Melanoma Detection using deep learning methods
Joana Barbosa Miranda de Garcia Baleiras
Instituto Superior Técnico, Lisbon, Portugal

ABSTRACT
Melanoma stands out as the deadliest skin cancer due to its propensity to metastasize. Early detection is crucial to deliver successful treatment. Computer-Aided Diagnosis (CAD) systems have been developed to provide an automated melanoma detection. This paper proposes a new CAD system to distinguish between melanoma and benign skin lesions. The novelty lies in the use and comparison of three different Convolutional Neural Networks (CNN) at a time, namely AlexNet, VGG and ResNet. These CNN were trained using transfer learning and data augmentation. The image dataset was extracted from the ISIC 2017 Challenge. The original dataset contains 2000 training images (1626 benign lesions and 374 melanomas), 150 validation images (120 benign lesions and 30 melanomas) and 600 test images (483 benign lesions and 117 melanomas). The training set was then augmented and contained 1626 benign examples and 1870 malign examples. For a test set of 600 images, ResNet reached a sensitivity of 79%, a specificity of 60%, a balanced accuracy of 69% and an area under the receiver operating characteristic curve of 77%. Hence, the proposed system demonstrates the potential of CNN in CAD systems for automated melanoma detection.

Keywords — Melanoma, CAD system, automated melanoma detection, CNN, AlexNet, VGG, ResNet, transfer learning, data augmentation, ISIC Challenge

1 INTRODUCTION
Skin cancer is a serious public health problem [1]. The IARC estimated 1.3 million new skin cancer cases and 125 thousand skin cancer deaths around the world in 2018 only [2].

Among the various types of skin cancer, melanoma stands out as the deadliest of all. Globally, melanoma accounts for 75% of skin cancer deaths [3,4], although it only accounts for 1% of all skin cancers [5].

Melanoma can be treated with a simple excision if it is detected in its early stage (melanoma in situ or stage I of melanoma) [3,5,6]. Otherwise, with signs of metastasis spread to other parts of the body, only 20% of people survive the first five years after being diagnosed, according to the American Cancer Society (ACS) [7]. Therefore, the early detection of melanoma is vital.

Clinically, the diagnosis of skin lesions is performed by visual examination with or without the help of a dermatoscope. This device is a dermatoscopy instrument allowing for detailed visualisation of diagnostic structures that are not detectable by the human eye [8]. When a skin lesion is ambiguous, dermatologists may call for histological analysis.

However, the clinical method is subjective since it depends on the doctor’s experience and his/her visual acuity. Therefore, there has been an endeavour to develop automated systems for the diagnosis of skin lesions, called Computer-Aided Diagnosis (CAD) systems [4,9,10,11,12,13]. Generally, a CAD system comprises four main steps: “image pre-processing”, “lesion segmentation”, “feature extraction (and selection)” and “classification”. The last step can be performed with one or a few Machine Learning (ML) methods. Common ML methods implemented in CAD systems to classify skin lesions are Artificial Neural Networks (ANN), Support Vector Machine (SVM) and Decision Trees (DT) [14,15,16].

In recent years, a special type of deep ANN (called Deep Learning, DL), the “convolutional neural networks” (CNN), has emerged as the favourite method to detect and classify objects due to their success triggered by the annual ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) [17]. In a CAD system, a CNN is an end-to-end approach, i.e., it is able to play a role in all four steps. A neural network can work either as a method in one of the CAD steps (e.g., as a classifier in the classification step) or as an all-in-one method (as an end-to-end approach). However, a huge image dataset is crucial to produce reliable CNN results. For this reason, only a few works of CNN-based CAD systems for melanoma detection exist. Section 2.2, presents those works.

The first public database of melanoma images became available merely in 2016 with the International Skin Imaging Collaboration (ISIC) Challenge [18].

The principal motivation of this work is to take advantage of the available ISIC database and to develop a CNN-based CAD system for melanoma diagnosis. This work tests and compares three different types of CNN: AlexNet, VGG and ResNet [19,20,21]. The objective of this CAD system is to classify skin lesions as melanoma and non-melanoma and potentiate the use of CNN as a melanoma early diagnostic.

