v. 5, n. 1, 2025

Journal of
Engineering
Research

Acceptance date: 20/01/2025

USING DATA AUGMENTATION TECHNIQUES ON DERMOSCOPIC IMAGES TO IMPROVE THE ACCURACY OF THE CONVOLUTIONAL NEURAL NETWORK

Ademar Takeo Akabane
Faculty of Computer Engineering

Luiz Henrique Souza
Custodio da Silveira Curso de Engenharia de Computação
http://lattes.cnpq.br/1506152684437729

Abstract: Computer vision and machine learning techniques capable of assisting specialists in different areas in their day-to-day tasks are the subject of several studies. For example, in the area of health, there is the diagnostic aid system that focuses on diagnosing certain diseases early. This early diagnosis is very important for improving patients' quality of life. One of the main fields using these techniques is the classification and detection of objects in images using *convolutional neural networks*. It is worth noting that when developing applications using *deep* learning models, large volumes of data are needed to train the networks. And one of the major problems is the difficulty of obtaining a data set large enough to adequately train convolutional neural networks. One way around this problem is to create synthetic data from the available images, i.e. to apply data *augmentation* techniques. In this work, *data augmentation* techniques will be applied to improve the classification accuracy of skin lesion images using a convolutional neural network. At the end of this work, it is hoped that the techniques applied can be used as inspiration for other diagnostic aid systems and also to improve existing applications in the medical field.
Keywords: Machine Learning, Convolutional Neural Networks, Automated Medical Diagnosis, Data Augmentation, Image Processing

INTRODUCTION

Machine Learning (ML) is a sub-area of Artificial Intelligence based on systems capable of acquiring knowledge automatically and improving with experience [1]. In AM, inductive learning is the most common, i.e. induction techniques are applied to increase the model's ability to generalize about a particular set of examples [2]. This type of learning can be divided into three categories: supervised, unsupervised and semi-supervised learning.

It is well known that image classification is one of the best-known tasks in several application areas. It is worth noting that the accuracy of the classification task is related to the number of labeled examples available in the model's training stage. In other words, the greater the number of examples, the greater the prediction capacity of the classifiers used. However, this scenario described in real-life situations is not reflected, i.e. obtaining data sets with a large number of labeled samples is not always an easy task, and is often costly [3]. To get around this problem, *data augmentation* (*DA*) techniques can be used. DA is a concept based on computational techniques with the aim of increasing the number of labeled samples in a pre-existing data set and thus improving the accuracy of classification models [3, 4, 5], see Figure 1.1.

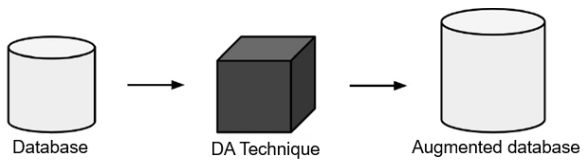

Figure 1.1: Illustrative example of DA.

Thus, the DA technique is a method capable of augmenting samples from the training data set while preserving their original label, and can therefore be represented as the mapping in Equation 1.1. As DA seeks to preserve the original labels of a sample, it means that if a given x has a label y, then $\varphi(x)$ will also have the label of y [6].

$$= \phi S + \rightarrow T \quad (1.1)$$

where S is the original training set and T is the set augmented from S. Thus, the augmented data set S', containing the original data and the data augmented after applying the φ technique, is represented by Equation 1.2.

$$S' = S \cup T \quad (1.2)$$

Therefore, there are several approaches to obtaining more samples in the training set. One of them is to apply image transformation

effects to generate new samples. The image transformation method is a simple idea and is widely used when the aim is to obtain more image samples [3, 4]. In image transformation, transformation effects or filters are applied to the images, thus generating other images. These image transformations are generally divided into two categories [6]: photometric techniques and geometric techniques. In the first category, they apply effects such as noise, blur, color effects (e.g. brightness, saturation, *color jittering*); while in the second, they apply rotations, translations, scales, *flips*.

Figure 1.2 shows an illustrative example of the application of photometric effects, where Figure 1.2(a) shows the original image. Figure 1.2(b) shows the same image with noise applied, Figure 1.2(c) shows changes in saturation and Figure 1.2(d) shows changes in brightness.

