Universidade de Brasília - UnB
Faculdade UnB Gama - FGA
Software Engineering

Skin Lesions Classification Using Convolutional Neural Networks in Clinical Images

Author: Danilo Barros Mendes
Advisor: Dr. Nilton Correia da Silva

Brasília, DF, Brazil
2018

Danilo Barros Mendes

Skin Lesions Classification Using Convolutional Neural Networks in Clinical Images

Work submitted to the undergraduate course in Software Engineering of the University of Brasília, as a partial requirement to obtain a Software Engineer Bachelor's Degree.

Universidade de Brasília - UnB

Faculdade UnB Gama - FGA

Supervisor: Dr. Nilton Correia da Silva

Brasília, DF, Brazil

2018

Danilo Barros Mendes
Skin Lesions Classification Using Convolutional Neural Networks in Clinical Images/ Danilo Barros Mendes. – Brasília, DF, Brazil, 2018-
128 p. : il. (some color.) ; 30 cm.

Supervisor: Dr. Nilton Correia da Silva

Final Bachelor's Project – Universidade de Brasília - UnB
Faculdade UnB Gama - FGA , 2018.

1. Neural Networks. 2. Skin Lesion. I. Dr. Nilton Correia da Silva. II. Universidade de Brasília. III. Faculdade UnB Gama. IV. Skin Lesions Classification Using Convolutional Neural Networks in Clinical Images

CDU 02:141:005.6

Danilo Barros Mendes

Skin Lesions Classification Using Convolutional Neural Networks in Clinical Images

Work submitted to the undergraduate course in Software Engineering of the University of Brasília, as a partial requirement to obtain a Software Engineer Bachelor's Degree.

Approved Final Project. Brasília, DF, Brazil, December 10th, 2018:

Dr. Nilton Correia da Silva
Advisor

Dr. Fabricio Ataides Braz
Invitee 1

Dr. Cristiano Jacques Miosso
Invitee 2

Brasília, DF, Brazil
2018

Acknowledgements

I am thankful, first of all, for all the help and support I got from my family and friends. Specially my girlfriend, Taís Goulart, who supported – in more ways and times that I can account for – and accompanied me in this crazy ride called university from start to its end. I am very grateful for every step we took together. My mother, who have always been with me supporting, caring and giving me strength even in the small things. My father, that never failed to encourage me in academic endeavours, giving me all the opportunities to go forward. My sister, that were always by my side, making me laugh and teaching me to not take life too seriously.

I want to thank all my friends – Rafaels, Tito, Matbriel, Dudubels, Thiago, Bosorio, Mazin, André, Lud, Hannah, Milenas, Falcão – who have experienced this journey with me, for the company, amazing experiences, knowledge and support in good and difficult times. I have learned most of what I know alongside you. Thank you!

Lastly, I want to thank all the great professors that made me the professional I am today. You have the greatest and most responsible job, to shape and create critical thinking professionals that can grow for themselves. Thank you Milene Serrano, Elaine Venson, Hilmer Neri and Carla Rocha. A special thanks to my advisor, Nilton Correia, for all the knowledge, support and trust deposited in me.

Thank you!

"Time is an illusion that help things makes sense
So were always living in the present tense
It seems unforgiving when a good thing ends
But you and I will always be back then
[. . .] Will happen happening happened
(BMO, Adventure Time)"

Abstract

Skin lesions are conditions that appear on a patient due to many different reasons. One of these can be because of an abnormal growth in skin tissue, defined as cancer. This disease plagues more than 14.1 million patients and had been the cause of more than 8.2 million deaths, worldwide. Furthermore, a solution capable of aiding early diagnosis may save lives and cut costs in treatment. Therefore, this work proposes the construction of a classification model for 12 lesions, being 4 of these malignant, including Malignant Melanoma and Basal Cell Carcinoma. Furthermore, we use a pre-trained ResNet-152 architecture, which then was trained over 88,090 augmented images, using different transformations. The predictions were then analyzed with GradCAM method, to generate visual explanations, which were consistent with a prior belief and general good practices for explanations. Finally, the network was tested with 956 original images and achieve an area under the curve (AUC) metric of 0.96 for Melanoma and 0.91 for Basal Cell Carcinoma, that is comparable to state-of-the-art results.

Key-words: neural networks. skin lesion. classification.

Resumo

Lesões de pele são condições que aparecem em um paciente devido a várias razões. Uma delas pode ser por causa de um crescimento anormal no tecido da pele, definido como câncer. Essa doença aflige mais de 14,1 milhões de pacientes e tem sido a causa de mais de 8,2 milhões de mortes no mundo todo. Sendo assim, uma solução capaz de ajudar no diagnóstico precoce pode salvar vidas e diminuir custos de tratamento. Visto isso, é proposto a construção de um modelo de classificação para 12 lesões, sendo dessas 4 malignas, incluindo Melanoma Maligno e Carcinoma Basocelular. Além disso, neste trabalho é utilizado uma arquitetura ResNet-152 pré-treinada, que então foi aprimorada com 88,090 imagens aumentadas, utilizando diferentes transformações. As predições foram então analizadas com o método GradCAM para gerar explicações visuais, que foram condizentes com conhecimentos prévios e boas práticas para explicações. Finalmente, a rede foi testada com 956 imagens e alcançou a métrica de área abaixo da curva (AUC) de 0.96 para Melanoma e 0.91 para Carcinoma Basocelular, comparaveis aos resultados de estado da arte.

Palavras-chaves: redes neurais. lesões de pele. classificação.

List of Figures

List of Tables

List of abbreviations and acronyms

BCC	Basal Cell Carcinoma
SCC	Squamous Cell Carcinoma
US	United States
AI	Artificial Intelligence
XAI	Explainable Artificial Intelligence
SVM	Support Vector Machine
DNN	Deep Neural Network
CNN	Convolutional Neural Network
RGB	Red Green Blue
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve

Contents

Introduction

Today, skin cancer is a public health and economic issue, that for long years have been approached with the same methodology by the dermatology field (HAMBLIN; AVCI; GUPTA, 2016). This can be seen when we analyze the significant increase of cases diagnosed with skin cancer in the last 30 years (American Cancer Society, 2018). It is more troublesome when money comes in the equation, seeing that millions of dollars are being spent in the public sector (SOUZA et al., 2011). All this, to analyze a patient individually, the lesion and take action on the pieces of evidence seen. If any of these steps were to be optimized, it could mean a decrease in expenditure for the whole dermatology sector.

Moreover, the machine learning field is an area of knowledge, that studies and have the goal to build computers that are capable of learning. The knowledge learned, can be whether for the capacity to solve a problem, take decisions, etc (MURPHY, 2012). This field of work gathers both, statistical and the artificial intelligence roles. Furthermore, the field that applies machine learning to images is called computer vision, which in return, has many years of research in its history applied to the analysis of images with the purpose of selecting features capable of building a classification model.

Nowadays, with the invention, and validation, of the neural network (MCCULLOCH; PITTS, 1943; HAYKIN, 1999), the step of searching for features in images has been reduced to defining operations in a layer of a neural network. Thus, the task of image classification gained a powerful ally. Therefore, many kinds of research had been using this methodology to classify several types of imaging, including medical imaging (KRIZHEVSKY; SUTSKEVER; HINTON, 2012; MATSUNAGA et al., 2017). These researches brought down many barriers that though to be too much complex to be solved in the near future.

Furthermore, recently some advances have been made in the subject of skin lesion classification, using techniques based on deep neural networks. The results generated from this were impressive and with it, many other barriers had been brought down. However, the full problem its far from being solved, but once found the solution it can mean a revolution in the dermatology field.

An automatic tool that is capable of detecting and correctly classify a skin lesion, can save lives, as it may shorten the development of a disease and increase treatment's effects. Also, there may be many more casualties that happen due to skin cancers, we just do not have the means to reach these people that may live on far and recluse locations. With the ubiquity and advances of technology this tool may reach the hands of patients that doctors can not. Furthermore, this technology has the power to alert people of the

gravity in their skins and where they should seek a doctor to begin the treatment, thus saving unsuspecting lives.

Therefore, an automatic tool that aid in the early diagnosis of skin lesions, especially skin cancer, may save countless lives. With the use of such a tool, it is possible to have a vicious cycle that raises awareness about skin cancer and spread the use of it to distant populations. Another use of this same tool is in the hand of doctors that want to increase their diagnosis and decision making effectiveness. Where such a tool would help doctors make a more informed decision, where the patient does not need to go through many exams and trial of medicines to finally found out which lesion is being analyzed. Furthermore, this tool can be used in scenarios where it is accessible in smartphones, accessible in websites or as self-check stations in hospitals. Therefore, the solution may have many forms, deploying value to patients and doctors in many scenarios.

Seeing the problems involved in diagnosing skin lesions, this work envisions to create a learning model to classify skin lesions in one of 12 conditions of interest. With this purpose, the classifier aims to correctly distinguish lesions analyzing clinical images with the condition. Furthermore, this can prove to be a useful tool to aid patients and doctors on a daily basis operation.

The related work on this field proved that there are many algorithms capable of tackling this problem, but there is an astonishing difference between shallow and deep methods in machine learning. With that in view, this work will guide its efforts in using deep neural networks to achieve its main objective. For this to happen, the gathering of good practices and techniques used to approach classification of clinical images is needed.

Work organization

This work is divided in four chapters, Background, Methods and Materials, Results and Conclusion. The first chapter, Background, does an introduction about the background around dermatology from the past until nowadays, explains what are skin lesions, as well as discusses the related works on classifying lesions. The chapter Methods and Materials describes what are the techniques and resources, with emphasis on the methods that make viable the training of deep neural networks on a small and specialized dataset. The chapter Explainability & Interpretability elucidates what is explainable artificial intelligence, what is a good explanation as well as how to produce one with which methods. The chapter Results discuss the partial results obtained in this work, going through the difficulties encountered, the experiments made, and the results analysis. Finally, in chapter Conclusion, the final thoughts regarding the material shown in this work are discussed and possible future works are exposed.

1 Background

Medicine is a science and field of practice that devotes to diagnosing, treating and preventing illnesses. It has been around for hundreds of centuries, when it was considered an art, divine gift and sometimes involved in mysticisms. Was common to see a practitioner of medicine apply herbs to an unhealthy person and say some prayers as it was being done. As humankind evolved, the practices and knowledge evolved and became more available for the society as a whole, demystifying the science. Contemporary medicine is applied in a much more ample field, gaining space in genetics, biomedicine, drug discovery, technological development of medical devices, neurology, dermatology and many more (Association of American Medical Colleges (AAMC), 2018).

Medicine is always evolving and has always benefited from advances in technology, that was no different when photography and cameras were invented. The capture of images helps doctors register, analyze and diagnose multiple diseases, and that technique may be referred as biological imaging. This technique may include examples of molecular imaging, radiography, magnetic resonance imaging and medical photography. The latter is an area of photography specialized in the documentation of patients and its clinical representation, surgical procedures, devices and specimens for autopsy (LARSSON; BRANE, 2007). Although this technique was first used in medicine in 1840 by Alfred François Donné for photographing teeth and bones (DONNÉ, 1845), it has been used extensively since then. However, was only in the early 19th century that medical imaging was first used in dermatology (NEUSE et al., 1996).

1.1 Skin Lesions

Dermatology is one of the most important fields of medicine, with the cases of skin diseases outpacing hypertension, obesity and cancer summed together. That is accounted because skin diseases are one of the most common human illness, affecting every age, gender and pervading many cultures, summing up to between 30% and 70% of people in United States. This means that in any given time at least 1 person, out of 3, will have a skin disease (BICKERS et al., 2006). Therefore, skin diseases are an issue on a global scale, positioning on 18th in a global rank of health burden worldwide (HAY et al., 2014).