This paper is organised as follows. Section 2 addresses the melanoma disease, and presents the clinical and automatic detection methods in use. Section 3 overviews the CNN structure, in general, and the specific architecture of the three particular CNN used in this work. Section 4 describes the proposed melanoma’s detection system. Section 5 discusses

---

1 The International Agency for Research on Cancer (IARC) is the World Health Organization’s arm specialized in carcinogenic issues.
the experimental results. Finally, Section 6 concludes and suggests possible future work avenues.

2 MELANOMA DETECTION

2.1 The Melanoma

Cancer is caused by abnormal and unexpected development of human cells [22]. Skin cancer, in particular, develops in abnormal epidermal or dermal cells since epidermis and dermis are the two principal layers of the skin [3]. Basal cell carcinoma, squamous cell carcinoma and melanoma are the three most common skin cancers [23].

Skin lesions, or moles/nevi, fall into two categories: melanocytic and non-melanocytic. Pigmentary lesions (agglomerates of melanocytic cells) represent the former. The latter includes all types of skin lesions that are not formed by melanocytic cells, such as seborrheic keratosis, vascular nevi and dermatofibromas. Each skin lesion is described by a set of dermatoscopic structures (e.g., asymmetry, (ir)regular dots/globules, and (a)typical pigment network) and sometimes malignant and benign lesions share a few of them.

Melanoma enters the melanocytic category. This tumour is triggered by a malignant growth of melanocytic cells, which are localized in the epidermis [24] and are responsible for the melanin production, *i.e.*, the pigment that gives colour to the skin and hair [3]. Until an abnormal development, every melanocytic lesion is benign. Otherwise, with signs of malignancy, it is called “malignant melanoma” (or “melanoma” for simplicity) and, for advanced stages, the patient may not survive [3].

There are a few clinical methods to detect melanoma, such as the ABCD Rule [25] and Seven-Point Checklist [26]. (For the interested reader may refer to Section 2.4 in [27]). Since clinical methods are subjective and fail melanoma detection in its early stage, automatic algorithms (CAD systems) for melanoma detection have recently been developed to assist with the clinical diagnosis. Section 2.2 reviews briefly these previous CAD system works on melanoma diagnosis.

2.2 Automatic Methods – CAD systems

Huge difficulties on the correct visual examination of cutaneous lesions by inexperienced dermatologists [12] and the need to accelerate the melanoma detection led researchers to develop computer-aided diagnosis (CAD) systems for automated melanoma detection. The main purposes of these systems are to help dermatologists in their clinical evaluation of dermatoscopic lesions [28] and to speed-up the screening process, by allowing a pre-selection of suspicious lesions and a faster detection. Hence, these systems may increase the diagnostic accuracy.

2.2.1 The four steps of classic CAD systems

CAD systems proceed in four consecutive steps or blocks: pre-processing, segmentation, feature extraction and selection, and classification [9,29]. The last step relies on ML algorithms (such as SVM, ANN, logistic regression, k-nearest neighbours and decision trees) and the first three steps rely on image processing methods (such as Gaussian and median filters, colour constancy and texture techniques).

In the Pre-processing step, image acquisition imperfections (such as hairs, air bubbles and acquisition symbols) are eliminated from each dermatoscopic image and the contrast between the skin lesion and the surrounding skin is enhanced [11].

The Segmentation block splits the dermatoscopic image into two regions: one region associated with the skin lesion and the other region associated with healthy skin [11]. The border between these two regions is labelled the “contour of the lesion”.

The third step focuses on the extraction of a set of features allowing to distinguish melanoma from non-melanoma lesions. It is important to mention that the term *feature* may have two different meanings. On the one hand, it can represent a dermatoscopic structure. On the other, it can be an image characteristic, such as colour, texture, and geometry (*e.g.*, size, area, and perimeter) [30]. A CAD system can extract both clinical and image features. Generally, following the feature extraction step, a Feature Selection comes into action to eliminate redundant features or features without meaningful information about melanoma detection.

Finally, the Classification block predicts the skin lesion diagnosis. There are two types of outputs: binary diagnosis (two classes: benign and malign) and *n*-ary diagnosis (*n* classes). The more accurate the set of extracted features, the more correct is the classification [11]. The accuracy of the extracted features set measures how well that set distinguishes across skin lesions.

2.2.2 The importance of a large image dataset

A large enough database of real dermatoscopic images is a crucial requirement to produce reliable automatic detections of melanoma. Thus, image databases are required, and each image must be segmented and annotated (*i.e.*, labelled with the diagnosis of the image lesion) by experienced dermatologists [28]. When a database contains only a few images, it is common to augment training data artificially by rotating, flipping or producing non-linear distortions of the original images, for example [1]. These data augmentation (DA) procedures tend to improve the diagnostic accuracy of the system [31].