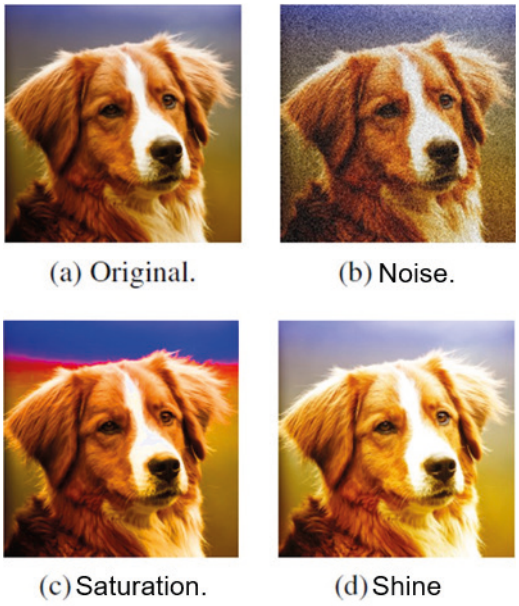

Figure 1.2: Illustrative example of the photometric transformation.

Figure 1.3 shows the geometric effects from an original image, Figure 1.3(a). Figure 1.3(b) shows the horizontal flip effect, Figure 1.3(c) shows rotation and Figure 1.3(d) shows scaling.

OBJECTIVES

This work plan aims to apply *data augmentation* methods to skin lesion images order to improve the accuracy of machine learning models. The main idea is to understand how the generation of new training data can increase the generalization capacity of the convolutional neural network.

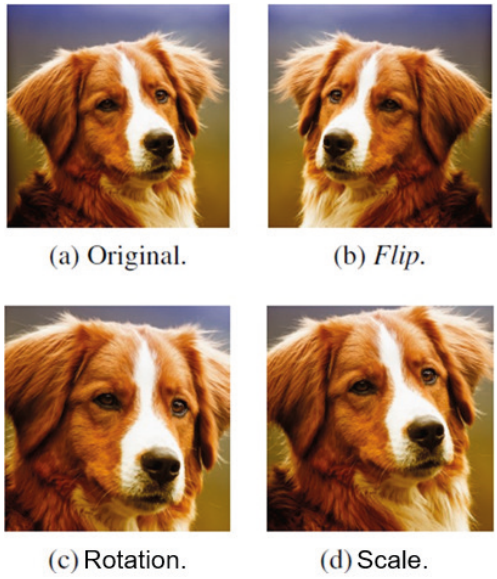

Figure 1.3: Illustrative example of a geometric transformation.

SPECIFIC METHODOLOGY

This section has been divided into three sections: Activities Foreseen in the Original Schedule (Section 3.1), Work Schedule (Section 3.2) and Activities Carried Out (Section 3.3).

ACTIVITIES FORESEEN IN THE ORIGINAL SCHEDULE

In order to achieve the objective proposed for this research, the following activities were proposed:

- **A1 - Verification of the state of the art**: Study of existing work in the literature, through *surveys*, *journals* and articles. There are two main objectives of the literature review: *(i) to* deepen knowledge of techniques and concepts in *data augmen-*

tation and *(ii) to* identify gaps. In addition, regular meetings were held with the supervisor;

- **A2 - Participation in the Meeting of Scientific Initiation and Meeting of Initiation in Technological Development and Innovation**: Participation consolidates the university's mission to qualify undergraduates and enable students to engage in scientific research;

- **A3 - Obtaining a dataset and studying it**: The aim of this stage is to obtain a dataset of skin lesions. Initially, the aim is to use the following HAM-10000 dataset [7] (International Skin Imaging Collaboration) and also to study this dataset. In this case, it is to understand if the data is unbalanced and if there is missing data, among other studies;

- **A4 - Submission of the Partial Report**: Writing of the Partial Report detailing the activities carried out up to the time of writing;

- **A5 - Applying *data augmentation* techniques to images**: This stage aims to apply traditional *data augmentation* techniques, for example geometric transformations (*flipping*, rotation and translation), cropping, scaling and shearing. It is worth noting that this technique will be applied to train a convolutional neural network;

- **A6 - Quantitative and** qualitative **evaluation**: Quantitative and qualitative evaluations will be carried out in order to assess the solution proposed in this research plan. In this way, the choice of characteristics to be evaluated will be defined, and it will also be examined which of them can influence the solution's performance;

- **A7 - Writing a scientific article**: The results obtained during the Scientific Initiation period will be published at scientific events. The solutions will be improved based on *feedback* from regular meetings, comments from reviewers and the state-of-the-art solutions themselves;

- **A8 - Submission of the Final Report**: Writing of the Final Report detailing all the activities carried out during the Scientific Initiation period.