Dermatology is a field that heavily relies on visual observation of the skin. That can be credited to the domain itself, as verbal descriptions cannot characterize well a lesion. Therefore, practitioners were conducted to examine patients analyzing their skin with the naked eye to diagnose the condition. In 1572, was published the first scientific work to be considered focused on dermatology (SIRAISI, 2003). That predates any means

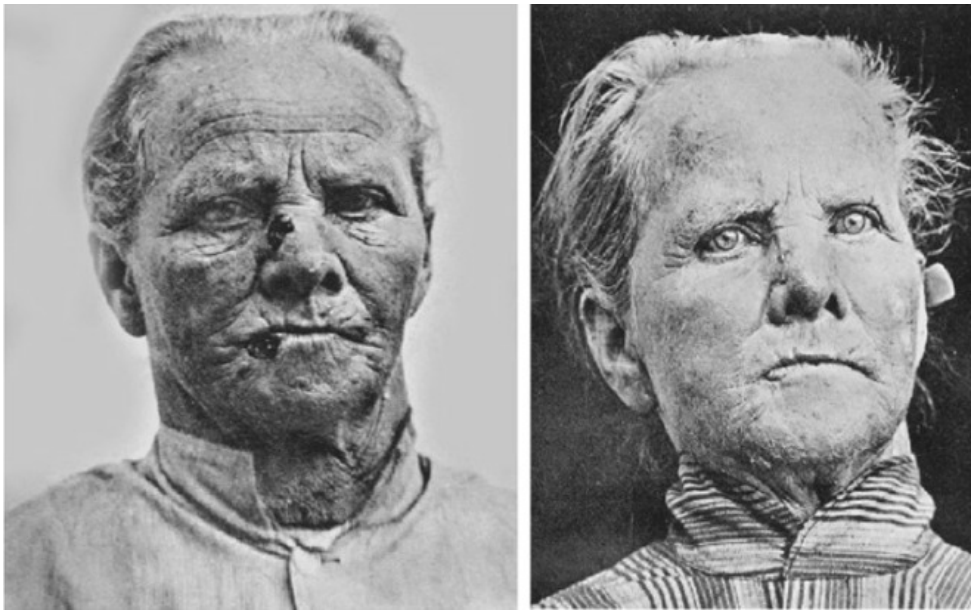


Figure 1 – Documentation of treatment for non-melanoma skin cancer around 1900.

Source – Prime (1900).

of photography, thus the science heavily relied on visual observation in its origins. Aside from visualizing the condition on-site, the practitioners had other two resources of communicating and illustrating these visual images, through drawings that required skilled artists and reproductions of the illness in wax sculptures.

Was only around 1900 that medical imaging was first used in a scientific publication (Figure 1). From there on out, the opportunities to capture images for medicine and dermatology increased significantly. For dermatology was above all the rest, since the object of study (the skin) is much more accessible, thus the field benefited from this technique immensely. After that, pictures played a major role in education as well, as it was possible not only to print the images but to present them in a classroom and in conferences to promote discussions.

With the advances of technology and the rise of the personal computers, every dermatologist with an affordable camera and a computer have a way to store images and further analyze them (HAMBLIN; AVCI; GUPTA, 2016).

Furthermore, medical imaging can show itself as a resource of high value, as dermatology has an extensive list of illness that it has to treat. In addition, the field has developed its own vocabulary to describe these lesions. However, verbal descriptions have their limitations and a good picture can replace successfully many sentences of description and is not susceptible to the bias of the message carrier.

Moreover, the recommended way to detect early skin diseases is to be aware of new or changing skin growths (American Cancer Society, 2017). This, support the idea that skin cancer often is detectable through naked eye and medical photography. However,

Figure 2 – Modern dermatoscope with double polarized light made by 3GEN.

Source – Northerncedar (2009).

these techniques are usually not used to diagnose a patient with an illness or not. There are many apparatus and new technologies that aid the diagnoses of skin diseases, including epiluminescence microscopy (dermoscopy), digital dermoscopy, confocal microscopy, infrared imaging and multispectral imaging (CELEBI; STOECKER; MOSS, 2011). From this list, dermoscopy is the most popular among the specialists. It is a noninvasive tool that enables to examine the morphological features beneath the first skin layers. The most modern versions, generally, consists of a magnifier, a polarised light source and a liquid between the instrument and the patient skin, that allows the light to be less reflective on the skin, as seen on Figure 2.

Although these technologies are capable of aiding the diagnoses of skin cancers, they are often expensive and/or requires extensive training in using them. Analysis with the naked eye is still the first resource used by specialists, along with techniques such as ABCDE, that consists of scanning the skin area of interest for asymmetry, border irregularity, uniform colors, large diameters and evolving patches of skin over time (NACHBAR et al., 1994). In this way, the analysis from medical images is analogous to the analysis with the naked eye and thus can be applied the same techniques and implications.

Contrary to the fact that medicine is always advancing, the humankind has never succeeded to mitigate cancer in all its forms. It has been a constant concern in all human

history (HAJDU, 2011), and a subject of fear and taboos. In 2012 there were 14.1 million[1] new cases of cancer and 8.2 million deaths worldwide (FERLAY et al., 2012). On these numbers, the types of cancer more incident were lung, breast, intestine and prostate (FERLAY et al., 2012). However, worldwide the most common case of cancer is skin cancer, been melanoma, basal and squamous cell carcinoma (BCC and SCC) the most frequent types of the disease (American Cancer Society, 2016). This type of the disease is most frequent in countries with the population with predominant white skin or in countries like Australia or New Zealand (STEWART; WILD et al., 2014).

In Brazil, it is estimated that for the biennium of 2018-2019, there will be 165,580 new cases of non-melanoma skin cancer (BCC and SCC mostly) (Instituto Nacional de Câncer José Alencar Gomes da Silva, 2018). Moreover, it is visible that the incidence of these types of skin cancer had risen for many years. This increase can be due to the combination of various factors, such as longer longevity of the population, more people being exposed to the sun and better cancer detection (American Cancer Society, 2016).

In the United States, the numbers add up to 9,730 deaths estimated for 2017 (American Cancer Society, 2017). Skin cancer accounts for more than 1,688,780 cases (not including carcinoma in situ, nor non-melanoma cancers) in the US alone in the year of 2017 (American Cancer Society, 2017).

Despite skin cancer is the most common type of cancer in the society, it does not represent a great death rate in its first stages, since the patient has a survival rate of 97%. However, if the patients are diagnosed in the later stages the 5-year survival rate decreases to 15%.

In Brazil, were expected to occur 114,000 new cases of non-melanoma skin cancer in 2010. From that, it was expected that 95% were diagnosed in early stages. However, even with early diagnosis this amount of cases means around R$37 million (Reais) to the health public system and R$26 million to the private system per year (SOUZA et al., 2011).

Seeing the applicability of medical photography on the detection of skin cancers, we can generalize that thought for all the illnesses and anomalies that may occur in the skin. That said, we can generalize any abnormality in the skin as a lesion, so a melanoma and a mole can be called lesions present in the skin.

1.2 Detecting skin lesions

The problems in detecting skin lesions accurately are that it exists many features and minutiae that need to be dealt with. For that task, many professionals train for part of

[1] Not including non-melanoma skin cancer

their life to specialize in detecting and differentiating these diseases. Is not uncommon that one may dedicate a lifelong effort to continually improve one's ability. However specialized one may be, one is still susceptible to human failures such as fatigue and mistakes. Another limitation that a specialist faces is the workload that is possible to take in any given day, for such is human nature to not be able to work several hours a day and process many pieces of information at a fast rate, not staggering once.

Thus many algorithms and tools have been created to aid these professionals in their task of detecting diseases in many fields (AERTS et al., 2014; ESTEVA et al., 2017; LEE et al., 2017; VANDENBERGHE et al., 2017; KERMANY et al., 2018). This has proven to add more reliability and confidence to doctors in their practices as they have more information to diagnose patients.

One subject and field that is growing more every day is artificial intelligence (AI), where it is revolutionizing many industries. Artificial intelligence has the power to reinvent the way we interact with each other and perform our tasks. As Ng (2017) stated once in his talk "Artificial intelligence is the new electricity".

The medical industry has the potential to become a major beneficed from this technology as it is abundant in data that need to be analyzed by humans, and many times the process of analyzing becomes mechanical. When this happens is a sign that artificial intelligence can play a big part. AI has the potential to analyze a lot of images and perform difficult classifications on it, helping the diagnosis of certain illness. Furthermore, detecting skin lesions are mainly done by scanning the patient with the naked eye and then execute different approaches to finally diagnose the patient. This expresses a major task of classification as the specialist tries to fit the lesion in a broad spectrum of possibilities, given only the symptoms and the appearance of the lesion in the skin.

Along with artificial intelligence came the rise of the data age, hundreds of new pieces of information and electronic data being generated. It became easier to utilize methods that leverage on hundreds of thousands of examples. The premise is that with enough data, we can learn how a domain behaves and what patterns it follows. And with this information, one can detect single patterns, predict future data and other outcomes (MURPHY, 2012). Those methods are known as machine learning algorithms.

Seeing this, a handful of approaches have been made to solve classifications problems in medical imaging. First, it was common to do this by image analysis, that consists of collecting handcrafted features from the images then classifying these features with some shallow machine learning algorithm designed for each specific class, finally try to classify the image itself (GOLDBAUM et al., 1996). This approach is exhausting as it requires many hours and high-level skills to achieve the results that were aimed at (CHAUDHURI et al., 1989).

For dermatology and skin lesions detection has not been different. History shows that many approaches had been made over the course of years, applications with shallow algorithms such as K-Nearest Neighbors (KNN) (BALLERINI et al., 2013) and Support Vector Machines (SVM) (GILMORE; HOFMANN-WELLENHOF; SOYER, 2010) had been proven to accomplish good results, but are as well tiresome to build applications that involve such approaches.

However, with the advancements in technology, it became easier to consume and compute more and more data. In addition, processing capacities became faster enabling feasible training time of algorithms that once took several days to test a hypothesis. This means that algorithms that were once costly and unfeasible to use, became reachable even to researchers with more modest equipment. Thus, became possible to compute more data with costly algorithms. And so, the long forgotten dream to build machines that learn became more tangible (EVANS; CARMAN; THORNDIKE, 2010).

Deep learning is the field that leverages the most from big data, it has the ability to process numerous examples and abstract patterns and general high-level abstractions. By building knowledge from data, it is possible to avoid the need to human operators to construct these concepts and knowledge in form of well-defined rules (GOODFELLOW et al., 2016). This solved part of a problem that was how to transfer informal and tacit knowledge to algorithms.

This field is based on the idea that given a layer of nodes capable of executing a mathematical function, we can stack together these layers on top of each other and pipe input and output from one another through every layer until we are deep in the last one. Therefore the name deep learning is applied.

Seeing this, some researchers have been applying this approach to classifying skin lesions with success. One common thing in this domain is the lack of quality and scarcity of open data, it is common to see works with only a couple hundred of examples. That is a characteristic of the medical field. There are many hospitals and clinics that hold huge amounts of data and do not make it public mainly because of privacy issues with patients. However, many authors still apply efforts to push forward the technology in such fields, overcoming these barriers. For the purposes of this work, we listed some related researches that uses deep learning in dermatology, applying neural networks to skin lesions.

Matsunaga et al. (2017) proposed an approach to classify melanoma, seborrheic keratosis, and nevocellular nevus, using dermoscopic images. In their work, they proposed an ensemble solution with two binary classifiers, that still leveraged from age and sex information of the patients, if they were available. Furthermore, they utilized techniques of data augmentation, using a combination of 4 transformations (rotation, translation, scaling and flipping). For the architecture, they chose the ResNet-50 implementation on the framework Keras, with personal modifications. This model was pre-trained with the

weights for a generic object recognition model and finally used two optimizers AdaGrad and RMSProp. This work was then submitted to the ISBI Challenge 2017 and won first place, ahead of other 22 competitors.

Nasr-Esfahani et al. (2016) showed a technique that uses imaging processing as a previous step before training. This result in a normalization and noise reduction on the dataset, since non-dermoscopic images are prone to have non-homogeneous lightning and thus present noise. Moreover, this work utilizes a pre-processing step using *k-means* algorithm to identify the borders of a lesion and extract a binary mask, which the lesion is present. This is done to minimize the interference of the healthy skin in the classification. Furthermore, Nasr-Esfahani et al. (2016) used a technique called data augmentation to increase the dataset, using three transformations (cropping, scaling and rotation) and multiplied the dataset by a factor of 36 times. Finally, a pre-trained convolutional neural network (CNN) is used to classify between melanoma and melanocytic nevus for 200 epochs (20,000 iterations, using a batch size of 64 and a dataset with 6,120 examples).

