# Evaluation of Melanoma Diagnosis using Imbalanced Learning

**Lucas Bezerra Maia, Alan Carlos Lima, Pedro Thiago Cutrim Santos,
Nigel da Silva Lima, Humberto Oliveira Serra
Geraldo Braz Junior, João Dallyson Sousa de Almeida, Anselmo Cardoso Paiva**

[1]Applied Computing Center, Federal University of Maranhão (UFMA)
Av. dos Portugueses, 1966 - Bacanga, São Luís - MA, 65080-805, Brazil

{lucas.maia, alanlima, thiagocutrim}@nca.ufma.br

{nigel.lima, humberto.serra}@huufma.br

{geraldo, jdallyson, paiva}@nca.ufma.br

***Abstract.*** *Melanoma is the most lethal type of skin cancer when compared to others, but patients have high recovery rates if the disease is discovered in its early stages. Several approaches to automatic detection and diagnosis have been explored by different authors. Training models with the existing data sets has been a difficult task due to the problem of imbalanced data. This work aims to evaluate the performance of machine learning algorithms combined with imbalanced learning techniques, regarding the task of melanoma diagnosis. Preliminary results have shown that features extracted with ResNet Convolutional Neural Network, along with Random Forest, achieved an improvement of sensibility of approximately 21%, after balancing the training data with Synthetic Minority Oversampling TEchnique (SMOTE) and Edited Nearest Neighbor (ENN) rule.*

## 1. Introduction

Melanoma, even though it is considered the most lethal type of skin cancer, it has high cure rates when diagnosed in its early stages. The disease may manifest itself in normal skins with the appearance of a dark mole and borders presenting irregularities, or from a pre-existing pigmented lesion where the tumor might evolve showing color changes and enlargement of the lesion area [Soares 2008].

Understanding the importance of the diagnosis of skin cancer in its early stages, digital image processing methods are being developed with the aim of improving existing diagnostic techniques. The idea is to provide a classification model for diagnosing skin cancer that may decrease cases of misdiagnosis and increase the chances of the patient receiving treatment in a timely manner.

The process of learning and the predictions of machine learning algorithms can be affected by the problem of imbalanced data set, which corresponds to the difference of the number of samples in different classes [Liu and Chawla 2011]. Available databases of skin cancer images such as ISIC [Gutman et al. 2016] and $PH^2$ [Mendonça et al. 2013] present Imbalanced Ration (IR) of 4.2 and 4 respectively, considering the quantity of non-melanoma examples (negative class) over melanoma samples (positive class), which makes harder the task of training models good enough to generalize real cases data.

Some approaches have been explored in order to provide automatic melanoma diagnosis and soften imbalanced data sets issues. In [Codella et al. 2016] and [Santos et al. 2017], convolutional neural networks architectures such as Deep Residual Network (ResNet), proposed in [He et al. 2016], Alexnet and VGG-F were used with dermatoscopic images as feature extractor. To avoid class imbalance problem, images were cropped, flipped or rotated to increase sample instances for training, based on the principle of replicate samples in dataset, named data augmentation. The classification of the extracted data was performed by traditional algorithms.

Another study found in [Lopez et al. 2017] used an adaptation of the architecture proposed in [Simonyan and Zisserman 2014] to evaluate the effects of initialization of the network parameters according to three paradigms: (i) weights initialized randomly and trained from scratch, (ii) weights initialized with values of other pre-trained networks and (iii) fine tune the weights of other networks by training on dermatoscopic images.

In [Moura et al. 2017], a hybrid descriptor for extraction of characteristics was proposed. Several methods such as Grey-Level Co-occurrence Matrix (GLCM), Histograms of Oriented Gradients (HOG), Local Binary Pattern (LBP) and others were explored. After combination of the main descriptors, a selection of 10% of the most relevant attributes with a Gain Ratio Information was performed to reduce feature space complexity of the different classes.