WORK SCHEDULE

Table 3.1 shows the timetable for the original activities in the Research Plan, which were followed during the course of the work. The filled-in circles refer to the activities carried out.

ACTIVITIES CARRIED OUT

Activity 1 (A1) was successfully completed, adding knowledge about the applications of the various *data augmentation* techniques and the necessary and appropriate changes to include such data in the training data.

While Activity 1 (A1) and Activity 3 (A3) were being carried out, Activity 2 (A2) was carried out. This work was presented in the poster section of the XXVIII Meeting of Scientific Initiation and XIII Meeting of Initiation in Technological Development and Innovation, on October 24, 2023.

Activity 3 (A3) has been completed, with data obtained from HAM-10000, a *dataset* specifically made up of carcinoma images. This is directly followed by Activity 4 (A4), which refers to the development of this partial report detailing the activities and their achievements.

Activity 5 (A5) was completed by manipulating the data from the HAM-10000 dataset, applying traditional *data augmentation* techniques such as geometric transformations and brightness transformations, preparing the data in an appropriate way for training a neural network.

		2023		2024			
		Sep.~ Oct.	Nov.~ Dec.	Jan.~ Feb.	Mar.~ Apr.	May.~ Jun.	Jul~ Ag.
Activities	**A1**	●	●	●	●	●	◖
	A2	●					
	A3	●	◖	◖			
	A4				●		
	A5			●	●	◖	
	A6				●	●	◖
	A7					●	◖
	A8						●

Table 3.1: Scientific Initiation work schedule.

	lesion_id	image_id	dx	dx_type	age	sex	localization
0	HAM_0000118	ISIC_0027419	bkl	histo	80	male	scalp
1	HAM_0000118	ISIC_0025030	bkl	histo	80	male	scalp
2	HAM_0002730	ISIC_0026769	bkl	histo	80	male	scalp
3	HAM_0002730	ISIC_0025661	bkl	histo	80	male	scalp
4	HAM_0001466	ISIC_0031633	bkl	histo	75	male	ear

Table 4.2: Sample of the first 5 HAM10000 metadata.

Activity 6 (A6) was completed by training ResNet50 models with different data, training with both augmented and un-augmented data, and analyzing the results, both overall accuracy and Melanoma hit rate.

Activity 7 (A7) is being developed, based on the results generated by the experiments and the *feedback* received.

Activity 8 (A8) was completed as the other activities progressed, ending with the delivery of the report.

RESULTS OBTAINED

This section is divided into 5 subsections, Analysis of the HAM-10000 Set Data (Subsection 4.1), *Data Augmentation* Techniques Applied (Subsection 4.2), Balancing the HAM-10000 Set Data (Subsection 4.3), Architecture used in the experiments (Subsection 4.4) and Analysis of the Training Results (Subsection 4.5).

DATA ANALYSIS OF THE HAM-10000 SET

HAM-10000 is a high-quality dermoscopic image dataset launched in 2018 with the aim of helping to develop and evaluate automatic melanoma diagnosis algorithms. The dataset consists of 10,000 clinical images 7 different types of dermatological lesions. Table 4.1 shows the classes and their respective frequencies of separate samples in the training and validation sets.

Classes	Training	Test	Total
Melanocytic nevi	4690	2015	6705
Melanoma	770	343	1113
Benign keratosis	700	399	1099
Basal cell carcinoma	357	157	514
Actinic keratosis	227	100	327
Vascular lesions	100	42	142
Dermatofibroma	81	34	115
Total			**10015**

Table 4.1: HAM-10000 *dataset* classes and their respective frequencies.

In addition to the images, the dataset also has metadata associated with each image, such as the patient's age, gender and the location of the lesion in the human anatomy. Table 4.2 shows how the metadata is arranged in the dataset.

