-pooling layers. The ReLU as activation function is applied to the output of every convolutional layer and also to the hidden fully-connected layer, which has 1024 hidden units. The last fully-connected layer, since is the output layer with class prediction, has three neurons one for each class (melanoma, benign nevi and seborrheic keratosis) and the activation function was set to be the softmax.

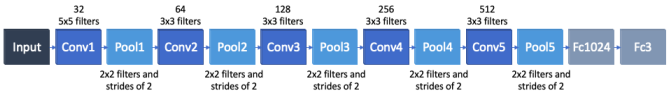


Fig. 4. An illustration of baseline architecture.

Input dimensions for each layer as well as information about non-linearity and which layer is followed by a max-pooling layer, can be consulted in Table II. More details about the baseline network will be discussed below.

The first layer, a convolutional one, applies 32 filters with kernel size $5\times5\times3$ to the input images dimensions, and is followed by a pooling layer with a $2\times2$ max-pooling filter and stride of 2. The pooling specifications of the other convolutional layers are the same.

The second convolutional layer, also followed by a max-pooling, takes as input the output of the first max-pooling layer and applies 64 filters with kernel size $3\times3\times32$. The third and fourth convolutional layers, follow the second max-pooling layer, and have the same kernel size as the second convolutional one, with respectively 128 and 256 filters. Both are followed by max-pooling layers.

TABLE II
LAYERS' DETAILS (NA STANDS FOR NOT APPLICABLE).

| Layer | Input Dimensions | Layer Type | Specificity | Non-Linearity |
|---|---|---|---|---|
| 1 | 256×256×3 | Convolutional | 32 filters 5×5×3 | ReLU |
| 1 | 252×252×32 | Max-Pooling | 2×2×1 filters and stride 2 | NA |
| 2 | 126×126×32 | Convolutional | 64 filters 3×3×32 | ReLU |
| 2 | 124×124×64 | Max-Pooling | 2×2×1 filters and stride 2 | NA |
| 3 | 62×62×64 | Convolutional | 128 filters 3×3×64 | ReLU |
| 3 | 60×60×128 | Max-Pooling | 2×2×1 filters and stride 2 | NA |
| 4 | 30×30×128 | Convolutional | 256 filters 3×3×128 | ReLU |
| 4 | 28×28×256 | Max-Pooling | 2×2×1 filters and stride 2 | NA |
| 5 | 14×14×256 | Convolutional | 512 filters 3×3×256 | ReLU |
| 5 | 12×12×512 | Max-Pooling | 2×2×1 filters and stride 2 | NA |
| 6 | 6×6×512 | Fully-connected | 1024 units | ReLU |
| 7 | 18432 | Fully-connected | 3 units | Softmax |

Finally, the fourth convolutional and max-pooling layers are followed by another one with 512 filters and a kernel size of 3×3×256. The max-pooling is applied after, and the output is transformed into a vector for the fully-connected layer. As said before, this layer has 1024 hidden units.

The final layer of the network, as said before, is a fully-connected layer and has three neurons, one for each target class.

*1) Training:* In order to train the model, the choice of the loss function and optimizer were required.

The loss function used is the categorical cross-entropy, which measures the difference between the distribution of the training data and the model.

The chosen optimizer was the Adam Optimizer algorithm with learning rate equal to 0.001. This algorithm is a variant of the stochastic gradient descent, among several benefits requiring little memory and being computationally efficient [8]. Several tests were made and this learning rate showed a better performance. This model was trained using mini-batches with a batch size of 25.

The equations of this algorithm are the following [8]. Equation 1 indicates the gradient at timestep $t$.

$$g_t = \nabla f_t(\theta_{t-1}) \tag{1}$$

The moving averages of the gradient ($m_t$) and the squared gradient ($v_t$) are computed as follow (see equations 2 and 3). Both $m_t$ and $v_t$ are estimates, respectively, of the the mean (1st moment) and the uncentered variance (2nd raw moment) of the gradient. The $\beta_1$ and $\beta_2$ are the hyper-parameters within $[0, 1]$ that control the exponential decay rates.

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{2}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \tag{3}$$

The $\hat{m}_t$ and $\hat{v}_t$ are bias-corrected first and second moment estimates, and are computed as in the equations 4 and 5.

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} \tag{4}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t} \tag{5}$$

The bias-corrected first and second moments are then used to update the parameters, returning the Adam's Update Rule (see equation 6).

$$\theta_t = \theta_{t-1} - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{6}$$

$$\alpha_t = \alpha \frac{\sqrt{1 - \beta_2^t}}{1 - \beta_1^t} \tag{7}$$

### C. Regularization Techniques

For the purpose of training and testing the baseline model and obtaining the best performance, there are some improvements which could allow the model to generalize better and, consequently, prevent overfitting. These improvements are frequently called as regularization techniques.

Several experimental tests were performed in order to evaluate the impact of the regularization techniques in the performance of the model.

Table III show the different model configurations used in each experiment, considering the applied regularization techniques. In all of them a training set was used to learn the model parameters and an independent validation set was used for evaluation.