Menegola et al. (2017) presented a thorough study for the 2017 ISIC Challenge in skin-lesion classification. In this work, it is presented experimentations with some pre-trained deep-learning models on ImageNet for a three-class model classifying melanoma, seborrheic keratosis, and other lesions. Models such as ResNet-101 and Inception-v4 were vastly experimented with several configurations of the dataset, utilizing 6 data sources for the composition of the final dataset. It was also reported the use of data-augmentation with at least 3 different transformations (cropping, flipping, and zooming). Also, it is reported that the points that were critical to the success of the project were mainly due to the volume of data gathered, normalization of the input images and utilizing meta-learning. The latter is elucidated as an SVM layer in the final output of the deep-learning models, that map the outputs to the three classes that were proposed in the challenge. Finally, this work won the first place in the 2017 ISIC Challenge for skin lesion classification.

Kwasigroch, Mikołajczyk and Grochowski (2017) present a solution similar to the previous 3. This is due to the inherent limits and problems that are existent in this domain, data scarcity. In this work transfer-learning is applied, using two different learning models, VGG-19 and ResNet-50, both pre-trained on ImageNet 1,000 classes dataset. These were used to classify between malignant and benign lesions, using 10,000 dermoscopic images. For the correct learning process, it was also used the up-sampling of the underrepresented class. This process was done using a random number of transformations, chosen between rotation, shifting, zooming, and flipping. Furthermore, in this paper, it was presented 3 experiments, first with the VGG-19 architecture with the addition of two extra convolutional layers, two fully connected layers, and one neuron with a sigmoid function. Second it experimented with the ResNet-50 model, and finally a implementation of VGG-19 with

an SVM classifier as the fully-connected layer. As a final result, the modified implementation of the VGG-19 had the best results. However, the main reason for the poor results in the ResNet-50 model was due to the small amount of training data. Maybe with larger amounts of data, it would be possible to train a small model and produce better results.

Esteva et al. (2017) presented a major breakthrough in the classification of skin lesions. This research compared the result of the learning model with 21 board-certified dermatologists and proven to be more accurate in this task. It was performed to classify clinical images, indicating whether a lesion is a benign or malignant one. For this result were used 129,450 images, consisting of 2,032 different diseases and including 3,372 dermoscopic images. Furthermore, it was used a data-augmentation approach to mitigate problems as variability in zoom, angle, and lighting present in the context of clinical images. The augmentation factor was by 720 times, using rotation, cropping, and flipping. Here, an Inception-v3 pre-trained model was utilized as the main classifier, fine-tuning every layer and training the final fully connected layer. Moreover, the training was done for over than 30 epochs using a learning rate of 0.001, with a decay of 16 after every 30 epochs. The classification was done in such a way that the model was trained to classify between 757 fine-grained classes, and then as the probabilities were predicted it was fed into an algorithm that selected the two different classes (malignant or benign). Using this approach, this work achieved a new state of the art result.

Seog Han et al. (2018) proposed to classify the skin lesions as unique classes, not composing meta-classes such as benign and malignant. It used the ResNet-152 pre-trained on the ImageNet model to classify 12 lesions. However, for training was used other 248 additional classes, that were added to decrease the false positive and improve the analysis of the middle layers of the model. Furthermore, this was done in such a way that the train sampling for the 248 diseases did not outgrow the main 12, thus when used for inference the model predicted one of the 12 illness, even when the lesion does not belong to one of them. For training was used 855,370 images, augmented approximately 20 to 40 times, using zooming and rotation. These images were gathered from two Korean hospitals, two publicly available and biopsy-proven datasets, and one dataset constructed from 8 dermatologic atlas websites. Furthermore, the training lasted for 2 epochs using a batch size of 6 and a learning rate of 0.0001 without decay before 2 epochs. This early stopping was done to avoid overfitting on the dataset. Finally, it was reported that the ethnic differences presented in the context were responsible for poor results in different datasets, thus it was necessary to gather data from different ethnics and ages to correct mold the solution to reflect the real world problem present in skin lesions classification.

Finally, we can observe that every one of these works has one aspect in common, data scarcity. This is a characteristic of the medical domain, there are very few annotated examples of data that are publicly available. The works that proven to have

more impact had to collect data from other sources, mainly private hospitals or clinics. Furthermore, this step of data collection did not fully mitigate the problem, it was still necessary to use techniques such as transfer-learning (PAN; YANG, 2010; YOSINSKI et al., 2014) and data-augmentation (SIMARD et al., 2003; DYK; MENG, 2001; KRIZHEVSKY; SUTSKEVER; HINTON, 2012).

For this work, the research done by Esteva et al. (2017) and Seog Han et al. (2018) will be used as guidelines to construct and, if possible, further improve a skin lesion classifier. Taking into account that data is scarce in the medical field, this work will focus on the available datasets. Therefore, the lesions of interest for this work will be limited by the available data.

2 Methods and Materials

Seeing the past history of dermatology and the implementation of artificial intelligence in the field, it is safe to say that many characteristics have evolved and developed over the years. It is also correct to state that the current approach has had the most promising results, but it has a long way to go until it is viable to use inside clinics and at homes. For this purpose, it is necessary to understand the underlining concepts underneath the current state of the art.

Furthermore, this chapter will discuss these concepts in a succinct manner, going through the raw material needed to construct it, knowledge about the technology that is implemented in the field of machine learning, how to tie the technology and the data together in a medical field, such as dermatology, what is the current state of the industry of artificial intelligence, and finally, how to integrate everything in one piece.

2.1 Neural Networks

Have seen the possible approaches used in the past to handle the problem of classifying skin lesions, this work will focus on techniques that involve neural networks, more specifically deep neural networks (DNN). This is mainly because of the prominent results that it has achieved in the recent experiments done by Esteva et al. (2017) and Seog Han et al. (2018). However for this, its necessary to understand what is a neural network, what composes it, and what variants exist in the current state of the AI field.

2.1.1 Basic Concepts

Neural networks are algorithms in the field of machine learning, that emerged from biomimetic studies of neurons in the cerebral cortex (ROSENBLATT, 1958). These neurons are arranged in a way that they form a nervous system, similar to the one present in the human brain. And analogous to the biological neurons, the artificial one (also called *perceptron*) uses the input from the last neuron and propagate the signal forward (a forward pass). This process is done using the synapses that connect the neurons, forming a network.

However, the forward passing of the neurons is not just passing along the impulse. The neuron has the task to process this signal before doing so. For this to happen the neuron sums up all the inputs received and process it (as shown in Figure 3). If the resulting impulse is strong enough to activate the neuron, it passes forward the impulse to the other neurons, whose dendrites are attached to any of the axon terminals. This is

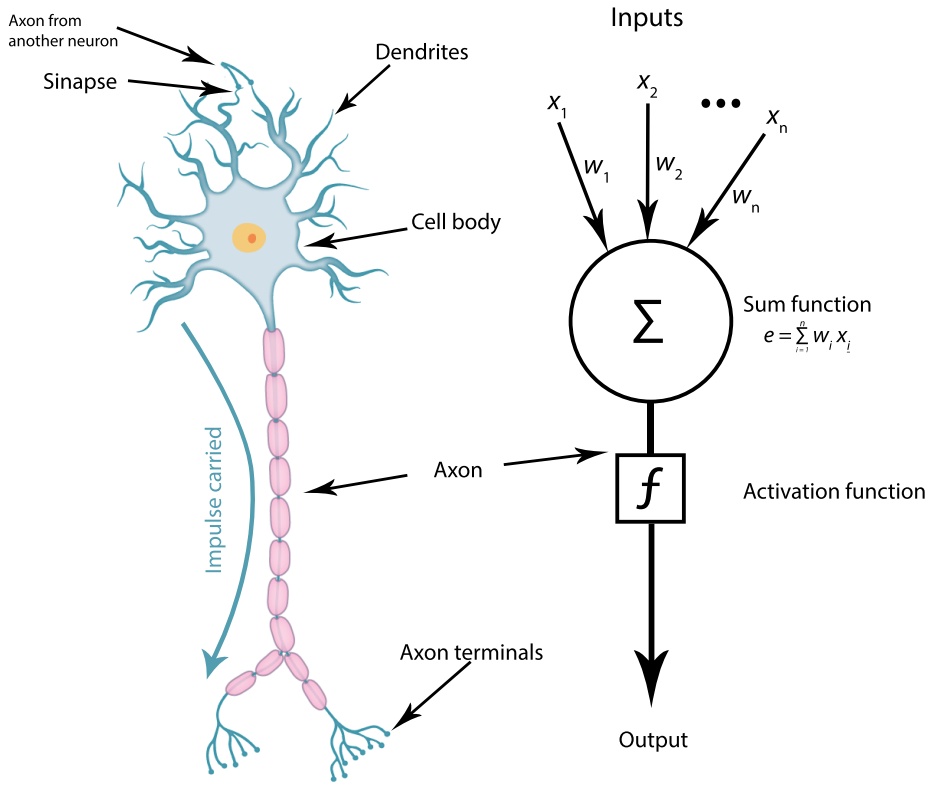


Figure 3 – Artificial neuron biomimetism.

Source – Author.

called the activation function of the neuron. Only when the requisites are met, the neuron is capable of producing a signal that is meant for the next neuron(s) in the network or is the final product. This simple arrangement of the neural system of the brain is responsible to conduct and perform extremely complex tasks. However, artificial neural networks have not come close to reproduce the total capability of the human brain.

The neural network technology has been around since the decade of 1940 when a simple network was modeled in an electrical circuit (MCCULLOCH; PITTS, 1943). In spite of this, the technology has not popularized until the year of 1982, when a dynamic feedback loop was reinvented using bidirectional lines connecting the neurons (WERBOS, 1982). Thus, the algorithm of backpropagation (or often called backprop) optimized the error term (cost) calculation and the propagation to the activations back through the network.

However, this optimization still did not make viable the training of these algorithms. The training process continued to be very computational costly. Was not until the advancements in software and hardware, such as parallel computing and graphics processing units, that the use of neural networks became more prominent. This is due to the fact that a fully connected (or dense) network has a quadratic number of connections

between layers of neurons, and thus makes the training slow for such a number of connections (HAYKIN, 1999). And is easy to increase in neuron numbers when it is necessary to compute many pieces of information, such as pixels in an image. A solution to this problem came with the invention of the convolutional neural network.

2.1.2 Convolutional Neural Network

With the research of Hubel and Wiesel (1968) on the animal visual cortex was possible to discover that the connections of these neurons were more sparse than those previously noted on the human brain. This led to an insight that even small regions on the visual space, the receptive fields, where responsible for generating the stimulus that triggered neurons individually.

Furthermore, this concept of sparse features makes the network to be more capable of recognizing individual features apart. This is due to the fact that more sparse correlation between neurons will create weaker links to characteristics that are distant from each other. This can be intuitively seen in a picture of a cat (Figure 4), as usually the top left corner of the image has nothing to do with the bottom right one, and both are different from the inner center of the image and hold no correlation between each other.

Figure 4 – A tabby cat.

Source – Jia et al. (2014a).

Although, this solution is fitted to process images, the assumption that the more sparse the features are, less correlation they will have, do not facilitate the use of this type of network in fields that have important relationships between features that are not spatially close. This restriction is often applied to images, as visual features only add semantic knowledge if they are spatially near, however other fields have been reported to benefit from this (ZHANG; LECUN, 2015).

When applied to images, this network has its neurons arranged in a three-dimensional manner $h \times w \times d$, such as that, it represents width, height, and depth respectively. Moreover, on this representation the convolution network does a linear operation called convolution (LECUN et al., 1989), that is where the name comes from. So, convolutional neural

networks (CNN) utilizes convolutions in place of the common matrix multiplication used in fully-connected networks in at least one of its layers (GOODFELLOW et al., 2016). Therefore, a typical CNN has a set of layers, normally composed of the convolutional layer, pooling layer, and a fully-connected layer. These layers are then stacked together to form a full convolutional neural network architecture.

2.1.2.1 Convolution

This layer is the most important layer in a convolutional neural network (hence the name). Furthermore, the convolution layer is responsible to compute and detect each specific characteristic in any point of the input. In general, the convolution is a mathematical operation used in the signal processing field. It is the integral of the product of two functions that operate in the real numbers, being one of them flipped horizontally. Furthermore, we can write the operation as a sum of the products of two matrices in an element-wise way.