The distribution of classes in the dataset is uneven, as can be seen in Table 4.1. In a dataset for classification tasks, uneven class distribution occurs when one class is much more frequent than the others. This can have a substantial impact on the learning process of the classification model, resulting in possible complications such as:

- **Biases in the Model:** The model may present artificial results, overestimating the probability of the majority class and underestimating the probability of the minority classes;
- **Difficulty in learning minority classes:** The model may have difficulty learning the distinguishing characteristics of minority classes due to the smaller amount of data available.

APPLIED *DATA AUGMENTATION* TECHNIQUES

To effectively address class imbalance, one of the strategies widely used is the data *augmentation* technique. This approach aims to balance the distribution of samples between the different classes in sets of disproportionately represented data. *Data augmentation* consists of generating new data samples using techniques such as rotation, magnification, cropping and other image processing transformations in order to increase the diversity of the training set. By applying these modifications in a controlled manner, the aim is to strengthen the classification model'ability to deal with minority classes, promoting more robust and generalized learning.

As planned, the data from the HAM-10000 set was obtained and a preliminary analysis of the images in the data set was carried out in order to study which *data augmentation* methods could be used on the images in the set without affecting the accuracy of the model to be trained. The methods chosen because they would not harm the model's accuracy were brightness manipulation, image orientation manipulation, image saturation manipulation and scale manipulation, as shown in Figure 4.1.

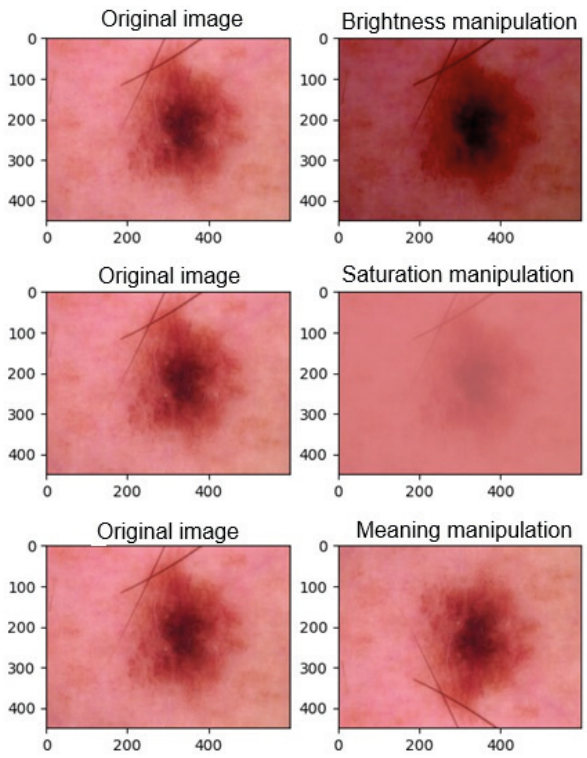

Figure 4.1: Demonstration of the brightness, saturation and upside-down transformations.

As you can see from the manipulations in Figure 4.1, it was possible to obtain three variations of the same image that can be used to improve the model's accuracy.

BALANCING THE HAM-10000 SET DATA

For the purposes of this comparison between training with and without *data augmentation*, and based on the analysis carried out in section 4.1, a significant imbalance in the majority class can be seen of the data set, as illustrated in Figure 4.2.