2.2.3 CNN-based CAD systems: literature recension

In the past few years, DL has been applied to several domains of study, namely natural language processing [32], image processing (such as in object detection [33]) and medical image analysis [34]. In the latter field, CNN have been widely used as an automatic method because CNN can learn to extract and classify digital images. For melanoma detection, CNN may learn by themselves which features perform better to distinguish melanoma from non-melanoma lesions; thus, this method can accomplish an automated melanoma diagnosis [1,13,35], whose diagnostic accuracy compares to the dermatologist’s diagnostic accuracy [13]. Hence, CNN-based CAD systems may help dermatologists to diagnose melanoma.
in its early stage of the clinical setting [13]. However, their results are still not totally reliable. In addition, there are only a few works using CNN in CAD systems due to the lack of public databases with a sufficient number of dermatoscopic images. The creation of a large public database for melanoma detection only appeared in 2016 associated to the ISIC Challenge. The latter was launched this year with the double purpose of providing the community with the largest set of publicly available high-quality dermatoscopic images [18] and promoting an annual competition to develop automated melanoma diagnosis algorithms. The following paragraphs review particular studies that provide a case for CNN-based CAD diagnostic systems.

Codella et al. proposed a system that became a new state-of-the-art in 2017 [1]. The authors combined machine learning techniques with neural networks to develop a CAD system that performs three steps: lesion segmentation, feature extraction and classification. The dataset was augmented. The classification results for average precision reach 64.5% and for the area under the receiver operator curve (ROC-AUC, explained in Section 5.2.) achieve 83.8%.

Esteva et al. published a remarkable study by applying deep neural networks to melanoma diagnosis [13]. The authors employed a single CNN-based CAD system to perform three classification tasks: differentiate across keratinocyte carcinomas and benign seborrheic keratoses, distinguish across melanoma and non-melanoma using only clinical images, and distinguish across melanoma and non-melanoma using dermatoscopic images [13]. For each task, the performance of the CNN exceeded the average performance of 21 certified dermatologists. In the first task this CNN reached a ROC-AUC of 96%, in the second task it achieved 94%, and in the last one attained 91%. However, the dataset for each task was different: 135 images, 130 clinical images (33 melanomas and 97 benign nevi), and 111 dermatoscopic images (71 melanomas and 40 benign nevi), respectively.

Haenssle et al. presented in 2018 a study in the Annals of Oncology (a medical journal of oncology) [34] that empower the CNN diagnostic accuracy above the diagnostic accuracy of a group of 58 dermatologists, similarly to the Esteva’s work. The classification task only required the distinction between melanoma and non-melanoma out of a 100-dermatoscopic images set. The dermatologists were given two opportunities to perform the clinical diagnosis, called level I and level II diagnoses. In level I, they had to provide the diagnosis and prescribe what the patient should do. Level II diagnoses were conducted four weeks later; the same doctors had to classify the same lesions and perform the same task, but at this moment, they were given more information about the patient as well as the chance of visualizing each lesion more closely. Even when compared to level II results, CNN showed better results. While in level I dermatologists had correctly identified 86.6% of all melanomas (i.e., sensitivity, SE, of 86.6%) and 71.3% of all benign nevi (i.e., specificity, SP, of 71.3%) and in level II they achieved 88.9% of SE and 75.7% of SP, CNN distinguished 95% of melanomas. Hence, CNN outperformed dermatologists on melanoma detection.

CNN-based CAD systems are thus a very promising method for melanoma diagnosis. Their main objectives are to help general practitioners to detect melanoma as early as possible [13,35] and to avoid unnecessary excisions [12].

3 CONVOLUTIONAL NEURAL NETWORKS

3.1 Main Concepts

The proposed solution in this work uses neural networks and CNN concepts. The explanation of the envisaged solution in Section 4 will benefit from a brief introduction to these notions.