Initially, the performance of the baseline architecture was evaluated, in order to have a base of comparison. The ensuing test was performed considering a dropout of 0.50, with the objective of preventing overfitting (Experiment 1 in Table III). A different test was performed in order to evaluate batch normalization as a technique of regularization: batch normalization layers were added after all the convolutional layers and right before the max-pooling layer (Experiment 2 in Table III). Since the training set is unbalanced, the class weighting technique was applied to balance the weights of each class (Experiment 3 in Table III). Finally, this thesis intended to evaluate the model performance using data augmentation, once again in an attempt to reduce overfitting (Experiment 4 in Table III). For this purpose, data augmentation online was considered, since the transformations were performed on the

TABLE III
MAPPING OF EXPERIMENTS PERFORMED WITH DIFFERENT MODEL CONFIGURATIONS, REGULARIZATION TECHNIQUES AND DATASET. (**DO** STANDS FOR DROPOUT, **BN** STANDS FOR BATCH NORMALIZATION, **CW** STANDS FOR CLASS WEIGHTING AND **DA** STANDS FOR DATA AUGMENTATION).

| Experience ID | Model Specifications | | | |
|---|---|---|---|---|
| | **DO** | **BN** | **CW** | **DA** |
| **Baseline Model** | | | | |
| **Exp. 1** | X | | | |
| **Exp. 2** | | X | | |
| **Exp. 3** | | | X | |
| **Exp. 4** | | | | X |
| **Exp. 5** | X | X | | |
| **Exp. 6** | X | | X | |
| **Exp. 6** | X | | | X |
| **Exp. 8** | | X | X | |
| **Exp. 9** | | X | | X |
| **Exp. 10** | | | X | X |
| **Exp. 11** | X | X | X | |
| **Exp. 12** | X | X | | X |
| **Exp. 13** | X | | X | X |
| **Exp. 14** | | X | X | X |
| **Exp. 15** | X | X | X | X |

mini-batches to be fed to the model, immediately prior to the training. The augmentation techniques used were [19]:

- **rotation**: images were rotated by 90 degrees, 4 times (performing a rotation of 0, 90, 180 and 270 degrees);
- **flip**: images were flipped randomly along X axis (horizontally: left to right) and along Y axis (vertically: upside down);
- **color augmentation**: images were adjusted according to its hue and saturation of RGB, brightness and contrast by a random factor.

All the following experiments after Experiment 4 (see in Table III) were carried out considering a combination of these regularization techniques.

## IV. EXPERIMENTAL RESULTS

In this section is discussed the experimental results of the evaluations carried out to the deep neural network configurations (see Table III for reference to the different experiments performed). The baseline model and the combinations of the regularization techniques discussed in Section III-C, were tested against the validation dataset.

In order to evaluate the classifier and see how well it performs, a set of metrics was adopted. These metrics were Accuracy ($ACC$), Balanced Accuracy ($BACC$), Loss, Sensitivity ($SE$) and Specificity ($SP$). The obtained results for the chosen metrics can be seen in Table IV.

The most relevant evaluation metric is balanced accuracy, $BACC$, since the validation dataset is unbalanced. By inspecting this table, it is possible to see that $BACC$ values range from $46,3\%$ to $59,9\%$. The $59,9\%$ value corresponds to the model that was trained using all of the possible regularization strategies. The applied techniques were dropout, class weighting, batch normalization and data augmentation, as can been seen in Experiment 15 (Table IV). Compared to the other models, it can be concluded that this model generalized better given the application of all these techniques. Also, the overall accuracy for this model, not considering the fact that the validation dataset is unbalanced, is $69,3\%$ outperforming all other experiments. In addition, the loss also has the lowest value compared to the other models, which confirms that this model is the one that achieves the best performance for the validation set.

Further, by analyzing the models when a single regularization technique is applied, it is easy to see that the model with best performance is the one that uses data augmentation (Experiment 4, with a $BACC = 57\%$) compared to batch normalization (Experiment 2, with a $BACC = 53,7\%$), class weighting (Experiment 3, with a $BACC = 52,3\%$) or dropout (Experiment 1, with a $BACC = 47,9\%$) techniques. However, when comparing these models through other evaluation metrics ($ACC$ and $LOSS$), also presented in Table IV, it can be concluded that the best performing model is the one that only considers batch normalization as a regularization technique (with $ACC = 65,3\%$ and $LOSS = 0,916$), against Experiment 1 ($ACC = 59,3\%$ and $LOSS = 0,947$), Experiment 3 ($ACC = 60,0\%$ and $LOSS = 0,973$) and Experiment 4 (with $ACC = 62,0\%$ and $LOSS = 0,934$).

The model with only data augmentation, besides its $BACC$ outperforming the models with a single regularization technique, also shows a better performance than some of those that used two or three techniques (ranging from $0,3\%$ to $5,8\%$ for $BACC$ value). The only exception is when combining every considered regularization technique, achieving a $BACC = 59,9\%$. This may suggest that data augmentation by itself can achieve a similar or better regularization effect, for instance in reducing overfitting, compared to others techniques or the combination between them.