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i+m, j+n)K(m,n) \tag{2.1}$$

Where in equation 2.1, I is the input (image), K is the kernel or filter, m and n being the height and width of the kernel matrix (usually $m = n$), and i and j the horizontal and vertical axis, respectively.

The kernel (or filter) is a set of weights that convolve through the whole depth of the input (usually the RGB channels in images) and act as small receptive fields and are responsible to establish a local connectivity relationship. These kernels are slid through the whole width and height of the input, which is done in the forward pass of the network (when the input is passed through all the layers until it reaches the output).

The use of these receptive fields makes it easier to spot visual patterns and ignore noise or otherwise disperse patterns that would influence negatively in the network ability to recognize patterns.

The output of the convolutional layer is dependent in some free parameters (or hyperparameters) that control how many neurons are in the output volume and how they are arranged. These hyperparameters are:

- **Depth**: The depth of the output volume corresponds to the number of kernels that are applied to the input. Thus, is the number of neurons that will be stimulated by the same receptive field, as seen in Figure 5. However, each one learns something different from the input.

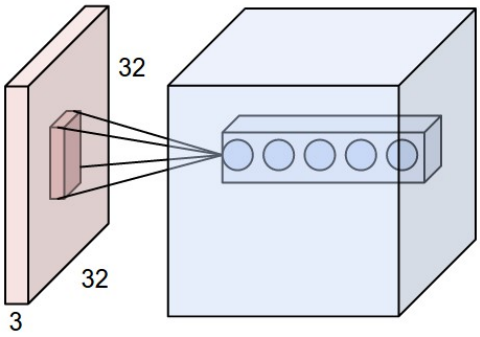


Figure 5 – Convolution applied to a input with a output of depth of 5 neurons.

Source – Karpathy (2018a).

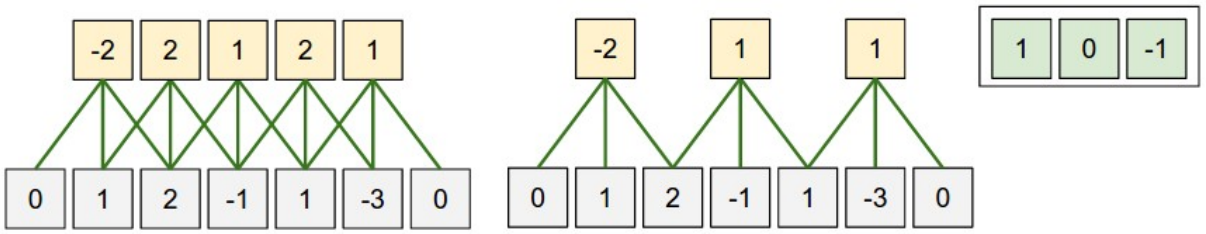


Figure 6 – Stride of 1 (left) and of 2 (right). Shared weights on top right ([1,0,-1]).

Source – Karpathy (2018a).

- **Size of the receptive field** (m and n): Is the size of the field that will be convolved. This is the same for every neuron in the same layer. Usually, the size is equal for both width and height, being 3×3 or 5×5 but it depends on the size of the input.
- **Stride**: This hyperparameter is responsible to specify the amount we move each filter until we apply a dot product again. With this, we define the gap between the overlap of receptive fields in each slide. Altering this value the size of the output alter accordingly, with a higher stride the smaller is the output produced spatially. As Figure 6 shows, the yellow neurons have a receptive field of size 3 (in one dimension only) and convolve an input (in green) of size 5 (with a zero padding of one). With the stride of two, the spatial size of the output is reduced from 5 to 3.
- **Zero Padding**: The zero padding determines the number of zeros that will be around the border of the input. This allows us to control the spatial size of the output. Usually, it is used along with stride to further determine the output.

Seeing this, we can compute the spatial size of the output using the knowledge of the hyperparameters along with the information of the input volume size. With these we can use the equation 2.2, where the input volume size (W), the receptive field size (F),

the amount of zero padding (P), and the amount of stride (S) are used.

$$\text{output\_size} = \frac{W - F + 2P}{S} + 1 \tag{2.2}$$

Furthermore, convolutional neural networks have a propriety that makes them even more efficient compared to the fully-connected ones. CNN's have what is called shared weights, the weights used in the kernels. These are learned and shared among the neurons on the same depth, reducing in several orders of magnitude the number of parameters. This reduces not only the memory footprint of the network, but also the time to train these parameters, further improving the performance of the update step.

2.1.2.2 Pooling

The pooling operation is usually used in regular intervals between convolution layers in a CNN. Moreover, pooling is responsible to generalize the position of the patterns often found by the convolution layers. This is done by reducing the spatial dimensions of the data, but not the depth.

Furthermore, this operation does a summary statistic of a spatial region of the output, this can be done using some operations. The most common are the maximum values, called max pooling, and the average of the values in the neighborhood, represented by the average pooling. In all cases, the pooling operation adds invariance to small translations in the input. This can be important if we care more about whether some feature is present in the input than where it is present. It also reduces the spatial information used (see Figure 7), further improving computation performance as the operations use fewer parameters.

The pooling layer also has hyperparameters, such as stride and size of the receptive field, that control its behavior. Along with these, the pooling function is determinant in the configuration of this layer. However, this layer does not have any parameters that are updated over the training process, it only implements a fixed function.

Furthermore, comparatively with the convolution operation, we can compute the spatial size of the pooling output with the equation 2.3. Where, H' and W' are the height and width of the output, respectively, H and W are the inputs, H_p and W_p are the filters height and width applied to the input. The output depth, however, stays the same.

$$H' = \frac{H - H_p}{S} + 1 \qquad \text{and} \qquad W' = \frac{W - W_p}{S} + 1 \tag{2.3}$$

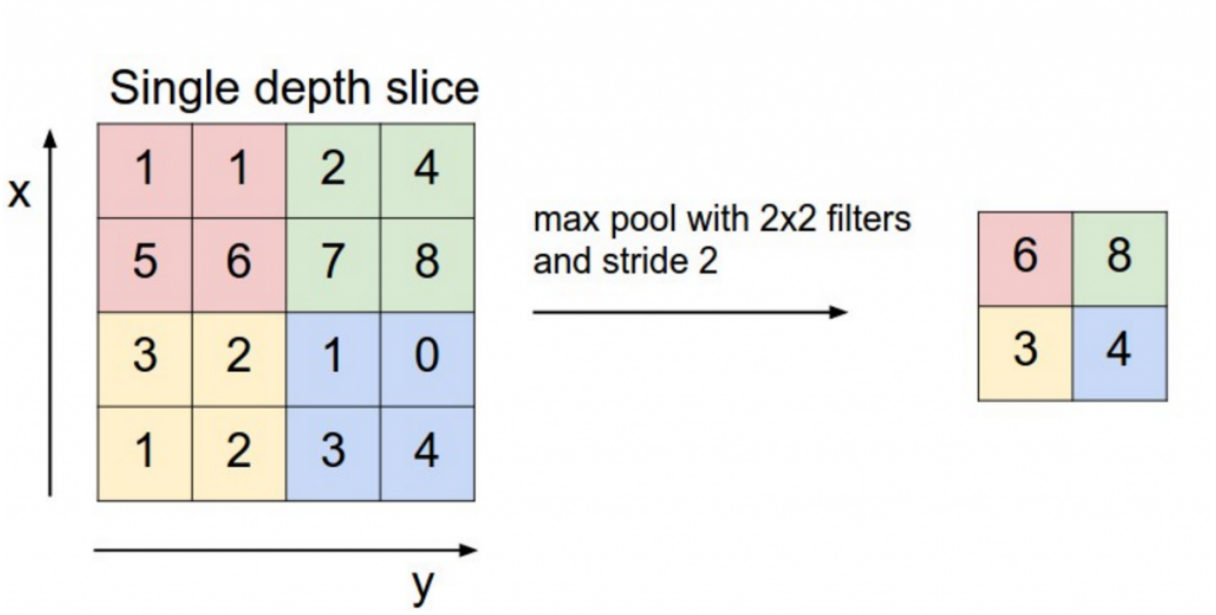


Figure 7 – Max pooling done with a filter of 2x2 and a stride of 2.

Source – Karpathy (2018a).

2.1.2.3 Activation Functions

Every activation function does a fixed non-linearity function over an input number. There are several forms that may be used in neural networks, but the most commonly used in convolution neural networks are $ReLUs$, the Rectified Linear Unity. This function has become very popular over the last few years. This is due to the fact that it has been found that it greatly accelerates the convergence of optimization algorithms such as stochastic gradient descent, compared to other activation functions (KRIZHEVSKY; SUTSKEVER; HINTON, 2012).

This function is a simple check whether the input is over the zero threshold. If true, the output is equal to the input. On the other hand, if false, the output is zero. In other words: $ReLU(x) = max(0, x)$. This operation is inexpensive compared to other activation functions that use exponential operations (KARPATHY, 2018a). Thus, it has a much better performance on larger matrices. See Karpathy (2018b) for a more in-depth notion of other activation functions.

2.1.3 Training neural networks

A large amount of data is usually needed to train a deep neural network or convolutional neural networks. This is essential for the good generalization of the network for new never seen examples of data. Furthermore, the approach used to train a neural network influences directly, as much as the right architecture, in how it will perform in use.

The training process of a neural network is divided into two steps: forward pass and backpropagation. The first is done when the input is intaken through the layers, one by one, and through every neuron necessary until it reaches the output. On this forward passing, the input undergoes many operations and the learned, or initialized, weights are taken into account in this operations (HAYKIN, 1999). On the other hand, the later is done when the forward pass reaches its end and the update of the weights is needed to fit the weights for the correct output given the input. Thus, the backpropagation is a kind of feedback used to update the parameters on the network (HAYKIN, 1999).

2.1.3.1 Backpropagation

Backpropagation is where the weights of the network are learned, where the knowledge is built. In a supervised training, the backpropagation corrects the "guess" that the network made of the output for a given input. This correction is made by calculating the error function between the correct answer and the given answer. Furthermore, the loss, or the error function, is what determines the quality of the output, or classification in the context of this work.

There are a few numbers of loss functions implemented in different types of problems. Mainly, there are two of them, regularization losses, that have the objective of penalizing complexity in a learning model. And data losses, that are used in supervised problems to compute the error between the truth and the prediction. For the latter, there are still some derivatives that are applied depending on the expected outcome as a product. Generally, the final loss is an average over every example of data, this can be represented as $L = \frac{1}{N}\sum_i L_i$, where N is the number of training data and L_i is the individual loss (KARPATHY, 2018a).

The most common derivatives are for classification problems, where we can have a binary, noted by equation 2.4, or multiple classification, represented by the equations 2.5 and 2.6. Where f is the activation function of the network output and y_i are the labels of the data examples.

$$L_i = \sum_j max(0, 1 - y_{ij} f_j) \tag{2.4}$$

$$L_i = \sum_{j \neq y_i} \max(0, f_j - f_{y_i} + 1) \tag{2.5}$$

$$L_i = -\log\left(\frac{e^{f_{y_i}}}{\sum_j e^{f_j}}\right) \qquad \text{or} \qquad L_i = -f_{y_i} + \log\sum_j e^{f_j} \tag{2.6}$$

In the case of a multiple class classification, we have a single correct label among a well-defined set. This then can be approached with the SVM loss function (equation 2.5)

or the cross-entropy classification loss function (equation 2.6). The later is normally used in conjunction with the Softmax function that, contrary to the SVM loss function that uses the outputs as scores for each class, normalizes the outputs, which adds a probabilistic interpretation to it.

Seeing that, the goal of the network is to minimize this loss function, thus the learning is given by its optimization. Presently, the optimization is mostly accomplished by the Stochastic Gradient Descent (SGD), that in turn pursue the optimal loss by altering the weights for the steepest slope in the gradient of the function. This method made possible to train larger models with large datasets and currently it powers nearly all deep learning algorithms (GOODFELLOW et al., 2016). This is due to the fact that the algorithm uses an expectation that approximates the estimated value only using a set of examples.