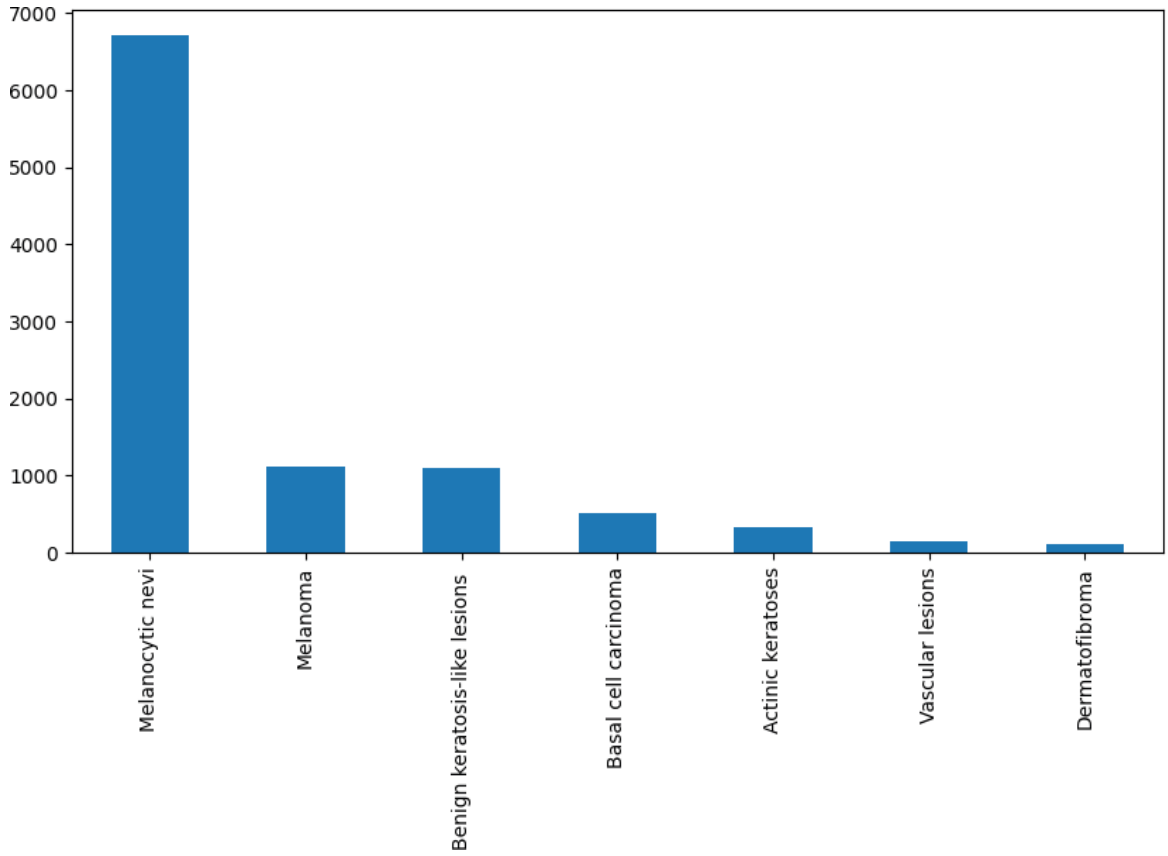

Figure 4.2: Graphical illustration of the count of items present in each class in the HAM-10000 set.