Nevertheless, combining two or three techniques into the evaluated models tends to improve the value of $BACC$. For instance, the model in Experiment 13 (with dropout, class weighting and data augmentation) shows a slight improvement in performance ($BACC = 56,7\%$) than the models in Experiment 6 (dropout and class weighting) with a $BACC = 56,5\%$, Experiment 7 (dropout and data augmentation) with $BACC = 55,1\%$, Experiment 8 (batch normalization and class weighting) with $BACC = 56,4\%$, Experiment 9 (batch normalization and data augmentation) with a $BACC = 55,7\%$ or Experiment 10 (class weighting and data augmentation) with $BACC = 55,5\%$ . It also outperforms Experiment 5 (dropout and batch normalization) with a $BACC = 51,2\%$.

By looking at the $BACC$ of the Experiment 13 and comparing it with the $BACC$ values for the models that use two of the three regularization techniques in this configuration (dropout, class weighting and data augmentation), the model

TABLE IV
EVALUATION METRICS FOR THE DIFFERENT MODEL CONFIGURATIONS, FOR VALIDATION DATASET.
(**DO** STANDS FOR DROPOUT, **BN** STANDS FOR BATCH NORMALIZATION, **CW** STANDS FOR CLASS WEIGHTING AND **DA** STANDS FOR DATA AUGMENTATION)

| Experiment ID | DO | BN | CW | DA | ACC | BACC | LOSS | $SE_0$ | $SE_1$ | $SE_2$ | $SP_0$ | $SP_1$ | $SP_2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Baseline Model | | | | | 56,0% | 46,3% | 0,984 | 76,9% | 16,7% | 45,2% | 52,8% | 83,3% | 88,9% |
| Experiment 1 | X | | | | 59,3% | 47,9% | 0,947 | 79,5% | 16,7% | 47,6% | 58,3% | 90,0% | 80,6% |
| Experiment 2 | | X | | | 65,3% | 53,7% | 0,916 | 91,0% | 20,0% | 50,0% | 50,0% | 94,1% | 91,7% |
| Experiment 3 | | | X | | 60,0% | 52,3% | 0,973 | 62,2% | 6,7% | 81,0% | 73,6% | 92,5% | 70,4% |
| Experiment 4 | | | | X | 62,0% | 57,0% | 0,934 | 74,4% | 46,7% | 50,0% | 68,1% | 89,2% | 80,6% |
| Experiment 5 | X | X | | | 63,3% | 51,2% | 0,899 | 88,8% | 20,0% | 45,2% | 47,2% | 91,7% | 92,6% |
| Experiment 6 | X | | X | | 65,3% | 56,5% | 0,949 | 83,3% | 26,7% | 59,5% | 62,5% | 90,8% | 87,0% |
| Experiment 7 | X | | | X | 62,1% | 55,1% | 0,920 | 78,2% | 30,0% | 57,1% | 62,5% | 90,8% | 83,3% |
| Experiment 8 | | X | X | | 67,3% | 56,4% | 0,871 | 89,7% | 20,0% | 59,5% | 59,7% | 93,3% | 81,5% |
| Experiment 9 | | X | | X | 65,0% | 55,7% | 0,908 | 80,8% | 26,7% | 59,5% | 63,9% | 93,3% | 81,5% |
| Experiment 10 | | | X | X | 61,3% | 55,5% | 0,925 | 69,2% | 56,7% | 40,5% | 75,0% | 75,8% | 86,1% |
| Experiment 11 | X | X | X | | 66,7% | 54,6% | 0,897 | 92,3% | 16,7% | 54,8% | 51,4% | 93,3% | 93,5% |
| Experiment 12 | X | X | | X | 64,0% | 53,5% | 0,904 | 87,2% | 23,3% | 50,0% | 59,7% | 90,0% | 88,0% |
| Experiment 13 | X | | X | X | 63,3% | 56,7% | 0,921 | 73,1% | 56,7% | 40,5% | 72,2% | 75,8% | 90,7% |
| Experiment 14 | | X | X | X | 63,3% | 55,1% | 0,923 | 82,1% | 33,3% | 50,0% | 68,1% | 88,3% | 83,3% |
| Experiment 15 | X | X | X | X | 69,3% | 59,9% | 0,858 | 87,2% | 23,3% | 69,9% | 73,6% | 93,3% | 82,4% |

that considers dropout and class weighting (Experiment 6) has only a slight decrease in its performance ($BACC = 56,5\%$), while both Experiment 7 and 10 have a larger decrease (with $BACC = 55,1\%$ and $BACC = 55,5\%$, respectively). This potentially suggests that the combination of dropout and class weighting may achieve a better performance.

When using batch normalization ($BACC = 53,7\%$) there's a better performance versus using dropout alone ($BACC = 47,9\%$). Additionally, when combining these two techniques together, the model seems to have a poorer performance, potentially due to both techniques nullifying the effect of each other. Though, the nullifying effect does not seem to occur when comparing the model with dropout and batch normalization (Experiment 5, $BACC = 51,2\%$) with the one with just dropout (Experiment 1, $BACC = 47,9\%$). However, comparing Experiment 5 with just batch normalization (Experiment 2, $BACC = 53,7\%$), the latter outperforms the first. Furthermore, by looking at the results of Experiment 11 (dropout, batch normalization and class weighting) and Experiment 12 (dropout, batch normalization and data augmentation), respectively with a $BACC = 54,6\%$ and a $BACC = 53,5\%$, these perform worst when compared to Experiments 6, 7, 8 and 9. When comparing both Experiments 11 and 12 with Experiment 5 (dropout and batch normalization), the formers outperform the latter. This was expected since the combination of dropout and batch normalization may nullify the regularization benefits of each other, and the combination of each one with one of the remaining techniques (class weighting or data augmentation) achieves a better performance. Therefore, when considering dropout, batch normalization and another technique, the models outperform the one only considering dropout and batch normalization ($BACC = 51,2\%$). It is easy to conclude that combining dropout and batch normalization with class weighting (Experiment 11) achieves better results than combining it with data augmentation (Experiment 12), potentially due to the combination of dropout and class weighting.