This set of examples are called batches or mini-batches, that are divided this way due to the computational limits of hardware used to train the networks. The size of this division is used as another hyperparameter in the training of a neural network.

After the appreciation of the data, it is calculated the loss and gradient of the output, and then the local gradients. This is the step where the backpropagation uses this results to update the free parameters in the network. All of this occurs locally in every operation, only using the propagated error in the output to have a global vision of the impact of a small change in the analyzed parameter. This vision can be seen with the chain rule, that take the output gradient and multiply it into every gradient it computes for all inputs.

This is shown in Figure 8, where a function in the form of $f(x, y, z) = (x + y)z$ takes place. It then can be divided in two other functions, $q = x+y$ and $f = qz$. From this we may calculate the gradient, that is simply the partial derivatives of the multiplication function f as $\frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$, and the sum function q as $\frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$. However, the important gradient is in respect of the inputs x, y, z, and not so much from $\frac{\partial f}{\partial q}$.

We can then use the chain rule to multiply the partial derivatives as $\frac{\partial f}{\partial x} = \frac{\partial f}{\partial q}\frac{\partial q}{\partial x}$ and compute the final gradient, that in practice is just a two number multiplication.

In a practical sense, the gradient is telling where the numbers should go to, so the output increases. Therefore, if we decrease the x or y inputs we then decrease the sum and increase the final output. However, we usually wish to minimize the loss function, so when updating our parameters we should walk in the negative direction. The signal tells us where to go, upward or downward, whereas the magnitude tells how far in that direction we should go, the force we should apply. We could think of the backpropagation being the vehicle of the gates communication.

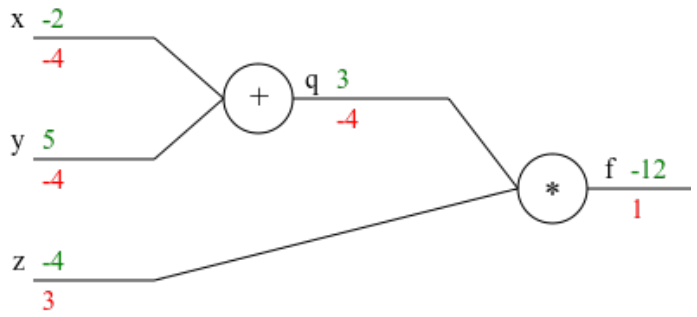


Figure 8 – Example of a forward pass (green), which starts with the inputs, and back-propagation (red), starting from the output backwards applying the chain rule.

Source – Karpathy (2018a).

2.1.3.2 Optimizers

After the calculation of the gradients, it still needed to update the parameters accordingly. Nowadays, there are several approaches reach an update, these approaches are another hyperparameter that we can choose to tune our architecture, these are called optimizers. The most common are:

- **Stochastic Gradient Descent (SGD)**: This method alone has some flavors of its own. Going through the vanilla version until applications with more sophisticated methods such as the Nesterov momentum. The vanilla method is the simplest form of this optimizer, it updates the parameters along the negative gradient direction, where it takes the learning rate (another network hyperparameter) and multiplies by the gradient and then add the result to the current weight. While the use of Nesterov momentum is made with the objective of calculating the step size that should be taken toward the minimum. For this, the product of the velocity and the momentum (that, in this case, acts something more as a friction coefficient) are added to the position to compute the lookahead (BENGIO; BOULANGER-LEWANDOWSKI; PASCANU, 2012). With this in hands, we can then calculate its gradient and then with the learning rate update the parameters.

- **Adagrad**: It is an adaptive learning rate method where the learning rate is normalized by the square root of the sum of the squared gradients (DUCHI; HAZAN; SINGER, 2011). This has an effect of reducing the learning rate of high gradients, while parameters that are less often updated will have an increase in the learning rate.

- **Adam**: This method is an implementation that adjusts the Adagrad method, reducing the aggressiveness of the decreasing learning rate. However, it changes the sum of the squared gradients for a momentum and velocity computation that aims

to reduce the noise in gradients before calculating the new parameter. Also, in its original form, it implements a bias correction for the initialized values of momentum and velocity.

2.1.4 Evaluating neural networks

After the training process of a neural network one needs to evaluate the results and question whether it is satisfactory or not. Furthermore, the threshold and the means to achieve a result, must be clear to every person in the process, since there need to be a consensus on what is a good model.

Moreover, the evaluation may happen not only in the testing, or final, phase of the model. It may happen during the training as a feedback on how well the model is doing, during training time. This is referred to as "Validation" step in deep learning, where a small subset of the training data is separated to be used to evaluate the training and get feedback on notions of overfitting and underfitting. Therefore, we may divide the used metrics in training and testing time.

For this work, we defined the metrics to be consistent throughout this work experiments. This decision was made to build the ground necessary to compare the results between different experiments. Therefore, two metrics were used in training time and three for the testing step.

2.1.4.1 Training metrics

For the training time, the main metric used was the accuracy metric. Nonetheless, as the model classifies 12 classes, the accuracy reported has two variants: top-1 accuracy and top-5 accuracy (or accuracy@5). Both compute the proportion of the true results (both positive and negative) among the total predictions. However, the first accuracy is interested as the top prediction (using softmax as prediction output) of the model, whereas the second calculates the accuracy among the top 5 predictions. So, if the true label is the second higher prediction, the top-1 accuracy will compute this as an error, on the other hand, the top-5 accuracy will compute as a correct prediction.

The formula to compute the accuracy is shown in equation 2.7. Where t_p is the true positive predictions, t_n the true negatives, and s the total of samples predicted.

$$\text{Accuracy} = \frac{\sum t_p + \sum t_n}{s} \tag{2.7}$$

2.1.4.2 Testing metrics

For the testing step, it was created a process that the predictions for the test dataset were generated. With these predictions in hand, as well as the true labels of the

examples, it was possible to create a confusion matrix for the model. Furthermore, with the confusion matrix at hand, was simple to compute other metrics, such as precision, recall (or sensitivity), and accuracy as well.

Both metrics, of recall and precision, can be seen on the equation 2.8. Where f_p is a false positive and f_n is a false negative.

$$\text{Precision} = \frac{t_p}{t_p + f_p} \qquad \text{and} \qquad \text{Recall} = \frac{t_p}{t_p + f_n} \tag{2.8}$$

Another metric used to evaluate the models was the AUC (Area Under the Curve), along with the ROC (Receiver Operating Characteristic) curve. The ROC curve is a mapping of the sensitivity (probability of detection) versus 1−specificity (probability of false alarm), using various thresholds points. Typically, this metric is implemented in systems to analyze how accurately the diagnosis of a patient state is (diseased or healthy) (SWETS, 1986). Furthermore, the AUC summarizes the ROC curve and effectively combines the specificity and the sensitivity that describes the validity of the diagnosis (KUMAR; INDRAYAN, 2011).

Alongside with de ROC curve analysis, is common to calculate the optimal cut-off point. This is used to further separate the test results, so that a diagnosis of diseased or not is provided. When the point is closest to where the sensitivity is equal one and specificity is equal zero, it has achieved the best result possible (HAJIAN-TILAKI, 2013; UNAL, 2017).

This was made to have a solid material on how the model is performing and what it should improve. Since the recall computes the probability of the detection of a lesion, the precision computes the degree to which a lesion will be classified as their true label (how precise is the recall), and the AUC computes how accurately a diagnosis is being delivered.

2.2 Deep Neural Networks Architectures

Deep neural networks are powerful and used in a wide variety of tasks (CIREŞAN et al., 2012; ABDEL-HAMID et al., 2014; MESNIL et al., 2015; WANG et al., 2018a). Many of these have achieved new state-of-the-art marks, because of that DNNs are very popular algorithms. Furthermore, for an application of such an algorithm, it is required the use of an architecture design. That is nothing more than a careful assembly of layers in a stack form.

Therefore, there are two possible options in this decision. One is to create a new design and build a custom architecture from the ground up, and the other is to choose one from the architectures already created and tested by other fellow researchers. This

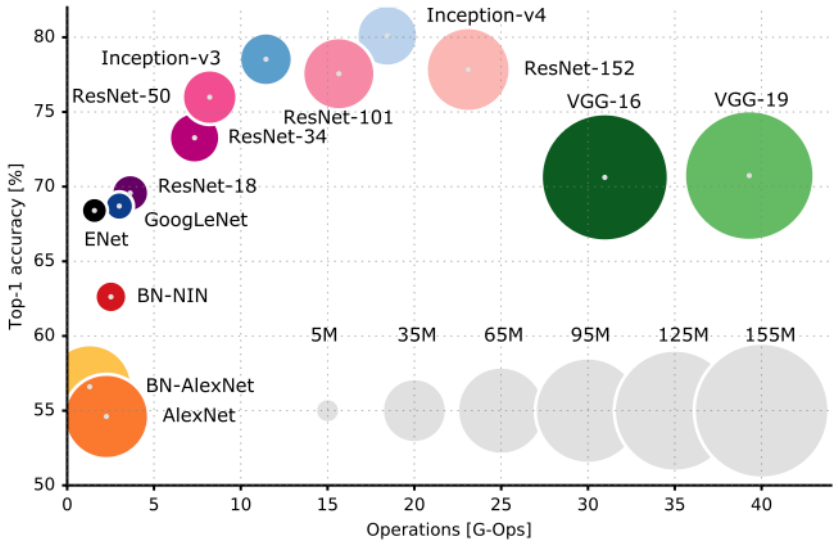


Figure 9 – Comparative of Top1 *vs* Operations between architectures.

Source – Canziani, Paszke and Culurciello (2016).

work chose for the second option, as the creation of a new design is not an easy task to do and demands much more time than it is available to the purposes hereby presented.

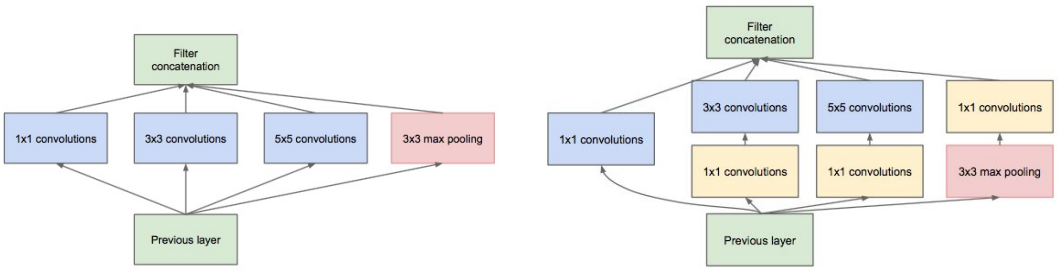
The growing popularity of DNNs can partially explain the appearance of so many new designs of architectures in the computer vision field, each one claiming to be better in some aspect than the other. Thus, this scenario of dozens of architectures (see Figure 9) to choose from can carry a have burden on a researcher. Furthermore, it is needed to research the applications that are more aligned with the one that is envisioned. For this work, we have already done this research (Section 1.2), and thus we have already a few options to choose from.

The most prominent results in the researches gathered are the ones performed by Esteva et al. (2017) and Seog Han et al. (2018), with two architectures that have won the ImageNet competition (RUSSAKOVSKY et al., 2015). These architectures are the *Inception* and *ResNet*.

2.2.1 Inception

The inception architecture was created with the purpose of improving the use of computing resources of deep neural networks. This design was created in 2014 and it won the 2014 ImageNet challenge with a new layer design, the inception layer. The idea of the inception layer is to analyze a bigger area of an image, but also keep information for small spatial fields.

This was achieved with an approach of parallel convolutions, beginning in a small and fine-grained space (1x1 convolution, introduced in Lin, Chen and Yan (2013)), growing



(a) Inception naïve module.

(b) Inception module with dimension reductions.

Figure 10 – Inception module.

Source – Szegedy et al. (2014).

to a medium sized filter (3x3 convolution) and going to a bigger area (5x5 convolution). These sizes have been chosen arbitrarily, only for the ease of calculation (SZEGEDY et al., 2014). Additional to that, there is a pooling layer with a maximization activation, that is responsible to summarize the information from the previous layer, as seen in Figure 10a. After the calculation of these parallel convolutions and pooling, they are all concatenated, forming a single output. Thus, the output means the integration of different scale features, this adds more invariance to object scaling.