CNN are neural networks that apply convolution operations to the input images. Neural networks consist of a set of neuron layers. The concept of “artificial neurons” was introduced by McCulloch and Pitts in 1943 [36]. A few others models of neural networks were proposed in the following years. The multilayer perceptron (MLP) stands out as it is the basis of many neural networks models in current days. It consists of several layers of neurons [37,38,39]. There are three types of layers: input layer, hidden layer and output layer. An artificial neuron involves two steps: i) a linear combination of inputs; ii) a non-linear function is applied to activate the neuron so that the signal can proceed to the next neuron. In CNN, the most frequent activation function is the Rectified Non-Linear Unit (ReLU) [19].

The training phase in a feed-forward neural network, such as a CNN, is done in a supervised way, which consists of computing the error (called “cost function”) between the desired and the network outputs for a training set. The purpose of the training step is to find a local minimum of the cost function [40]. This minimisation is usually performed by the gradient algorithm, which updates the weights (connections between neurons) in each training iteration. This algorithm works in two phases: forward propagation and backpropagation. In the first one, the input signal is passed from the input to the output layers through hidden layers, and the error between the desired and the network outputs is computed. The second phase features the calculation of the gradient, which is normally performed by the Backpropagation algorithm [ClassesofNeuralNet1998]. One way to accelerate the training phase is to use more efficient gradient optimization algorithms, such as Adam optimizer [41].

There are a few weight initialisation techniques [42], such as the transfer learning. This technique consists of reusing weights from an already trained network for a different, yet similar, problem [43].

The training dataset must be large enough so that the network does not memorise the training examples; otherwise, overfitting will occur. If the dataset is small, a few regularisation techniques may help to prevent overfitting [44], like weight decay and dropout [45].

The classification or generalisation phase consists of applying the trained model to classify a different dataset (test set). If the network classifies correctly the majority of test examples, then it has learned well during the training phase [46].
3.2 CNN Architecture

Given an image dataset, CNN can learn by themselves the most significant features to extract from each image. In their first hidden layers, CNN typically extract low-level features, such as edges, corners and curves, and then extract more and more high-level features, such as an object or pieces of objects [47,48].

Figure 1 depicts a CNN example. Comparing with a MLP, the fully-connected layers are converted into convolutional layers in a CNN. In convolutional layers, there are 3D convolution operations between the output of a previous layer and a set of kernels/filters [48]. The result of a convolution is called a “feature map”. After a convolutional layer, in which neurons are activated through an activation function (e.g. ReLU), there is a pooling layer. This layer is responsible for reducing the spatial dimensions of the receiving signal by splitting the input map into non-overlapping cells and replacing each cell by a number (e.g., average number or maximum [49]). After a sequence of convolutional and pooling layers, there are a few fully-connected layers. While the former focus on extracting the relevant features, the role of fully-connected layers is the classification. The neurons of these layers combine the extracted features and convert them into class scores. This conversion is done by the softmax function, instead of the ReLU function.

![Figure 1 – A simple CNN architecture.](image)

CNN have seen the light of glory days thanks to the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) in 2012 [17] and to the development of Graphic Processing Units (GPUs). That challenge focuses on algorithms for object detection and image classification and offers a huge public image dataset with 1000 object categories. Until 2012, the winners of this contest presented hand-crafted feature methods. Since that year, the winning algorithms have been CNN and, hence, nowadays they are widely used in several domains of study. AlexNet, VGG and ResNet are three CNN presented in ILSVRC at the 2012, 2014 and 2015 editions [19,20,21], respectively, and each of them is used in this work to build a CNN-based CAD system.

3.3 Three particular CNN: AlexNet, VGG and ResNet

This paper’s proposed detection system combines those three CNN, and thus it seems useful to highlight in the following paragraphs what they are about.

AlexNet [19] was the most promising CNN at the time, reaching a winning top-5 error rate of 15.3%\(^3\). It contains five convolutional (conv) and three fully-connected (fc) layers. Three of conv layers are followed by one pooling layer. The activation function is ReLU, except in the last fc layer that uses the softmax function.

VGG [20] was not the winner in the 2014 edition, but this CNN highlighted the importance of deepness in a CNN and achieved a top-5 error rate of 7.3% with the deepest configuration. VGG is a deep neural network and has a few possible configurations. VGG-16 is one of them and contains 13 conv and three fc layers. The activation function is ReLU, except in the last fc layer that uses the softmax function.

ResNet [21] was the winner in 2015 by attaining a top-5 error rate of 3.57% with the deepest configuration. Although this CNN is more complex than VGG, it is less deep [21] because of the introduction of the residual blocks, which allow to skip some layers. There are also a few possible ResNet configurations. ResNet-50 is one of them and it is described by 49 conv and one fc layers. The activation function is ReLU, except in the last fc layer that uses the softmax function.