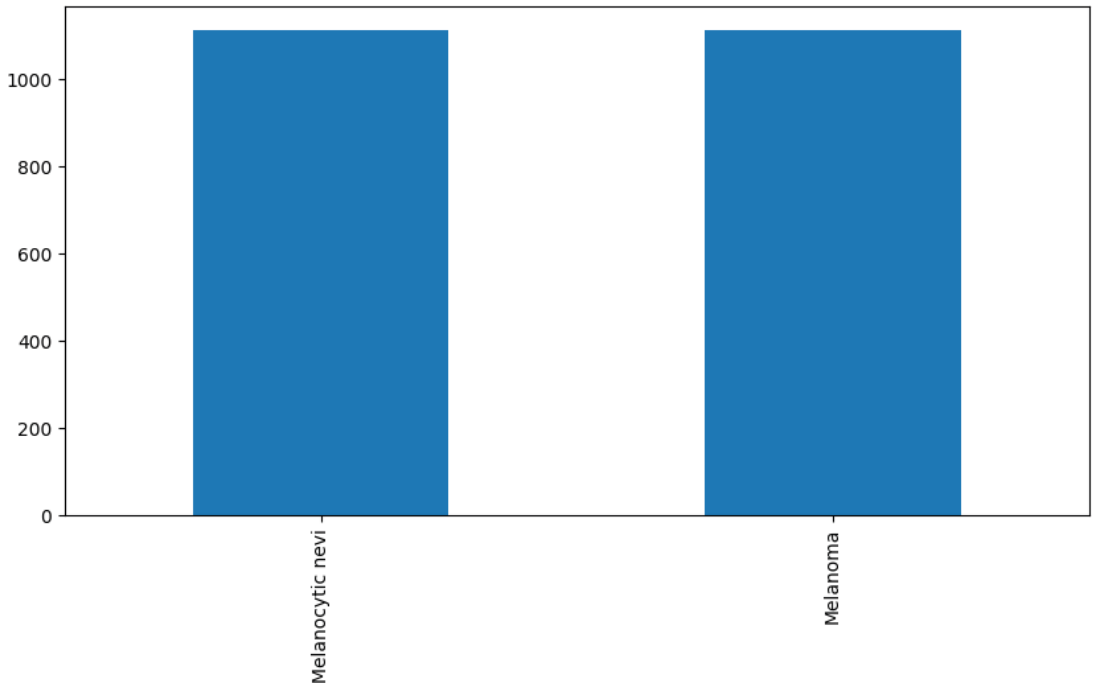

Figure 4.3: Graphical illustration of the count of items present after processing of the data.

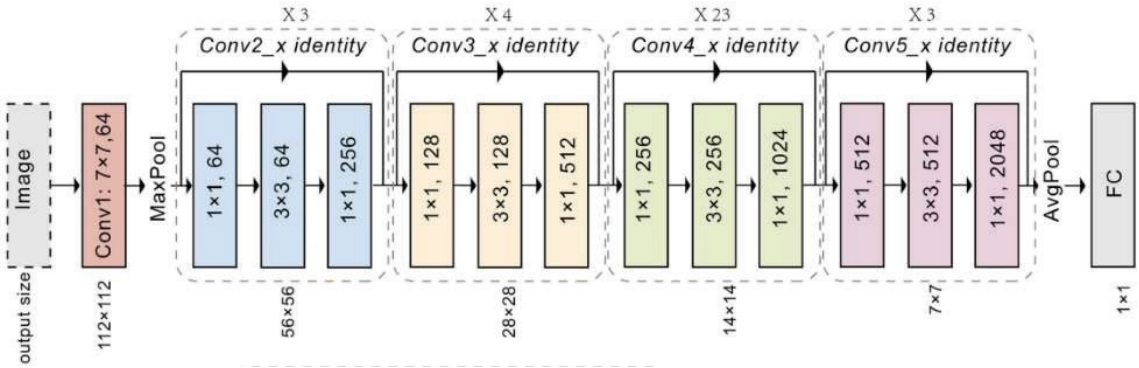

Figure 4.4: Illustration of the general design of the ResNet architecture. [9]

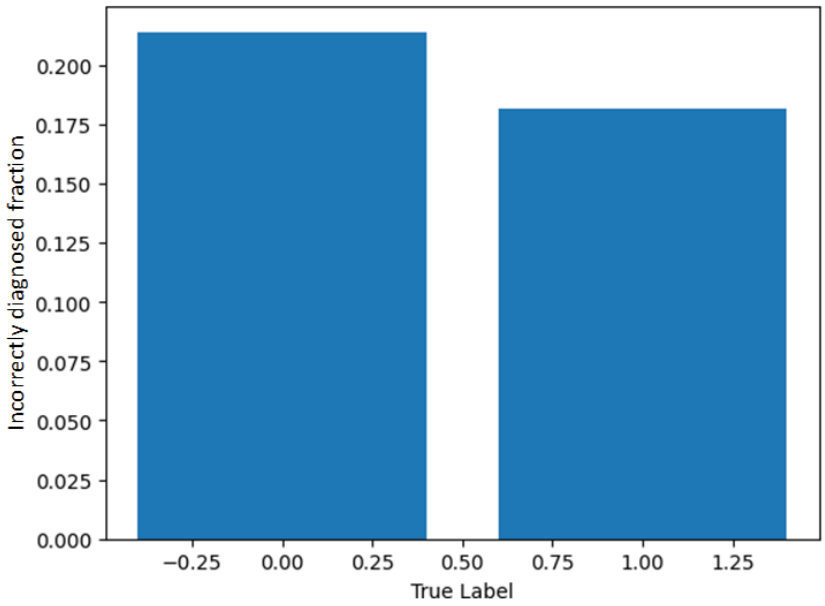

Figure 4.5: Graphical illustration of the classification error rate for the model trained without the augmented data.