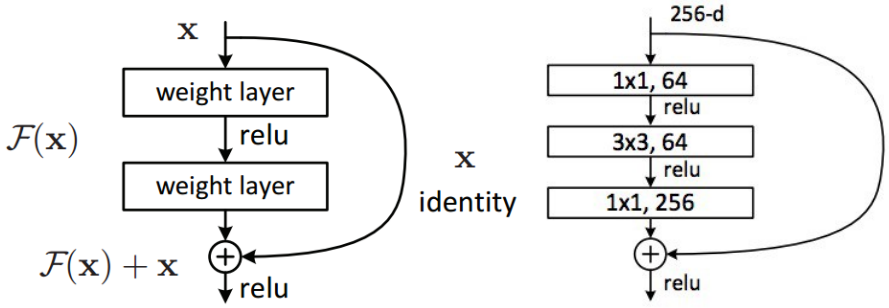
However, this adds a high risk to the architecture that is, after many layers, the inception module output will become huge, and this will be responsible for many of the computational cost done onward. Therefore, this module had to be changed.

A optimization of the inception module came in the form of 1x1 convolutions before the costly 3x3 and 5x5 convolutions, see Figure 10b. These 1x1 convolutions where added to effectively reduce the dimensions, as they use fewer parameters to express more information in the form of a combination of features across filters (LIN; CHEN; YAN, 2013). This reduces the complexity of the final model in terms that the dimensions of filters are also reduced.

Furthermore, since 2014 this architecture has improved significantly in its final versions, although always maintaining the same principle of the modules. We can say that the modules explained above were from version 1.0, and the most state-of-the-art Inception architecture is in version 4.0 (SZEGEDY et al., 2015; SZEGEDY; IOFFE; VANHOUCKE, 2016). For the work done in Esteva et al. (2017), the architecture of Inception-v3 was used for the classification of 2,032 skin lesions between benign or malignant.

2.2.2 ResNet

The ResNet architecture was conceived in 2015, where it has won the ImageNet challenge (RUSSAKOVSKY et al., 2015). Since then, the concepts introduced with this



(a) Residual Block. (b) "Bottleneck" Block.

Figure 11 – Building blocks of a ResNet.

Source – He et al. (2015).

architecture has been used extensively. This is because of the problem that the ResNet proposed to solve, the degradation of the accuracy that is inherent of deep neural networks. Since the addition of many layers leads to higher training error (SRIVASTAVA; GREFF; SCHMIDHUBER, 2015).

Furthermore, the ResNet architecture proposed by He et al. (2015) introduced a new method to train deeper neural networks (more than 1,000 layers), this method was called residual block. The residual block is a direct passage of the input to the next block or layer, thus adding the input with the previous output (see Figure 11a). However, was found that bypassing two layers had a more effective result, as these two layers can be interpreted as a small classifier.

This is done to mitigate the problem of stacked functions (e.g. $y(x) = f(g(h(x)))$), that tries to map directly the input x to the output y. So, instead of just stacking the functions, the residual block adds with the previous learning function in a form of $y(x) = f(x) + x$.

Thus, the residual counterpart of the DNN was found to be easier to train and learn the characteristics shown to it. However, the residual block with 2 weighted layers between the skip, that works well for networks with approximately 30 layers, does not work weel for deeper models (> 100 layers). This is due to the problem of the computational cost introduced by the number of parameters, the same problem experienced by the inception team. Furthermore, bottleneck layers (1x1 convolutions), as seen in Figure 11b, were added to the residual block to reduce the number of parameters in each operation (HE et al., 2015). In addition, the use of these layers improved the combination of features used.

For the initial layers, the architecture uses a simple 7x7 convolution with stride 2, that tries to analyze more sparse features early on, followed by a max pooling function of stride 2, that further reduces the input. Moreover, is applied an average pooling layer, followed by a fully connected layer, and softmax activation to generate the classifications

for the final output.

Another thing that cooperated to the success of the deep models of this architecture was the extensive use of batch normalization layers. The first use of batch normalization was introduced by Ioffe and Szegedy (2015) with the purpose to eliminate the internal covariance shift problem that occurs on the training of DNNs[1]. The use of such layer reduces the dependency on initialization, improves the flow of the gradient for deep models, and allows for higher learning rates usage (IOFFE; SZEGEDY, 2015). This operation can be seen as a preprocessing in the layers level, as it fixes the means and variances of the inputs in its normalization step. Furthermore, insights on the residual network are appearing constantly (VEIT; WILBER; BELONGIE, 2016; LIAO; POGGIO, 2016).

Finally, this architecture has been used in the work done by Seog Han et al. (2018) to train over 855,370 skin lesion images and achieved good results in the skin lesion classification problem.

2.3 Datasets

Based on the data scarcity present in the medical field, the choice of datasets was not made by the selection of the best on a collection of options. The process of choosing one mainly took into account the criterion of public availability. Aside from that, the only pre-requisite was that the dataset was composed with only clinical images (photos taken from cameras without other tools or distorting lenses).

From these criteria, only two datasets fitted the description. The datasets contained 10 (ten) distinct lesions, containing 4 malignant illnesses at maximum. Another additional dataset was gathered from dermatologic websites, using a script for scrapping pages. The latter dataset was acquired from Seog Han et al. (2018) and is not publicly available due to copyrights owned by the websites. Finally, these datasets are further discussed below.

MED-NODE

The first dataset used is provided by the Department of Dermatology at the University Medical Center Groningen (UMCG) (GIOTIS et al., 2015). This dataset contains 170 images that are divided between 70 melanoma and 100 nevus cases. Furthermore, these images were processed with an algorithm for hair removal.

[1] This claim, however, is still discussed. See Santurkar et al. (2018).

Edinburgh

The second dataset is provided by the Edinburgh Dermofit Image Library and is publicly available for purchase, under an agreement of a use license[2]. This dataset is the more complete one found on the web. It contains 1,300 images, that are divided into 10 lesions, including melanoma, BCC, and SCC. These images are all diagnosed based on experts opinions. In addition, it is also provided the binary segmentation of the lesion, for each one. It is valid to note that the images are not all in the same size.

Furthermore, the lesions and its respective numbers are listed in the table 1.

Table 1 – Lesion sampling for Edinburgh dataset.

Lesion Type	***Number of images***
Actinic Keratosis	*45*
Basal Cell Carcinoma	*239*
Melanocytic Nevus (mole)	*331*
Seborrhoeic Keratosis	*257*
Squamous Cell Carcinoma	*88*
Intraepithelial Carcinoma	*78*
Pyogenic Granuloma	*24*
Haemangioma	*97*
Dermatofibroma	*65*
Malignant Melanoma	*76*
TOTAL	***1,300***

Atlas

This last dataset, was acquired from running several scripts for scrapping different dermatological websites[3]. So that is the reason that this dataset was baptized as Atlas. This dataset was obtained from Seog Han et al. (2018) in a personal submitted request. It contains 3,816 images downloaded from websites and distributed between six lesions.

The difference from the Edinburgh dataset is that this contains two lesions that are not present on the first, as it can be seen on table 2. That lesions are Wart and Lentigo, both benign lesions. This, alongside with the Atlas and MED-NODE datasets, sums up to 12 lesions, that are the interest of this work.

[2] Available at <https://licensing.eri.ed.ac.uk/i/software/dermofit-image-library.html>.

[3] These websites included, <http://dermquest.com>, <http://www.dermatlas.net>, <http://www.dermis.net/dermisroot/en/home/index.htm>, <http://www.meddean.luc.edu/lumen/MedEd/medicine/dermatology/melton/atlas.htm>, <http://www.dermatoweb.net>, <http://www.danderm-pdv.is.kkh.dk/atlas/index.html>, <http://www.atlasdermatologico.com.br>, <http://www.hellenicdermatlas.com/en>.

Table 2 – Lesion sampling for Atlas dataset.

Lesion Type	***Number of images***
Basal Cell Carcinoma	*1,561*
Lentigo	*69*
Malignant Melanoma	*228*
Melanocytic nevus (mole)	*626*
Seborrheic keratosis	*897*
Wart	*435*
TOTAL	***3,816***

Another difference is the quality of the images, since the dataset was collected from web pages, is not all the images that present the same quality, nor the same common viewpoints observed on the Edinburgh dataset. Therefore, this dataset is the most heterogeneous in matters of quality of imaging, viewpoints, the age of patients and ethnicity. However, this dataset in its entirety is not officially diagnosed by specialists, but on the other hand, these photos were displayed on websites that are reliable and used by students. So, there is a heuristic that these images were revised before putting to display in these websites and can be trusted.

Dermoscopic datasets

Another source of data that is publicly available is the International Skin Imaging Collaboration (ISIC) project archive. The project is an open source archive with public access of clinical and dermoscopic images of skin lesions. This project was created as an industry and academia partnership, with a goal to reduce melanoma deaths and unnecessary biopsy exams and procedures (International Society for Digital Imaging of the Skin, 2018).

This archive contains a well-organized set of data, containing metadata for every picture available. This metadata contains the diagnosis of the lesion, the age approximation, and sex of the patient, site of lesion and images metadata. This information can be used to further expand the classification in a more fine-grained approach.

Although this project was created with a goal to provide dermoscopic and clinical images, no dataset containing clinical images has been uploaded yet[4]. Therefore, this dataset is listed as a potential extra source of images for further testing and experimenting with a mixture of dermoscopic and clinical images.

[4] Last accessed in 20/03/2018.

2.3.1 Lesions of interest

Seeing the lesions seen in tables 1 and 2, we can divide skin lesions into two major groups, one being malignant lesions and the other benign lesions. The first is composed mostly of skin cancers and the latter being composed with any lesion that does not pose a major threat. One counterexample of this division is the actinic keratosis, that presents itself as a potential SCC, as it has the potential to develop into it. Thus, actinic keratosis is classified as a precancerous lesion (PRAJAPATI; BARANKIN, 2008). Seeing this we can create a visualization of the lesions in its respective groups – the Figure 12 shows this visualization. For this work, 12 lesions were chosen and analyzed, 4 malignant and 8 benign (being 1 precancerous).

They are well distributed over malignant and benign lesions. Therefore it is useful to map it to a real-world problem. The lesions that will be studied and are of interest for this work are the ones listed and further discussed below.

2.3.1.1 Actinic Keratosis

Actinic keratosis (Figure 14e), or solar keratosis is most frequent on those who are exposed constantly to the sun (MOY, 2000), thus is more frequent on those who are fair-skinned. For these factors, this illness is also more frequent on places more closer to the equator, such as Brazil and Australia.

This lesion presents itself as a patch of hard, scaly and often notable pigmentary alterations on the nearby skin, denoting yellowish discoloration, showing damages of sun exposure. Often detected on the head, neck, back of hands and forearms (MOY, 2000).

Although actinic keratosis is mostly benign, if left untreated it has a potential 20% of risk to progress into squamous cell carcinoma, a malignant disease, so treatment is highly recommended by dermatologists (PATTERSON, 2014).

2.3.1.2 Basal cell carcinoma

This lesion is a type of non-melanoma cancer, that it is originated from the sudden and uncontrolled growth of basal cells. These cells are located in the lower part of the epidermis, on the basal cell layer (American Cancer Society, 2016).

Furthermore, this is the most common type of skin cancer, where 80% of skin cancers are diagnosed as basal cell carcinomas. Each year is estimated that 4.32 million new cases of this disease are diagnosed, the numbers are not accurate since it is not required for the professionals to report it to the cancer registries, contrary to melanoma cancers (American Cancer Society, 2017).