4 THE PROPOSED MELANOMA’S DETECTION SYSTEM

4.1 Problem Formulation

The purpose of this paper is to present a CAD system for melanoma detection. The idea is to distinguish across melanoma and two benign skin lesions: melanocytic nevi and seborrheic keratosis. The latter are non-melanocytic. The two benign skin lesions are herein mentioned as “non-melanoma”, because the system only identifies lesions as melanoma and non-melanoma. It is not designed to differentiate between the two types of benign lesions. Given a dermatoscopic image, the system classifies it as a melanoma or as a non-melanoma. Thus, this is a binary classification problem, in which the training and test sets correspond to dermatoscopic images of skin lesions.

The problem under study faces three difficulties. First, the skin lesions, that the system has to differentiate, share a few morphological features. Second, images in the database were obtained from different image acquisition methods. Hence, there are differences with respect to colour and luminance, as well as imperfections in some images (e.g.: hairs and strange objects) associated with the quality of the acquisition method that may impair the results. These differences may lead to a

---


3 The top-5 error rate is the fraction of test images for which the correct label does not belong to the top-5 labels’ list (the ones with the highest probabilities).
misunderstanding of the relevant features to extract. Third, the training set of the image database is an imbalanced set because the number of benign examples is much higher than the number of malignant examples. Therefore, it is essential to develop a pre-processing step to mitigate these difficulties.

The image database was downloaded from the 2017 edition of the International Skin Imaging Collaboration (ISIC) challenge [18], as explained in Section 5.1.

### 4.2 System Architecture

Figure 2 presents the system built for automated melanoma diagnosis.

For each input dermatoscopic image, the system outputs a decision: melanoma or non-melanoma. This system was implemented for each CNN used. The following steps must be ran before that decision is made. First, the full dataset is preprocessed so that each image contains the skin lesion only. Second, the training phase is executed. The preprocessed training set is artificially augmented and, then, the CNN is trained to distinguish between melanoma and non-melanoma. The performance of the CNN is evaluated by two metrics using the validation set. At the end of the training, the best configuration of the CNN is selected. Then, in the classification phase, the system classifies each preprocessed test image as melanoma or non-melanoma, using the best configuration.

In the pre-processing block, all images are converted into the CNN’s image format and then normalized. The pre-processing modifications are:

- **Segmentation:** this transformation calculates the binary mask of each dermatoscopic image, but all images of the database already contain a segmentation mask. Thus, this phase was not implemented.
- **Cropping:** each image is cropped around the skin mole. First, the bounding box for each image is calculated from the segmentation mask. Then, the crop is done using the dimensions of the bounding box.
- **Resizing:** the input image of each CNN has 224x224 pixels. Therefore, each image is resized to that size.
- **Normalization:** in each CNN there is a specific required normalization method. In AlexNet the average intensity calculated from all images in the training set is subtracted at each pixel. Then, each pixel is divided by the standard deviation. In VGG and ResNet, images are mean-subtracted, i.e., the average intensity of each colour channel is subtracted at each pixel.

Data augmentation is frequently used for imbalanced datasets due to the improvement of the classifier’s performance and helps to prevent overfitting [50]. Hence, the set of malignant lesion images was augmented by rotations and vertical and horizontal flips since the training set is an imbalanced set. Additionally, a weight value was associated to each type of the lesion taking into account the total number of images of the new training set and the total number of images of the corresponding type ((1) for benign lesions and (2) for malignant lesions).

\[
\text{weight}_{\text{benign}} = \frac{\# \text{benign images of augmented training set}}{\# \text{benign images of training set}} \quad (1)
\]

\[
\text{weight}_{\text{malign}} = \frac{\# \text{malign images of augmented training set}}{\# \text{malign images of training set}} \quad (2)
\]

The classification of skin lesions is performed three times using one different CNN at a time: AlexNet [19], VGG [20], ResNet [21]. The reasons for using these and not other networks are their high-quality classification results on the ILSVRC competition\(^4\) and the possibility of training these networks with low-cost computing equipment.