This work has as main objective to analyze the use of existing techniques for class balancing and noise reduction as well as their effects on an imbalanced dermatoscopic images data set. Two classical machine learning algorithms are trained and tested in order to evaluate the performance on different training set scenarios. As contribution, it is expected to find a suitable pipeline model to aid in the automatic melanoma diagnosis with the use of dermatoscopic exams.

## 2. Proposed Methodology

The block diagram of the methodology proposed in this study is summarized in Figure 1. The diagram shows the use of a data set that will undergo a simple preprocessing step, followed by the phase of feature extraction, whose vectors will be used to train the classifiers on two different scenarios, and so it will be compared the performance of the algorithms.

### 2.1. Data Set Preparation

It was used the $PH^2$ [Mendonça et al. 2013] dataset. It contains 200 dermatoscopic images cataloged in three different classes: Normal lesions (80 samples), Atypical lesions (80 samples) and Melanoma (40 samples). However, for this study, the dataset was divided into Non-melanoma and Melanoma images, merging the classes of normal and atypical lesions.

To prepare the images for feature extraction phase, regions of interest (ROI) were extracted by using a binary mask containing the specialist marking of the lesion. Due the fact that the images present different dimensions, vertical and horizontal edges of pixel value 0 were added in order to create a squared image without changing aspects of the lesion itself.

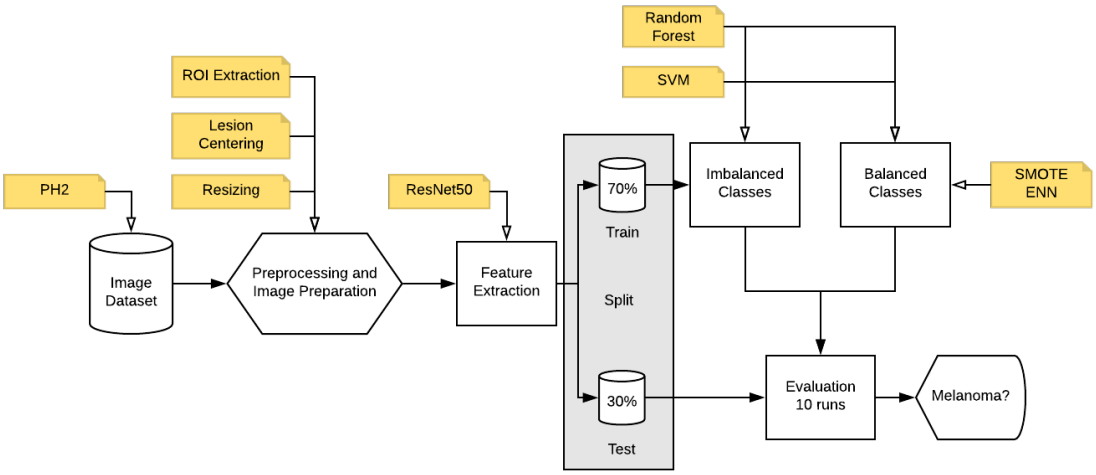


**Figure 1. Flowchart of the proposed methodology.**

## 2.2. Feature Extraction

In this work, a Convolutional Neural Network (CNN) was used to extract image traits forming a feature map that will serve as input of the machine learning algorithms. CNNs aim to continuously divide images in order to extract characteristics from them, inspired by the visual cortex of animals [Hubel and Wiesel 1968]. Such networks carry out learning through signal propagation and backpropagation of error methods [Rumelhart et al. 1986].

The architecture used was the Deep Residual Network (ResNET) [He et al. 2016] implemented by Keras library [Chollet et al. 2015] loading weights trained on the ImageNet challenge [Deng et al. 2009]. This technique is called Transfer Learning [Menegola et al. 2017], where, due to the insufficient amount of data to create a new model, knowledge from a large amount of known data (source domain) is transferred to a set of a new data (target domain). A vector of 2048 features was then generated for each image of the data set.