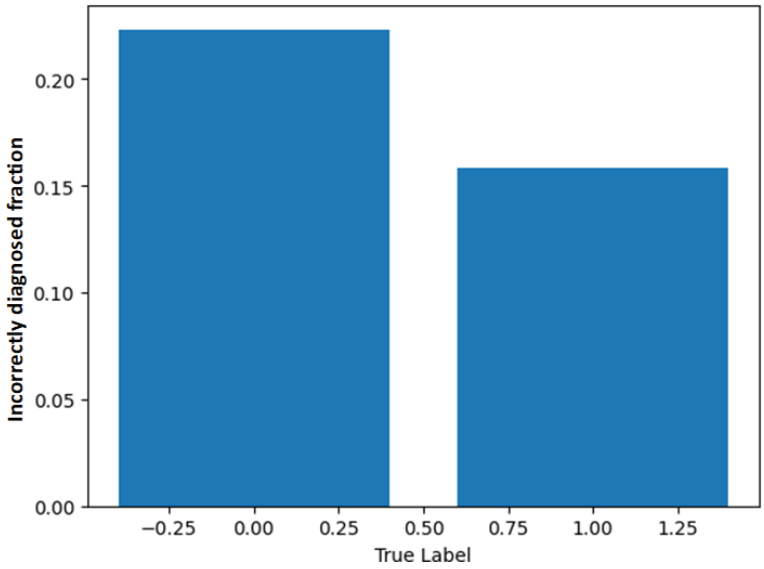

Figure 4.6: Graphical illustration of the classification error rate in the model trained with the augmented data.

For training without *data augmentation*, the data in the *Melanocytic nevi* class was *undersampled* so that the amount of data in it was equal the amount of data in the Melanoma class, and the other classes were discarded from the training dataset, as shown in Figure 4.3.

In the case of training with *data augmentation*, another adjustment was made: the other classes were discarded, but instead of reducing the number of data from the majority class, the *data augmentation* techniques discussed were used to increase the number of data to match the data from the majority class.

ARCHITECTURE

The architecture being used is ResNet (*Residual Network*), more specifically ResNet50, a 50-layer version of this architecture. It has been pre-trained with ImageNet to speed up training using *transfer learning*.

ResNet is a deep neural network that excels in image classification and brings as its main innovation the concept of residual blocks, which allow the model to learn mapping identities through residual connections, which allow the input of a layer to be passed directly to a layer further on in the network, skipping one or more intermediate layers, illustrated in Figure 4.4. In the case of ResNet50, these jumps have a distance of 3, advancing 3 blocks ahead with each jump. The use of residual blocks and connections provides more efficient training and prevents the model from degrading as the number of layers increases [8].

ANALYSIS OF TRAINING RESULTS

For the purposes of this analysis, three training sessions were carried out with different data: one with the HAM-10000 data set without any changes to its original data, another with the data balanced so that the number of elements in both classes was equal, and finally, a training session in which the Melanoma class data was artificially *augmented* using *data augmentation*.

After training, it can be seen that the model trained with the unchanged data set obtained an overall accuracy of 68.16%, the model trained with the balanced data set obtained an overall accuracy of 68.16%.

The model trained using artificially *augmented* data obtained an overall accuracy of 79.6%, and the model trained using artificially *augmented data* obtained an overall accuracy of 81.2%.

In addition to this general accuracy data, in the model trained using the augmented data, it was possible to observe a 2.5% improvement in the hit rate for Melanoma images (81.7% hit rate without data augmentation to 84.2% hit rate with data augmentation), compared to the model without the augmented data, as can be seen in Figure 4.4, which illustrate .5 5 and Figure .6the result for images labeled Melanocytic nevi and Melanomaand Figure 4.6 which illustrate the classification error rate, representing the result for images labeled *Melanocytic nevi* and Melanoma, the data being on the left and right respectively. For comparison purposes, the hit rate of the unbalanced and unenhanced model in the Melanoma and *Melanocytic nevi* cases was 65% and 30% respectively.

DISCUSSION

As the process of obtaining these carcinoma images with their appropriate classification is laborious and costly, the application of *data augmentation* techniques offers the possibility of improving the accuracy of the model without the additional cost of obtaining new data, which needs to be collected from the appropriate patients and validated by doctors. Another point consider when using *data augmentation* techniques is the appropriate use of these technologies. Certain manipulations of the data set can result in new images that bear little resemblance to the original image, which can have a negative effect on the accuracy of the model, as in Figure 5.1 below, where the augmented images show an extreme divergence from the original image.