However, the CNN models implemented here are a modification of the original architectures. In AlexNet and VGG, the last three fully-connected layers (with 4096, 4096 and 1000 units, respectively) are replaced with one fully-connected layer containing two neurons only (since the problem under study only needs two classes). The original ResNet only contains one fully-connected layer of 1000 units. Thus, this layer is replaced with another of the same type, but with just two neurons. For each network, dermatoscopic images have fixed dimensions and the dropout technique is applied during the training phase.

The training block is extremely important and determines how good each CNN performs in the classification step. This block is performed in the same way for each CNN, except for two training parameters: number of epochs and number of subsets of the validation set. In the forward propagation step, the network weights are initialised by transfer learning because the training set is small. In the backpropagation step, the gradient is defined by the Adam optimizer and the calculation of the gradient is implemented with the Backpropagation algorithm, in which the cost function is the cross-entropy (3) since this is the most common cost function used for classification problems [44].

\[
E(y, p) = - \sum_i y_i \log p_i \quad (3)
\]

where \( i \) represents each class (in this context, 1 or 2 since it is a binary classifier), \( y \) is a binary indicator that expresses if a class \( i \) is the correct classification for each example and \( p \) the predicted probabilities vector.

Since using all training examples in each iteration (batch mode) is computationally slow, the training set is divided into smaller sets (mini-batch mode). The dimension of each subset is defined by the batch size. The best configuration of each CNN is chosen with respect to one of two metrics measured in the validation set: loss and balanced error. The first metric is related to the cost function used. The second metric

\(^4\) Presented in Section 3.3.
corresponds to 1 – balanced accuracy, where balanced accuracy is the average accuracy of the two classes [51].

5 EXPERIMENTAL RESULTS

5.1. Image Database

The image database belongs to the 2017 edition of the ISIC Challenge [18]. The dataset contains 2000 training images, 150 validation images and 600 test images. The training set consists of 374 melanoma images and 1626 non-melanoma images (from which 254 of image lesions are seborrheic keratoses). After the data augmentation, the total number of malign lesions is 1870. The validation set is split into 120 benign lesions and 30 melanoma lesions. Finally, the test set is decomposed into 483 benign lesions and 117 melanomas.

For each lesion image, a binary mask (i.e. the segmentation images) is provided. These masks are essential in this work to obtain the dimensions of bounding boxes that are used in the cropping step of the pre-processing phase. The masks have the same image dimensions of the associated lesion image. A specialised doctor made the segmentation of this set either manually or through semi-automated methods.

Besides the original skin lesions images, the dataset contains the clinical diagnosis of each image, which indicates if the lesion is benign (non-melanocytic or seborrheic keratoses) or malignant. But the proposed system only discriminates melanoma from the other two types of benign lesions.

5.2. Performance Metrics

There are four types of results of a binary classifier: true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

- TP: a positive example is correctly classified as positive;
- TN: a negative example is correctly classified as negative;
- FP: a negative example is classified as positive;
- FN: a positive example is classified as negative.

For binary classifiers, there are a few performance measures, such as sensitivity, specificity, area under the Receiver Operating Characteristic (ROC) curve and balanced accuracy (BACC) [52].

Sensitivity (SE) [52], or true positive rate, is the proportion of correct classifications of positive cases taking into account all positive cases.

Specificity (SP) [52], also known as a true negative rate, quantifies the proportion of correct classifications of negative cases taking into account the entire negative population.

The ROC curve [53] is a probability curve. It relates graphically the SE and probability of false alarm (corresponding to 1 − SP) of the classification test. The area under the ROC curve (ROC-AUC) represents the classifier’s capacity of distinguishing between classes. The bigger the ROC-AUC value, the better is the classifier’s capacity. Thus, if ROC-AUC=1, the model distinguishes perfectly between the positive class and the negative class. But if ROC-AUC=0.5, the model classifies positive examples as negative examples and negative examples as positive examples.

The balanced accuracy (BACC) expresses the average accuracy obtained for each of the classes [51]. It is used when the dataset is imbalanced in one of the classes.

5.3. Description of Experimental Results

5.3.1. CNN Training

Each CNN has its own architecture and training convergence rate. Therefore, the number of epochs for each CNN is different: 1000 epochs for AlexNet, 600 for VGG and 500 for ResNet. The CNN weights were previously trained with the ImageNet database. In the proposed system, they are initialised by transfer learning. Each CNN is trained with mini-batch mode with batches of 46 images.

5.3.2. Hyperparameter Selection

Dropout and learning rate are the two hyperparameters evaluated for each CNN in the proposed system. Additionally, the data augmentation technique is analysed only for AlexNet.