## 2.3. Balancing Techniques

In order to balance the number of examples from different classes, several oversampling algorithms have been proposed. One of which is called Synthetic Minority Oversampling TEchnique (SMOTE) [Chawla et al. 2002] uses a certain neighborhood of size *k* in respect to a pivot example to create new samples by interpolation. Mathematically:

$$X_{new} = X_{pivot} - \lambda(X_{ki} - X_{pivot})$$

where $\lambda$ is a random number in the range [0, 1] and $X_{new}$ is the interpolation that will be created as a sample on the line between $X_{pivot}$ and its neighbor $X_{ki}$. Implementations of SMOTE differs mostly in the way the set of pivots are selected.

After the generation of synthetic data, possible noisy samples may appear. For this reason, under-sampling algorithms also became important in the process of balancing databases. This study investigates Edited Nearest Neighbor (ENN) rule that was proposed

in [Wilson 1972], which is used as neighborhood cleaning strategy capable of removing any example whose class value differs from the class of at least two of its three nearest neighbors. However, in order to increase data cleaning and provide more control to the algorithm, ENN can be modified to under-sample a data set by erasing samples which are not in compliance with the *k*-nearest neighbors. The larger the value of *k* is, more samples are cleaned up from the dataset. As a result, a simpler border among the classes is created.

For the purpose of this work, the set of pivots chosen for SMOTE was the Support Vectors (SVs) found by a separate Support Vectors Machine trained with the imbalanced data. The value for *k* was configured to be 3 for both SMOTE and ENN algorithms.

### 2.4. Experiments

As validation method, the data was split in 70% for training and 30% to test the models. This process was repeated 10 times, then the mean and standard deviation were calculated for the metrics: accuracy, recall, specificity and f1-score.

Feature vectors generated by the ResNET was used to execute two types of training in two different machine learning algorithms: Support Vectors Machine (SVM) [Hearst et al. 1998] and Random Forest (RF) [Ho 1995]. The first type of training was conducted by passing the imbalanced training set to the algorithms.

The second one was made by creating a pipeline with oversampling the training data using SMOTE and then cleaning noisy samples with ENN. Later, the remaining examples became input to the classifiers.

Each learner had the parameters automatic estimated using auto-sklearn library [Feurer et al. 2015] in order to generate the best classification model.

## 3. Preliminary Results

In this section, preliminary results obtained by each classifier are presented, showing the different scenarios of training. For a good classifier, the ability to hit the percentage of melanoma (recall) must be maximized with the correct percentage of examples of normal cases (specificity). The reason for this choice is related to the fact that it is better to detect melanoma, even generating more false positives than not perceiving melanoma in a risk patient. Table 1 summarizes the performances of each model presenting their average and standard deviation.

**Table 1. Initial results using imbalanced learning for melanoma diagnosis.**

| Model | Accuracy | Recall | Specificity | F1-Score |
|---|---|---|---|---|
| SVM | $0.8683 \pm 0.0621$ | $0.4250 \pm 0.3178$ | $0.9792 \pm 0.0380$ | $0.4906 \pm 0.3554$ |
| RF | $0.8783 \pm 0.0273$ | $0.6083 \pm 0.1802$ | $0.9458 \pm 0.0522$ | $0.6557 \pm 0.1074$ |
| SMOTEENN + SVM | $0.8483 \pm 0.0535$ | $0.5667 \pm 0.4002$ | $0.9187 \pm 0.0886$ | $0.5037 \pm 0.3514$ |
| SMOTEENN + RF | $0.8817 \pm 0.0346$ | $0.8167 \pm 0.1165$ | $0.8979 \pm 0.0585$ | $0.7366 \pm 0.0586$ |

Firstly, analyzing initial results, the training process of both algorithms tends to favor the class with more sample data, shown by the specificity mean values of 97.92% and 94.58% for SVM and RF respectively, whereas the percentages of melanoma correctly classified (recall) were 42.50%, reached by SVM, and 60.83%, by RF. On the other hand,

when simple balancing techniques (SMOTE + ENN) were included before the training of the machine learning algorithms, an increase of recall can be noticed. SVM had an improvement of approximately 14%, whereas Random Forest reached 81.67% with an increase of almost 21%. Also F1-Score indicates that the best produced model was found when used this last approach.