In addiction, Table IV shows the results of sensitivity ($SE$) and the specificity values ($SP$) for each lesion and for the different experiments performed on the validation dataset. As it can be seen, there are three different classes of skin lesions. Each of the identified classes (0, 1 and 2) corresponds to a specific lesion: benign nevi, melanoma, and seborrheic keratosis.

By analyzing this Table, it can be observed that, for the best configuration model (Experiment 15), the $SE$ values for each class vary significantly (ranging from $23,3\%$ to $87,2\%$) and the $SP$ values, although not so considerably, also differ among the three classes (ranging from $73,6\%$ to $93,3\%$). Figure 5 shows the confusion matrix for this classification model, giving an idea of the overall performance.

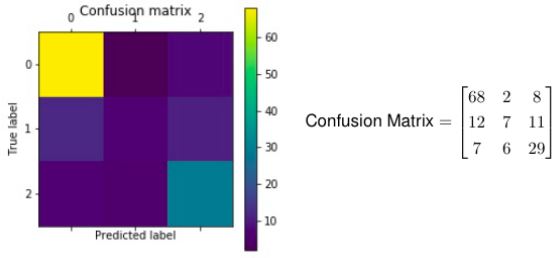


Fig. 5. Confusion Matrix for Experiment 15 (validation dataset).

Comparing the results presented in Table IV for this model (Experiment 15) with those of the confusion matrix (Figure 5), it is fair to conclude that the model is more accurate at detecting benign nevi and seborrheic keratosis, respectively showing a $SE = 87,2\%$ and a $SE = 69,0\%$. However, it only detects $23,3\%$ of melanomas as actually being melanomas, missing the rest. Nevertheless, the model has a $SP = 93,3\%$

for melanoma, which means it correctly identifies $93,3\%$ of the lesions as non-melanomas (i.e., it identifies $93,3\%$ the lesions as being one of the other two considered - benign nevi or seborrheic keratosis). For benign nevi lesions and seborrheic keratosis the model achieves, respectively, a $SP = 73,6\%$ and a $SP = 82,4\%$.

When considering the configurations with only one technique (Experiments 1, 2, 3 or 4), and comparing the results obtained for $SE$ and $SP$ for each lesion, it is possible to verify that the model using only class weighting (Experiment 3) achieves a $SE = 81,0\%$ for seborrheic keratosis lesions increases by around $30\%$ versus the models with dropout (Experiment 1 with a $SE = 47,6\%$), batch normalization (Experiment 2 with $SE = 50,0\%$) or data augmentation (Experiment 4 with $SE = 50,0\%$), considering the same lesion. Given seborrheic keratosis being a minority class, class weighting associates a higher weight leading to a better detection of this lesion. On the other hand, when comparing the same configurations, the model with only batch normalization shows a better performance at correctly detecting benign nevi, which can be observed when comparing Experiment 2 ($SE = 91,0\%$) versus Experiments 1, 3 and 4, respectively with $SE = 79,5\%$, $SE = 69,2\%$ and $SE = 74,4\%$. Furthermore, data augmentation alone outperforms when it comes to correctly identified melanoma, as it can be perceived by looking at Experiment 4 ($SE = 47,6\%$) versus Experiments 1 ($SE = 16,7\%$), 2 ($SE = 20,0\%$) and 3 ($SE = 6,7\%$). Confusion matrices for the configuration models in Experiment 1, 2, 3, and 4, can be seen in Figures 6, 7, 9, and 8, respectively.

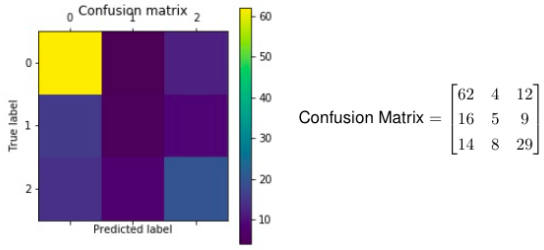


Fig. 6. Confusion Matrix for Experiment 1 (for validation dataset).

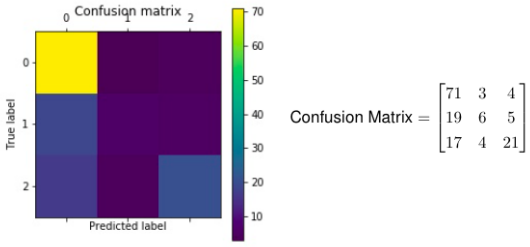


Fig. 7. Confusion Matrix for Experiment 2 (for validation dataset).