For dropout, four probability values are tested: 0%, 30%, 50%, 70%. For learning rate, a fixed rate of 1x10^-4 and a non-fixe rate (which value varies from 100 to 100 epochs) are tested.

The selection of the best set of hyperparameters is made based on the performance of each CNN in the validation set, which is measured by the loss and balanced error. For each measure, the network performance is evaluated by SE, SP and BACC. Then, the best set of hyperparameters (i.e., the best configuration) of each CNN are selected by the metric that presents the best performance results.

AlexNet

The Figure 3 shows the convergence curves for the best configuration of AlexNet: dropout=30%, a non-fixed learning rate and with data augmentation.

![Figure 3](image320x224 to 563x307)

Figure 3 – Best configuration of AlexNet on the validation set, given by the balanced error metric (on the left) and the loss metric (on the right).

The classification results of AlexNet under the data augmentation technique are undoubtedly better than the results obtained without data augmentation. The best configuration with data augmentation reached SE=70%, SP=75%, BACC=73%, while without data augmentation the CNN attained SE=20%, SP=98%, BACC=59%. Thus, AlexNet without the data augmentation technique is not able to distinguish melanoma from non-melanoma. Furthermore, the CNN merely converges and is influenced by overfitting.

In general, the results with a non-fixed learning rate are better when using a fixed learning rate. The CNN converges,
the overfitting is prevented and the oscillation frequency is smoother. Thus, the classification of the AlexNet is more precise.

The dropout accelerates the convergence of the CNN and also helps to prevent overfitting. Under a fixed learning rate, the bigger the dropout value, the better is the CNN’s SE. Under a non-fixed learning rate, the effect of the dropout is not significantly noticed on the classification results.

VGG

The Figure 4 shows the convergence curves for the best configuration of VGG: dropout=70% and a non-fixed learning rate with SE=87%, SP=73% and BACC=80%.

The best performance results of VGG are reached by varying the learning rate, rather than using a fixed value. The non-fixed learning rate reduces the amplitude of oscillation between epochs, presents lower values for the convergence curves and also prevents overfitting.

The VGG configurations obtained with dropout achieved a faster convergence rate and a lower number of oscillations (specially the oscillations of the loss curve) than the experiences without dropout.

ResNet

The Figure 5 shows the convergence curves for the best configuration of ResNet: dropout=50% and a non-fixed learning rate with SE=80%, SP=78% and BACC=79%.

Comparing the performance results of ResNet by varying the learning rate with the ones of using a fixed value of learning rate, the former results are slightly better. This variation of this hyperparameter prevents overfitting and accelerates the CNN’s convergence rate.

The best configurations of ResNet were obtained with dropout (30% and 50%). By comparing the ResNet experiences without dropout with ones influenced by dropout, it is possible to analyse that the network converges faster. Between the two best configurations, the oscillation frequency of the network given by the probability value of 50% is smoother than the network given by the probability value of 30%. Also, the convergence rate is higher.

5.4. Performance Evaluation

After selecting the best configuration for each CNN, each network is tested using the test set. Then, the CNN that best suits the classification between melanoma and non-melanoma is selected.

The classification results for each CNN is described in Tables 1, 2 and 3. Figure 6 illustrates the ROC curve in the test set for each network.

Table 1 – Classification results of the best AlexNet configuration.

<table>
<thead>
<tr>
<th>AlexNet</th>
<th>dropout=30%, non-fixed learning rate</th>
<th>epcoh_max=960</th>
<th>epcoh_max=300</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>0.54</td>
<td>0.56</td>
<td></td>
</tr>
<tr>
<td>Specificity</td>
<td>0.71</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>BACC</td>
<td>0.64</td>
<td>0.63</td>
<td></td>
</tr>
<tr>
<td>ROC-AUC</td>
<td>0.69</td>
<td>0.68</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 – Classification results of the best VGG configuration.

<table>
<thead>
<tr>
<th>VGG</th>
<th>dropout=70%, non-fixed learning rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>epcoh_max=599</td>
<td>epcoh_max=200</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.65</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.69</td>
</tr>
<tr>
<td>BACC</td>
<td>0.76</td>
</tr>
<tr>
<td>ROC-AUC</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Table 3 – Classification results of the best ResNet configuration.