In addition, if not removed and treated completely it can recur in the future on

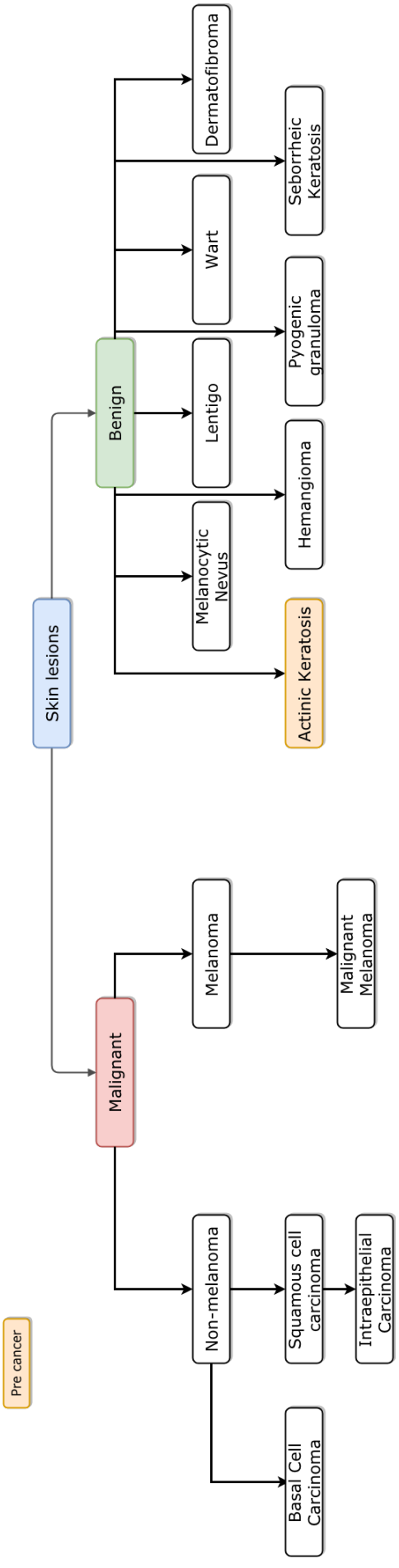


Figure 12 – Skin lesions groups.

Source – Author.

the same place (American Cancer Society, 2016). The rate of death for the non-melanoma cancers is low, however with its high incidence it can have as many casualties as melanoma cancers, in Brazil on 2015 were 1,958 deaths for non-melanoma cancers and 1,794 deaths for melanoma (BRASIL; Ministério da Saúde, 2017). Most people that died from this cancer is due to the lack of a doctor opinion, since the patient had not consulted a specialist until the lesion had grown and evolved (American Cancer Society, 2016).

2.3.1.3 Dermatofibroma

It can appear in many different colors, but usually are brownish or tanned, and are often elevated spots in the skin. It usually presents itself in the extremities of the body, arms, legs, and forearms. Normally it is asymptomatic, however, it can present tenderness and a chronic itchy skin. It is described to be the most common painful skin lesion (NAVERSEN et al., 1993).

The cause of this condition is unknown, however historically it has been attributed to be related to traumas on the skin (e.g., insect bites and tattooing) (EVANS et al., 1989; LOBATO-BEREZO et al., 2014). One other factor that is likely to be related is the alteration of the immunity system.

2.3.1.4 Hemangioma

Hemangioma is a benign lesion and it is the most common cutaneous vascular proliferation. The tumor is formed by an excess of blood vessels or a proliferation of dilated venules. As seen in Figure 14g, it has a red coloring, due to the high amount of blood near the skin, varying from a small red macule to a larger dome-topped lesion (ODOM; JAMES; BERGER, 2000).

2.3.1.5 Intraepithelial carcinoma (Bowen's disease)

Was first described by John T. Bowen in 1912 (BOWEN, 1983), and is actually a squamous cell carcinoma *in situ* with the potential for significant lateral spread. This means that it is pre-invasive and if left untreated, there is a chance that it can evolve into the deeper layers of the skin. This disease is most reported on fair-skinned people that are exposed to the sun, it rarely occurs on darker-pigmented skin patients (GUPTA et al., 2009).

This lesion is often located in the outermost layers of the skin, resulting on a reddish patch, and sometimes raised spots on the skin, as can be seen on Figure 14b.

This condition is more common to affect older patients, over the age of 60 years old, and the prognosis of this disease is favorable, but it has a risk of progression to invasive squamous cell carcinoma (MORTON; BIRNIE; EEDY, 2014). The risk of progression

rises with the delay in seeking assistance, and this is mostly because the lesion is asymptomatic, not causing discomfort for the patient. The early skin changes may be subtle and sometimes overlap with characteristics of other lesions (e.g., seborrheic keratosis) (BÖER-AUER; JONES; LYASNICHAYA, 2012).

2.3.1.6 Lentigo

It is a benign lesion, a small, sharply delineated, pigmented macule ranging from brown to black. Lentigines (plural for lentigo), proliferate linearly on the basal layer of the skin with an increased pigmentation that can be homogeneous or not. In contrast to moles (melanocytic nevi), that aggregates melanocytes in a single spot. The distinction of a lentigo to other melanocytic lesions (e.g., melanocytic naevus, melanoma) is of most importance since it is a marker for ultraviolet damage and systemic syndromes.

The study of patients that have had multiple lentigines serves to identify a population with a higher risk of developing malignant melanoma (DERANCOURT et al., 2007). That is a subject of study since it is also associated with long-term exposure to the sun (GOOROCHURN et al., 2017), and for that, it is more common to occur in adults than in children. Supporting this, was found that lentigo and seborrheic keratosis increased on the drivers' side face on a preliminary study of truck drivers in Turkey (KAVAK et al., 2008).

Another source of lentigines is aging, and for this matter, they are often referred to as senile lentigo or marks and are more pronounced in Japanese than in German women (TPCN, 2011). Therefore, lentigines are benign by nature, and treatment is often recommended for cosmetic reasons.

2.3.1.7 Malignant melanoma

Malignant melanoma is the abnormal growth of melanocytes cells or the cells that develop from it (National Cancer Institute, 2018c). These cells are the ones that make the pigment melanin and give the coloring to the skin. They are located on the basement membrane, on the division of the epidermis with the dermis as it can be seen in Figure 13. This poses a major risk, as the neoplasm is present deeper in the skin and can gain the ability to metastasize.

Although it was once considered uncommon, the incidence rates and deaths have increased significantly over the past 30 years (American Cancer Society, 2018). In 2018 is expected to be 9,320 deaths from melanoma and 91,270 new cases in the United States (American Cancer Society, 2018). In Brazil is estimated to be 6,360 new cases for 2018 (Instituto Nacional de Câncer José Alencar Gomes da Silva, 2018).

The causes of this malignancy can be intrinsic or extrinsic, both can contribute

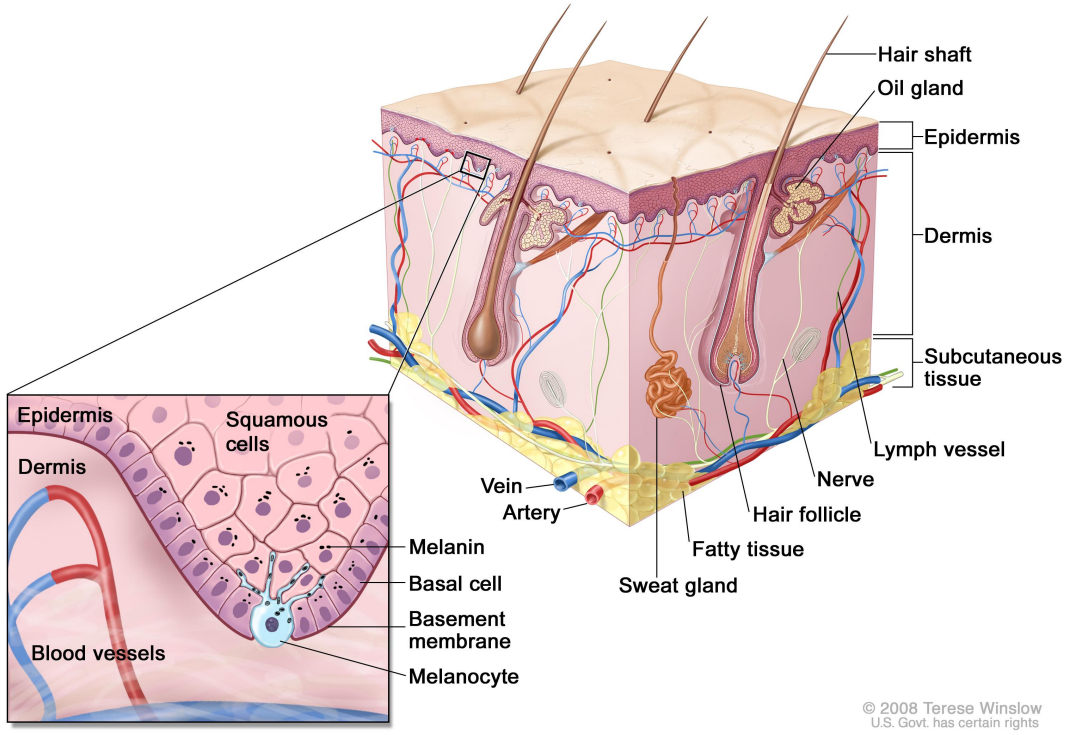


Figure 13 – Representation of normal skin.

Source – National Cancer Institute (2018c).

to its emergence (e.g., sun exposure, multiple nevi, family history). The lesion can occur anywhere, including mucosal surfaces, and they present darker or variable discoloration, growth, and evolution in the edges and bleeding. The latter being on later stages where the survival rates have dimmed. If not treated soon, the 5-year survival can decrease from 97% to 15% (National Cancer Institute, 2018a). Hence the importance to use the techniques and tools (e.g., ABCDE) to early detect skin lesions.

2.3.1.8 Melanocytic nevus

Melanocytic nevus is a benign lesion, composed of melanocytes, the cells that produce pigment. It occurs in all mammalian species with a special incidence in humans, dogs, and horses. When grown in clusters, the melanocytic nevi are called the common mole, and its usual to an adult to have between 10 and 40 in its body. They form in early childhood and are common to appear until later in that life phase. After that they continue to develop until the age of 40 and then tend to fade (National Cancer Institute, 2018b).

2.3.1.9 Pyogenic granuloma

It is a relatively common benign vascular lesion of the skin and mucosa, but its root cause is unknown (MILLS; COOPER; FECHNER, 1980). The lesion is rather poorly

named since the lesion is neither pyogenic (forms pus) nor a granuloma (inflammatory lesion). The main issue with the pathology is the proneness to bleeding and ulceration.

As seen in Figure 14j, the lesion presents itself as a glistening red lesion and are most commonly located on the head and neck. It is also common in the fast evolution over the course of a few weeks and the spontaneously bleeding, however, if left untreated they usually atrophy and slowly regress.

2.3.1.10 Seborrheic keratosis

This lesion is the most common benign tumor in elderly people, as they appear with the increase of age. They are mostly benign, but secondary neoplasms may arise from within the lesion, such as SCC, Bowen's disease, and melanoma. In addition, the lesion is mostly predominantly asymptomatic but can itch and develop into inflamed lesions if scratched violently (HAFNER; VOGT, 2008).

Although it is benign and treatment is not mandatory, it is recommended to consult a specialist for further examination. If the lesion is darker than usual it may be mistaken with melanoma, also it may develop a case of basal cell carcinoma, although it is very rare (MIKHAIL; MEHREGAN, 1982).

2.3.1.11 Squamous cell carcinoma

After basal cell carcinoma, it is the most common type of skin cancer in the US and together accounts for more than 5.4 million cases diagnosed (American Cancer Society, 2018). This cancer is developed from a neoplasm in the squamous cells, that are closest to the surface of the epidermis, as seen in Figure 13. It often appears in sun-exposed areas of the body, primarily in the head and backs of hands (American Cancer Society, 2016). It often presents itself as a shallow ulcer, often covered in a scaly plaque that may range from a flesh tone to a red.

In addition to general appearance, is recommended to look for signs of this lesion if the patient has a history of actinic keratosis. The patients that have had multiple actinic keratosis are more prone to develop this skin cancer (HOWELL; RAMSEY, 2017; MOY, 2000).

2.3.1.12 Wart

Warts are mostly benign and are caused by certain types of the human papillomavirus (HPV) (ORTH; FAVRE; CROISSANT, 1977). This virus is spread by direct or indirect contact and can resist a long period of incubation, ranging from 1 month to as long as 6 months.

This lesion is characterized by its rough and irregular surface that is similar in color to the skin, further features are determined by the type of virus that is causing the infection. If left untreated is common to resolve spontaneously in some months or years, although is common to treat it for cosmetic issues (LOO; TANG, 2014).

2.3.2 Data Difficulties

The difficulties faced with this dataset can be traced back to the problems surrounding medical datasets. Thus these difficulties are similar to other medical imaging datasets. Furthermore, the main adversities can be found into one of the points listed below.