When it comes to correctly classifying melanoma, considering two or more techniques, there are two models presenting

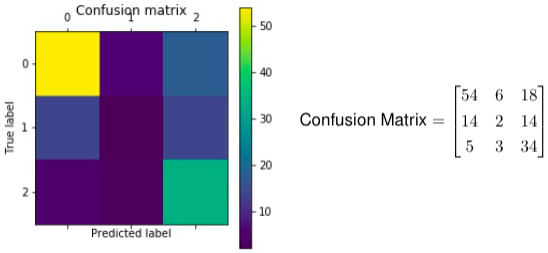


Fig. 8. Confusion Matrix for Experiment 3 (for validation dataset).

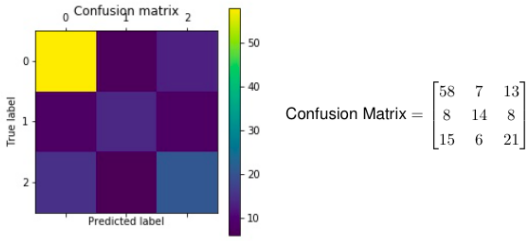


Fig. 9. Confusion Matrix for Experiment 4 (for validation dataset).

the best result. One uses dropout, class weighting and data augmentation (Experiment 13 with $SE = 56,7\%$), and the other considers only class weighting and data augmentation (Experiment 10, with $SE = 56,7\%$), both showing an increase of $10\%$ versus the second best model at detecting this lesion (Experiment 4 with $SE = 46,7\%$), and an increase of $50\%$ concerning the worst performing model (Experiment 3 with $SE = 6,7\%$). Figures 10 and 11 show the confusion matrices for Experiment 10 and Experiment 13, respectively. By analysing the sensitivity for the three lesions for these models it is possible to conclude that adding dropout, in this particularly case, may only influence the sensitivity for benign nevi lesions ($SE = 69,2$ for Experience 10 versus $SE = 73,1\%$ for Experience 13), as the other two sensitivities have the same value for both configurations ($SE = 56,7\%$ for melanoma and $SE = 40,5\%$ for seborrheic keratosis).

Likewise, the model that uses dropout, batch normalization and class weighting (Experiment 11), shows the best results for correctly detecting benign nevi ($SE = 92,3\%$), see Figure 12 for confusion matrix.

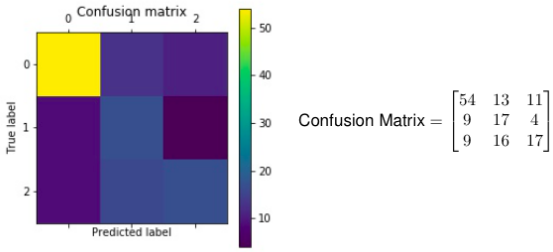


Fig. 10. Confusion Matrix for Experiment 10 (for validation dataset).

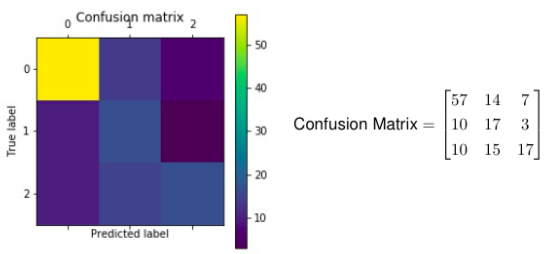


Fig. 11. Confusion Matrix for Experiment 13 (for validation dataset).

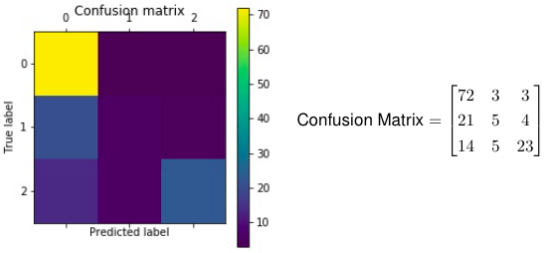


Fig. 12. Confusion Matrix for Experiment 11 (for validation dataset).

Models with batch normalization have a higher result at correctly detecting benign nevi. This can be observed when combining batch normalization and other techniques versus the same configuration without batch normalization:

- Experiment 5 ($SE = 88,5\%$) versus 1 ($SE = 79,5\%$)
- Experiment 8 ($SE = 89,7\%$) versus 3 ($SE = 69,2\%$)
- Experiment 9 ($SE = 80,8\%$) versus 4 ($SE = 74,4\%$)
- Experiment 11 ($SE = 92,3\%$) versus 6 ($SE = 83,3\%$)
- Experiment 12 ($SE = 87,2\%$) versus 7 ($SE = 78,2\%$)
- Experiment 14 ($SE = 82,1\%$) versus 10 ($SE = 69,2\%$)
- Experiment 15 ($SE = 87,2\%$) versus 13 ($SE = 73,1\%$)

The higher result of correctly identifying a lesion as benign nevi, can also be observed when bringing together dropout with other regularization techniques. This can be seen by comparing the models with dropout against the same without dropout, respectively:

- Experiment 6 ($SE = 83,3\%$) versus 3 ($SE = 69,2\%$)
- Experiment 7 ($SE = 78,2\%$) versus 4 ($SE = 74,4\%$)
- Experiment 11 ($SE = 92,3\%$) versus 8 ($SE = 89,7\%$)
- Experiment 12 ($SE = 87,2\%$) versus 9 ($SE = 80,8\%$)
- Experiment 13 ($SE = 73,1\%$) versus 10 ($SE = 69,2\%$)
- Experiment 15 ($SE = 87,2\%$) versus 14 ($SE = 82,1\%$)

In the abovementioned case, the only exception is when comparing Experience 5 ($SE = 88,5\%$) with Experience 2 ($SE = 91,0\%$), since the last presents better results at correctly detecting this lesion.