<table>
<thead>
<tr>
<th>RESNET</th>
<th>dropout=50%, non-fixed learning rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>epcoh_max=400</td>
<td>epcoh_max=280</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.79</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.61</td>
</tr>
<tr>
<td>BACC</td>
<td>0.70</td>
</tr>
<tr>
<td>ROC-AUC</td>
<td>0.77</td>
</tr>
</tbody>
</table>

ResNet is the best CNN for melanoma detection. It presents the best classification results, SE=79%, SP=60%, BACC=70% and ROC-AUC=77% (Table 3). For the malign test set, this network identified 93 lesions as malign and misclassified the remaining 24 lesions as benign. VGG reached SE=78%, SP=59%, BACC=68% and ROC-AUC=75% (Table 2). AlexNet is the poorest one. On the one hand, its capacity to distinguish melanomas from non-melanomas lies below 60%. On the other, it attained SE=56%, SP=70%, BACC=63% and ROC-AUC=68% (Table 1).
Figures 7 and 8 show three images of malign lesions and three of benign lesions (after the application of the bounding box for each image), respectively: on the left side, there are lesions correctly classified by ResNet, on the right side there are the misclassifications performed by ResNet.

![Melanoma examples](image1)

**Figure 7** – Melanoma examples: on the left, correctly classified; on the right, misclassified as benign lesions.

![Non-melanoma examples](image2)

**Figure 8** – Non-melanoma examples: on the left, correctly classified; on the right, misclassified as malignant lesions.

## 6 CONCLUSION AND FUTURE WORK

This section summarizes the main achievements and highlights the main conclusions of this work. Also, it presents some suggestions for future work.

The main purpose of this work was to develop a system for automatic melanoma detection based on convolutional neural networks (CNNs). The proposed system consisted of two principal phases: training phase and classification phase. For each one, the following three CNNs were used since they require low-cost computing equipment: AlexNet, VGG and ResNet.

The performance of each CNN was discussed and the best CNN was selected. The classification results were categorised by four well-known performance measures: sensitivity (SE), specificity (SP), balanced accuracy (BACC) and area under the ROC curve (ROC-AUC).

ResNet reached the best classification results by attaining SE=79%, SP=60%, BACC=70% and ROC-AUC=77%. This CNN presents the most complex architecture, but it needs less number of epochs to converge when compared with VGG and AlexNet. VGG was the network which required more computational memory and, consequently, needed more time to train. The AlexNet’s architecture is the simplest one. Therefore, is the most suitable CNN if computational memory and time are very restricted. However, this network needs a higher number of epochs to converge and the classification results are not so good. From a set of 117 malign lesions, this CNN classified correctly 66 lesions and the remaining 51 were misclassified as benign lesions.

The selection metrics evaluated on the validation set to select the best configuration for each CNN were the loss and balanced error. Although the comparison of these two metric has not been the object of study in this work, the balanced error was more robust and consistent. The best loss’s results occurred mostly at the end of the training phase. Hence, those results might be improved by overfitting, since this phenomenon occurs in the last training epochs.

The image database was limited and imbalanced, since there were very few images of malignant skin lesions. For these two reasons, the data augmentation technique was applied on the training set of malignant lesions.

The results obtained in this work show that it is in fact possible to automatically distinguish skin lesions between melanoma and non-melanoma. However, there is still room for better results. Thus, some suggestions are herein presented as future work.

First, it would be worthy to implement the proposed system using one or more GPUs in order to accelerate the CNNs’ training, which is a time consuming when computing the CNN hyperparameters. In addition, it would be very interesting to either use more complex CNNs with better classification results in image analysis problems or do some modifications in CNNs which were implemented in studies of automatic melanoma detection.

It would be interesting to explore other ways to enhance the hyperparameters optimization. The selection of other hyperparameters could be beneficial since it might boost sensitivity of the CNN, thus, its melanoma detection performance. The activation function, the batch size and the weight initialisation are a few examples of different hyperparameters.

On the other hand, it would be useful to evaluate the impact on classification results having three classes, rather than two, since the database still had a relevant number of seborrhoeic keratoses.

It would be very useful to explore image databases that contain an huge number of different dermatologic images and are preferentially balanced.

Finally, it would be interesting to deeply study the subject of selection metrics. On the one hand, understand which selection metric is the most suitable to choose the best CNN configuration. On the other, explore if there are other selection metrics.

## 7 REFERENCES