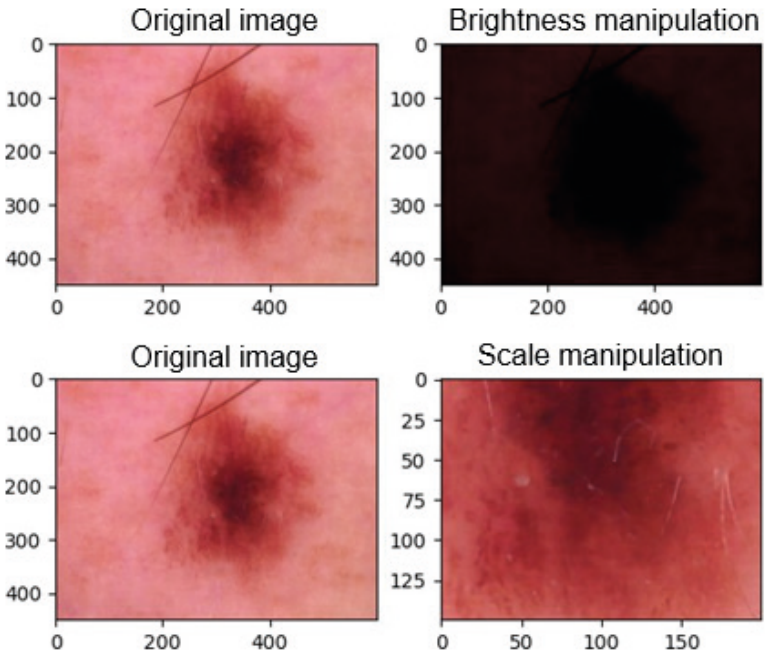

Figure 5.1: Demonstration of the transformations that could jeopardize the model's accuracy.

The entire implementation proposed in the research project is being carried out on the Google Collaboratory[1] platform with the help of the PyTorch[2] and scikit-learn[3] libraries in the Python language.

It should be noted that these libraries are widely used in the field of machine learning, especially in the context of deep neural networks.

When analyzing the results, it was possible to see an increase in the hit rate of the Melanoma images mentioned of 2.5%, which although it may seem like a small increase, in the medical context as in the identification of dermoscopic images, it is a notable increase, given that data to train dermoscopic image models is difficult to acquire, having to go through the process of image collection and image labeling by a medical specialist in the area, which is also a time-consuming and costly process.

CONCLUSION

Manipulating the *data* set using *data augmentation* techniques makes it possible generate several variations from a limited set of training images. These variations have the potential to improve the model training process by providing a wide range of examples for learning purposes.

By generating several variations from a restricted set of training images, these techniques substantially increase the diversity of the data available. These variations provide the model with an extensive range of examples to assimilate, increasing its capacity for generalization and strengthening its robustness.

Although a 2.5% increase in accuracy may seem small, it can be significant depending on the context, especially in critical areas such as Melanoma detection, where even small improvements can have a major impact on clinical practice, where acquiring data to train a dermoscopic model is very difficult.

1. https://colab.research.google.com
2. https://pytorch.org
3. https://scikit-learn.org

REFERENCES

[1] T. M. Mitchell and T. M. Mitchell, *Machine learning*, vol. 1. McGraw-hill New York, 1997.

[2] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, *Machine learning: An artificial intelligence approach*. Springer Science & Business Media, 2013.

[3] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[4] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621*, 2017.

[5] A. Mikołajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," in *2018 international interdisciplinary PhD workshop (IIPhDW)*, pp. 117–122, IEEE, 2018.

[6] L. Taylor and G. Nitschke, "Improving deep learning with generic data augmentation," in *2018 IEEE Sympo- sium Series on Computational Intelligence (SSCI)*, pp. 1542–1547, IEEE, 2018.

[7] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source derma- toscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint ar- Xiv:1512.03385v1*, 2015.

[9] "Mastering resnet: Deep learning breakthrough in image recognition." https://www.ikomia.ai/ blog/mastering-resnet-deep-learning-image-recognition. Acessado: 01/09/2024.