Additionally, when combining data augmentation with other configurations, the models tend to perform better at correctly detecting melanoma than the same configurations without data augmentation. This is observable when respectively comparing:

- Experiment 7 ($SE = 30,0\%$) versus 1 ($SE = 16,7\%$)
- Experiment 9 ($SE = 26,7\%$) versus 2 ($SE = 20,0\%$)
- Experiment 10 ($SE = 56,7\%$) versus 3 ($SE = 6,7\%$)
- Experiment 12 ($SE = 23,3\%$) versus 5 ($SE = 20,0\%$)
- Experiment 13 ($SE = 56,7\%$) versus 6 ($SE = 26,7\%$)
- Experiment 14 ($SE = 33,3\%$) versus 8 ($SE = 20,0\%$)
- Experiment 15 ($SE = 23,3\%$) versus 11 ($SE = 16,7\%$)

It can also be observed that models with three or more techniques using batch normalization tend to perform worse at correctly detecting melanomas than the same models without this technique. This can be noticed when comparing the following experiments, being the last the same configurations firstly referred, without batch normalization:

- Experiment 11 ($SE = 16,7\%$) versus 6 ($SE = 26,7\%$)
- Experiment 12 ($SE = 23,3\%$) versus 7 ($SE = 30,0\%$)
- Experiment 14 ($SE = 33,3\%$) versus 10 ($SE = 56,7\%$)
- Experiment 15 ($SE = 23,3\%$) versus 13 ($SE = 56,7\%$)

Each one of the models has a different $SP$ which varies based on the lesion. For benign nevi, Experiment 10 achieves the highest true negative rate having a $SP = 75,0\%$. For melanoma the model with batch normalization (Experiment 2) achieves the highest true negative rates with $SP = 94,1\%$. For seborrheic keratosis, Experiment 11 achieves the highest true negative rates with $SP = 93,5\%$.

The following Figures (13, 14 and 15) show the representation of sensitivity (true positive rate) versus $1-$specificity (false positive rate) for all models and lesions.

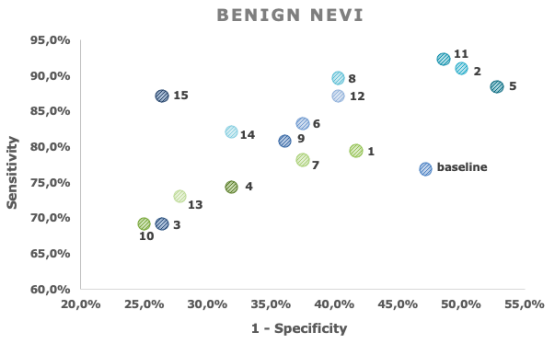


Fig. 13. Representation of each configurations for benign nevi lesions detection.

Each one of the points illustrated in the these Figures, represents a compromise between sensitivity and specificity, since the increase of sensitivity is coupled with a decline of specificity. In that sense, maximizing both sensitivity and specificity offers the best performance, i.e. the perfect model would have a $SE = 100\%$ and $1 - SP = 0\%$.

By looking into these Figures, it can be concluded that for benign nevi lesions (Figure 13), the model from Experiment 15 presents the best overall results, and the Baseline model presents the worst. When it comes to melanoma lesions (Figure 14), the best overall performance is achieved by the model from Experiment 4. The models with the worst compromise between sensitivity and specificity, for these lesions, are the Baseline and the one in Experiment 3. Finally, but not least, for

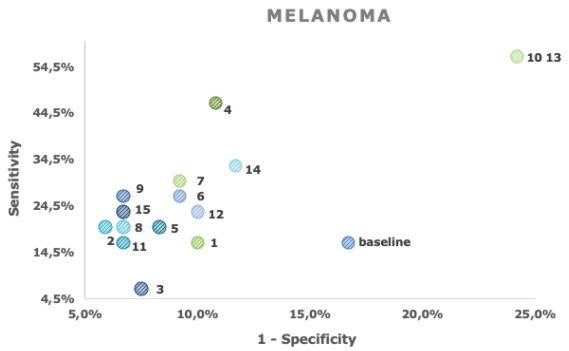


Fig. 14. Representation of each configurations for melanoma lesions detection.

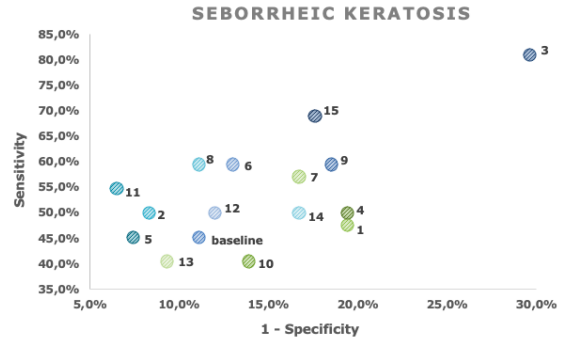


Fig. 15. Representation of each configurations for seborrheic keratosis lesions detection.

seborrheic keratosis (Figure 15), the model from Experiment 15 shows the best compromise, while models from Experiments 1 and 10 present the worst.