It also can be noticed that RF, using SMOTE and ENN, achieved the best overall mean accuracy of 88.17%, while keeping a balanced ratio of correctly classified instances for both normal and melanoma class.

## 4. Conclusion

The main objective of this under construction study was to evaluate the performance of machine learning algorithms combined with dataset balancing techniques, regarding the task of skin cancer classification. Preliminary results have shown that features extracted with ResNet, along with Random Forest, achieved best recall value after using a pipeline by adding synthetic data and cleaning noisy samples before training the models.

As future work, it is planned to experiment the method on other available databases of dermatoscopic images, for example: ISIC; use other variations of over and under-sampling techniques; estimate the best parameters of those sampling algorithms; explore features extracted by other CNNs architectures (perhaps with fewer layers or combination among them, such as InceptionResNet); Test other classifiers, such as: Linear Regression, Dimensionality Reduction Algorithms, Bagging; In addition, use image quality enhancement and noise reduction techniques that were not explored in this study yet, such as: Histogram Equalization and Median Filter.

## References


Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Chollet, F. et al. (2015). Keras. `https://keras.io`.

Codella, N., Nguyen, Q.-B., Pankanti, S., Gutman, D., Helba, B., Halpern, A., and Smith, J. R. (2016). Deep learning ensembles for melanoma recognition in dermoscopy images. *arXiv preprint arXiv:1610.04662*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., and Hutter, F. (2015). Efficient and robust automated machine learning. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 2962–2970. Curran Associates, Inc.

Gutman, D., Codella, N. C., Celebi, E., Helba, B., Marchetti, M., Mishra, N., and Halpern, A. (2016). Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28.

Ho, T. K. (1995). Random decision forests. In *Document analysis and recognition, 1995., proceedings of the third international conference on*, volume 1, pages 278–282. IEEE.

Hubel, D. H. and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243.

Liu, W. and Chawla, S. (2011). Class confidence weighted knn algorithms for imbalanced data sets. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 345–356. Springer.

Lopez, A. R., Giro-i Nieto, X., Burdick, J., and Marques, O. (2017). Skin lesion classification from dermoscopic images using deep learning techniques. In *Biomedical Engineering (BioMed), 2017 13th IASTED International Conference on*, pages 49–54. IEEE.

Mendonça, T., Ferreira, P. M., Marques, J. S., Marçal, A. R. S., and Rozeira, J. (2013). Ph2 - a dermoscopic image database for research and benchmarking. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, 2013:5437–40.

Menegola, A., Fornaciali, M., Pires, R., Bittencourt, F. V., Avila, S., and Valle, E. (2017). Knowledge transfer for melanoma screening with deep learning. In *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*, pages 297–300. IEEE.

Moura, N., Rodrigo, V., Kelson, A., Machado, V., and Santos, L. (2017). Proposta de um descritor híbrido para aprimoramento da identificação automática de melanoma. *XXXVII Congresso da Sociedade Brasileira de Computação – 17º WIM - Workshop de Informatica Médica*, pages 2034–2043.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533.

Santos, A., Kelson, A., Rodrigo, V., Uchôa, V., and Santos, L. (2017). Uma abordagem de classificação de imagens dermatoscópicas utilizando aprendizado profundo com redes neurais convolucionais. *XXXVII Congresso da Sociedade Brasileira de Computação – 17º WIM - Workshop de Informatica Médica*, pages 2010–2019.

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

Soares, H. B. (2008). *Análise e classificação de imagens de lesões da pele por atributos de cor, forma e textura utilizando máquina de vetor de suporte*. PhD thesis, Universidade Federal do Rio Grande do Norte.

Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, 3:408–421.