A final evaluation was carried out, to assess the performance of the best models on the test set. The selected models were those from Experiments 15, as it achieved the highest $BACC$ in the validation dataset ($BACC = 59.9\%$), and 10 and 13, as they outperformed all other models at correctly detecting melanoma (both with $SE = 56.7\%$ and respectively with a $BACC = 55,5\%$ and a $BACC = 56,7\%$). The results for these models, for test dataset, are demonstrated in Table V.

TABLE V
METRICS FOR THE MODELS THAT WERE EVALUATED WITH THE TEST DATASET.

| Experiment ID | Experiment 10 | Experiment 13 | Experiment 15 |
|---|---|---|---|
| **ACC** | 60,8% | 65,8% | 62,8% |
| **BACC** | 53,1% | 53,7% | 53,1% |
| **LOSS** | 0,924 | 0,889 | 0,909 |
| $\boldsymbol{SE_0}$ | 68,7% | 79,6% | 75,1% |
| $\boldsymbol{SE_1}$ | 49,9% | 32,5% | 23,1% |
| $\boldsymbol{SE_2}$ | 41,1% | 48,9% | 61,0% |
| $\boldsymbol{SP_0}$ | 60,7% | 49,6% | 54,1% |
| $\boldsymbol{SP_1}$ | 79,1% | 88,9% | 88,6% |
| $\boldsymbol{SP_2}$ | 89,6% | 92,6% | 85,7% |

As it can be observed, the model of Experiment 13 achieved a higher $BACC$ when compared with the other two evaluated, respectively $BACC = 53,7\%$ versus $BACC = 53,1\%$ for both Experiments 10 and 15. It also outperformed when comparing the other evaluation metrics ($ACC$ and $LOSS$). The model from Experiment 13 has a $ACC = 65,8\%$, showing, respectively, an increase of $3\%$ and $5\%$ versus the configurations of Experiments 15 and 10, and has a $LOSS = 0,889$, the best achieved value comparing with the last two referred.

The model that considers class weighting and data augmentation (Experiment 10) achieves a higher performance at correctly detecting melanoma when considering the test dataset (showing a $SE = 49,9\%$ versus $SE = 32,5\%$ for Experiment 13 and $SE = 23,1\%$ for Experiment 15). However both this Experiment and Experiment 13 maintain the highest values of sensitivity for melanoma lesions, as stated in the results of the three configurations against the validation dataset.

On the other hand, the model that considers dropout, class weighting and data augmentation (Experiment 13), presents the best performance at correctly identifying benign nevi lesions (with a $SE = 79,2\%$ versus $SE = 68,7\%$ for Experiment 10 and $SE = 75,1\%$ for Experiment 15). Likewise, the model considering all regularization techniques (Experiment 15) outperforms when it comes to correctly detecting keratosis seborrheic lesions (with a $SE = 61,0\%$).

Figures 16, 17 and 18 show the confusion matrices for these models, considering the test dataset.

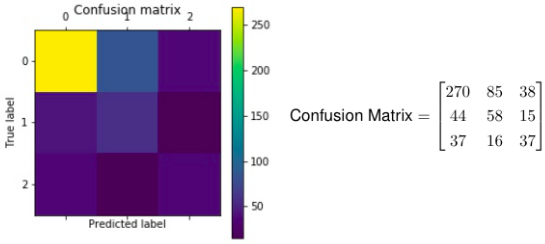


Fig. 16. Confusion Matrix for Experiment 10 (for test dataset).

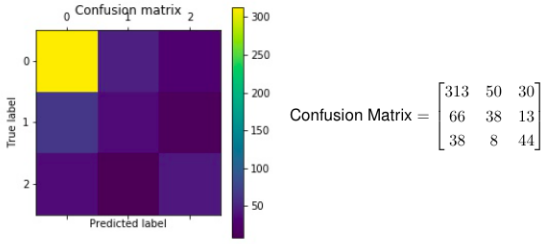


Fig. 17. Confusion Matrix for Experiment 13 (for test dataset).

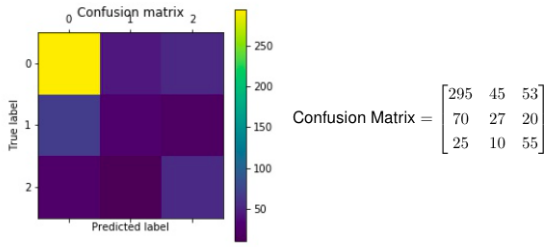


Fig. 18. Confusion Matrix for Experiment 15 (for test dataset).

## V. Conclusions

This thesis aimed to distinguish between melanoma, benign nevi and seborrheic keratosis lesions, by developing an end-to-end architecture based on neural networks. The main goal is to improve the performance of the automatic detection of the mentioned lesions.

By analysing the results of the experiments performed, it can be observed that every time a new regularization technique is added to the model, the performance improves when compared to the baseline model. Furthermore, it can be concluded that the model which achieves the best promising results, when considering an independent validation set, is the one that combine all the regularization techniques, obtaining a $BACC = 59,9\%$ for the validation dataset and $BACC = 53,1\%$ for test dataset.

Apart from the model combining all techniques, data augmentation alone seems to achieve a higher performance when compared with any other regularization technique.

Models that combine dropout and batch normalization tend to perform worse, which suggests that these techniques nullify each other. On the other hand, models that combine dropout and class weighting usually achieve a better performance. It seems that combining dropout or batch normalization with another regularization technique achieves an alike overall performance, potentially indicating a similar regularization effect.

The best configurations at correctly detecting melanoma are the ones considering dropout, class weighting and data augmentation ($SE = 56,7\%$ for validation and $SE = 32,5\%$ for test datasets), and class weighting and data augmentation ($SE = 56,7\%$ for validation and $SE = 49,6\%$ for test datasets). When considering data augmentation (alone or combined), models perform better at correctly identifying this lesion. However, when including batch normalization this detection worsens, a likely reason at why the best overall performing model ($BACC = 59,9\%$) is not the best one at detecting the mentioned lesion.

The configuration only using class weighting achieves the best sensitivity ($SE = 81\%$) in the validation dataset for seborrheic keratosis lesions. However, when adding other techniques to class weighting, the accuracy of correctly identifying this lesion varies significantly, not showing any clear improvement pattern.

The best configuration at accurately detecting benign nevi considers dropout, batch normalization and class weighting, with a $SE = 92,3\%$. Both batch normalization and dropout when used independently with other techniques correlate with a performance improvement at correctly identifying this lesion.

The model with the best compromise between $SE$ and $SP$ for benign nevi and seborrheic keratosis lesions is the one that considers all the regularization techniques. In regards to melanoma, the model with the best trade-off is the one that uses only data augmentation.

The best performing model in the test dataset considers dropout, class weighting and data augmentation ($BACC = 53,7\%$), although not being the highest performing configuration for the validation dataset ($BACC = 56,7\%$ versus the highest $BACC = 59,9\%$).

## References


[1] A. Krizhevsky, I. Sutskever, G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", Advances in neural information processing systems, 1097-1105, 2012.

[2] A. Lihachev, I. Lihacova, E. V. Plorina, M. Lange, A. Derjabo, J. Spigulis, "Differentiation of seborrheic keratosis from basal cell carcinoma, nevi and melanoma by RGB autofluorescence imaging", Biomedical optics express, 9(4), 1852-1858, April 2018.

[3] A. Masood, and A. Ali Al-Jumaily, "Computer Aided Diagnostic Support System for Skin Cancer: A Review of Techniques and Algorithms," International Journal of Biomedical Imaging, vol. 2013, 22, 2013.

[4] "American Cancer Society", https://www.cancer.org.

[5] Cancer Research UK, https://www.cancerresearchuk.org.

[6] Catarina Barata, Margarida Ruela, Mariana Francisco, Teresa Mendonça, and Jorge S. Marques, "Two Systems for the Detection of Melanomas in Dermoscopy Images using Textures and Color Features", IEEE Systems Journal, 8, 965-979, 2014.

[7] Devansh Bisla, Anna Choromanska, Russell S. Berman, Jennifer A. Stein, David Polsky, "Towards Automated Melanoma Detection with Deep Learning: Data Purification and Augmentation", The IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019.

[8] Diederik P. Kingma, Jimmy Lei Ba, "Adam: A Method for Stochastic Optimization", arXiv:1412.6980, 2015.

[9] E. Nasr-Esfahani, et al., "Melanoma Detection by Analysis of Clinical Images Using Convolutional Neural Network", 38th Annual International Conference of the IEEE EMBC, 1373-1376, August 2016.

[10] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M. and Thrun, S., "Dermatologist-level classification of skin cancer with deep neural networks", Nature, 542(7639), 115, 2017.

[11] I. Maglogiannis, C. Doukas, "Overview of Advanced Computer Vision Systems for Skin Lesions Characterization", IEEE Transactions on Information Technology in Biomedicine, 3(5), 721-733, September 2009.

[12] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, Deep Learning, http://www.deeplearningbook.org, MIT Press, 2016.

[13] "ISIC Archive", https://isic-archive.com.

[14] Julie A. A. Salido, Conrado Ruiz Jr., "Using Deep Learning for Melanoma Detection in Dermoscopy Images", IJMLC, 8(1), 2018.

[15] K. Simonyan, A. Zisserman, "Very Deep Convolutional Networks For Large-Scale Image Recognition", arXiv:1409.1556, 2014.

[16] N. C. F. Codella, Q. B. Nguyen, S. Pankanti, D. Gutman, B. Helba, A. Halpern, and J. R. Smith, "Deep learning ensembles for melanoma recognition in dermoscopy images", IBM Journal of Research and Development, 61, arXiv:1610.04662, 2017.

[17] Roberta B. Oliveira, et al., "Computational methods for the image segmentation of pigmented skin lesions: A Review, Computer Methods and Programs in Biomedicine", 31, 127-141, July 2016.

[18] Sergey Ioffe, and Christian Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift", International Conference on Machine Learning, 37, 448-456, arXiv:1502.03167, 2015.

[19] "TensorFlow", https://www.tensorflow.org.

[20] Y. Li, and L. Shen, "Skin lesion analysis towards melanoma detection using deep learning network", Sensors, 18(2), E556, 2018.