Data nature

Since the dataset consists solely of clinical images of skin lesions, this brings a very unique set of challenges that are inherent from this domain and type of data. Thus, it is necessary to understand these challenges, so we can solve and evaluate them in the best way possible. Furthermore, the different datasets that are used in this work add more variability, but also they add more challenges, such is the way of the real world of clinical imaging.

Furthermore, we found that the dataset of Edinburgh is the most organized one, in a question of reproducibility of the work done and quality of an image. Since the images were put through standardized conditions that controlled the final output. The MED-NODE dataset is controlled in the manner that it does not contain images that are not far away from the lesion. However, it does contain a high variability in the images clearness and focus, thus leading to some images sharp clear and others blurry. In addition, the Atlas dataset is the one with most internal variance, since its composition being from several different sources. Therefore, the main challenges found in this work were:

- **Ethnics**: As exposed by Seog Han et al. (2018), the diversity in ethnics on the domain of skin lesion is very important since the symptoms and appearances of the same lesion can vary on different ethnics and different ethnics has different skin tones. Thus, for a good generalization of the problem, it is necessary a significant sampling of different ethnics of the same lesion. It is not possible to know every patient ethnicity from these datasets since this information is not disclosed. Maybe due to the patient's privacy. However, an analysis is known that the datasets used in this work are composed mainly of Caucasians. This lack of diversity can lead to a poor test accuracy on new samples that are from other ethnicities.
- **Age range**: Another fact of consideration is the wide range of age that a lesion can be present on a subject. This also interferes with the appearance of the lesion and

(a) Basal Cell Carcinoma (b) Intrapithelial Carcinoma (c) Malignant Melanoma

(d) Squamous Cell Carcinoma (e) Actinic Keratosis (f) Dermatofibroma

(g) Hemangioma (h) Lentigo (i) Melanocytic nevus

(j) Pyogenic Granuloma (k) Seborrheic Keratosis (l) Wart

Figure 14 – Lesions of interest for this work.

Source – Edinburgh dataset.

is a factor that must be taken into account if a classification model is to be built for this domain. It was found in Seog Han et al. (2018) that the most of the images in a private dataset where from elderly people, so the model did not generalize well enough for young individuals that presented the same illnesses.

- **Different set of cameras**: This is more a variance and diversity problem. This is a challenge in a way that with more cameras, more different specifications and capturing setups are used to create an image. This means that the same photo taken with two different cameras may have different viewpoints, different colors capturing system and different quality of an image.

- **Different abilities for photography**: Along with the different set of photographic cameras, it still needed to have abilities that corroborate to the success of the final clinical images. This leads to images with blurry effects, a different framing of the lesion (some with more healthy skin, some with less), and different body posing.

- **Posing**: This problem can be caused partially by the lack of abilities to take a clinical image. Other times, the local of the lesion obligates the picture to frame some other body parts (e.g. such as parts near the nails). Furthermore, is observable that some photos include body parts that were not meant to be in the picture (e.g. lesion in the cheek and the eye corners or nose are framed). This can pose a difficulty for the learning model to generalize.

- **Body hair**: Another common case of a challenge posed by this dataset, is the partial occlusion of some lesion images by the body hair of the person. It is common to see this phenomenon occurring mostly on the scalp, where the hair occluded the lesion and poses an obstacle to properly analyze it.

- **Skin elasticity and reflexivity**: Another fact that permeates this domain it the inherent characteristics of a skin lesion, and photographies of patches of skin. The challenges brought by these natures are the elasticity of the skin and its reflexivity. The first can be the means to distort shapes that are on the top of the skin. It can be also accounted for this fact that as humans grow old the elasticity of their skin changes (CUA; WILHELM; MAIBACH, 1990). The latter is due to the property of the skin to reflect direct focal lights. Depending on the body part the skin may have more reflexive glare or backscattered light proprieties (DENGEL et al., 2015).

As the nature of this data is very unique, the use of methods such as transfer learning became a challenge. Since the common pre-trained networks are created using the ImageNet dataset, that is not close to the proprieties of this one. Therefore, to use such a technique on ImageNet network weights is challenging.

Minutiae

Another factor that must be taken into account is the minutiae between lesions. Since a lesion must appear in the patient skin can sometimes be just a tone darker than the skin color, leading to the mislead that there does not characterize well the skin lesion. Another point of confusion is the inter minutiae between lesions. This is easy observable in the melanocytic lesions since they share many features.

Data Sampling

As exposed in the section 2.3, these datasets do not have the optimal sampling number, having various fluctuations in the number of examples for the lesions. As an example, the *Pyogenic Granuloma* only has 24 examples in all the datasets. Furthermore, this low number on the examples of one class can mean an underfitting of this learning model. Moreover, it is necessary to rebalance some of these lesions with the purpose to get a more fair representation of the world as well as gathering enough data so that the models learn. For this, the method of data augmentation will be used for the up-sampling of data.

Labeling

Another issue regarding this application and dataset, is the labeling process that the data underwent until it reached its final stage. Since this domain has delicate implications on what may arise from a erroneous labeling, this topic is one of important attention. The dataset of Edinburgh has proved the labels through an opinion of a specialist, the MED-NODE does not state which method is used to label their data, and the Atlas data differ from website to website. However, none of these states to label their data from a response of biopsy, which would be a important and valuable information to have, since biopsy is the last resort to decide which lesion the tissue is from.

Another thing that is not described in the dataset providers is whether there is an aggregation of lesions under the primitive lesion class or not (e.g. blue nevus is a melanocytic nevus), this could be adding intra-class confusion.

2.4 Handling data scarcity

As noted previously, for the correct generalization of the weights and biases of a network, a huge amount of data is needed. However, the medical field lacks this amount of images and if only used the data public provided, a good generalization of the problem cannot be met if we wish to train a network from zero. Therefore, the need for new approaches arise. How do we approach a problem that lacks the data needed for the proper training on a blank network? One option is to gather new data. But if it is not

possible to do it we "forge" new data. But this processes needs to be done in such a way that the final product is not altered to the point that the original label does not fit it anymore. We can call these as non-invasive transformations. That although we alter the original data, the label can still be applied to it.

Another option to this is to use previous knowledge of other similar problems and build the new concepts needed on top of it. However, to do this another problem with huge amounts of data is needed. Furthermore, this new problem has to hold some kind of relation to the problem that needs to be solved. This is necessary for the transferability of knowledge itself. For an instance, if we train a network to detect objects, a simple subset of random objects, and then used this built knowledge in a problem to detect faces, this will surely help. It is intuitive to think, if someone knows how to detect an object - a thing - then it surely can be taught to detect a pencil in a table, as the perception to spot edges, curves and differences in colors is already a learned ability.

This two options are called **Data Augmentation** and **Transfer Learning**, respectively.

2.4.1 Transfer Learning

In practice, the domains that are faced in the industry, rather than the academia, usually have low numbers of labeled data. This poses a major obstacle to train a deep convolutional neural network from scratch, since the data may not demonstrate a true representation of the real world. Thus, it is common to see works that utilize the pre-trained weights of a previously trained architecture, this can lead to 2 major approaches.

The approaches are: using a CNN as a fixed feature extractor and fine-tuning the architecture. The first is mostly used to collect features of images and then use to train a linear classifier in a new dataset. The second strategy is to continue the training of the network, replacing completely the final layer, but updating the parameters through backpropagation.

A common use of pre-trained models for object classification is from models that are trained on the ImageNet dataset. Some recent work done by Kornblith, Shlens and Le (2018) shows that ResNets take the lead in performance when treated as feature extractors, while only fine-tuning some models to other datasets, they achieved a new state-of-the-art. All these tests used the pre-trained weights and fine-tuned them with Nesterov momentum for 19,531 steps, which sometimes corresponded as more than 1,000 epochs using a batch size of 256. Finally, it was proven, empirically, that the Inception-v4 architecture achieves overall better results for this task than the other 12 pre-trained classification models.

Therefore, transfer learning optimizes and cuts short most of the time in the train-

ing of new applications. However, this can add some constraints to the work. One example of this is when using a pre-trained network is not possible to extract and change arbitrarily the layers of the network. Another point is that normally, small learning rates are applied to CNN weights that are being fine-tuned. This is because we already expect that the weights are good, and we do not want to distort them too much (YOSINSKI et al., 2014).

2.4.2 Data augmentation

Data augmentation is a technique used for application where we do not have an infinite amount of data to train our models. This can be done by introducing random transformations to the data. In image classification, this can be translated as rotating, flipping and cropping the image. These perturbations add more variability to the input, thus this could mean an overfitting reduction in our model by teaching it about invariances in the data domain (KRIZHEVSKY; SUTSKEVER; HINTON, 2012; PEREZ; WANG, 2017; CUBUK et al., 2018). Therefore, these transformations do not change the meaning of the input, thus, the label originally attributed to it still holds its importance.

Although some transformations in an image can be done agnostic to the field of application (e.g. translation), some other transformations are entitled to domain-specific characteristics. For this work we used an additional transformation that randomizes the natural light effect in the picture, this was done to mimic the transformations seen in indoors clinics due to different light sources. “Random distortion” is another applied transformation that was chosen based on the domain, in order to mimic the flexibility and the malleability of the skin. Furthermore, the probability of application and magnitude variability are added to the transformations, having in mind the variability increase added in the data.

2.4.2.1 Augmentation Methods

Have seen the needs and importance of augmentation in the medical field. It was searched a framework or tool that could aid in this task. Finally, we found the Augmentor Python library (BLOICE; STOCKER; HOLZINGER, 2017) for implementing the augmentation process. The library has predefined transformations (e.g. rotation, flipping, translation, ...), and has a hot-spot for new transformation implementations. Which was quite useful when implementing the method to add light variance to the augmentations. Aside from this, several other transformations were applied.

Transformations

Each decision to choose the transformations to be applied had been based on general guidelines of data augmentation (PEREZ; WANG, 2017; CUBUK et al., 2018) or

on the nature of the data. Furthermore, the transformations were aligned in a pipeline fashion, where each had a probability that defined the likelihood of being applied to the image. Finally, the new image was saved at the destination. Moreover, the operations, and probabilities, used for this work were the ones listed in table 3.

Table 3 – Transformations applied for data augmentation.

Transformation	**Probability**
Rotation	0.5
Random zoom	0.4
Flip horizontally	0.7
Flip vertically	0.5
Random distortion	0.8
Lightning variance	0.5

Seeing this, the algorithm that implements this pipeline can be seen in Appendix A.

2.5 Datasets Preparation

The dataset preparation is an important step when training machine learning models. This procedure takes into consideration the sampling of the dataset so that the training data is separated from the validation and test. This task has major importance, because, if the training data is also used as testing data the results will be tampered with and no conclusions can be taken from it. Therefore, a well-supervised procedure of separating these datasets must happen.

For this work, we implemented a concise methodology. First of all, a test set is separated, usually, a 10% of each lesion, before of any transformation is applied on the dataset. Following this, if the experiment requires, is done the data transformation process. Then, for each experiment, the sample is analyzed to see how much is necessary to upsample or downsample each class.

After processing the necessary images to compose the training and test datasets, the images for the training dataset are processed to create an LMDB file (CHU, 2011) for fast access to the data in training time. In this process the training dataset is divided between a training set and a validation set, that is used to verify the results of the training in train time. Thus, this split is done in a way that 80% of the data is used for training and 20% is for validation. However, this split is done in a stratified way, so that each split has a fair amount of each class.

Finally, these slices of the original dataset are kept separated and are used as such for the experiment. The code that implements the creation of the LMDB file and the split of the training/validation set can be seen in Appendix B.

3 Explainability & Interpretability

AI is becoming more and more a part of our everyday life, and the more we involve AI in our daily lives, the more we need to be able to trust decisions autonomous systems make. Right now too much of what AI systems do are "black-boxes". We have little visibility into how decisions are being made, how conclusions are drawn, objects identified, and more. Moreover, the need for an explanation rises, even more, when AI systems are integrated into sensitive environments that affect people personally, such as financial, education, job hiring, and medical field (CARUANA et al., 2015). The ability to explain a certain decision has also been rated as the most desirable feature of a decision-assisting software (TEACH; SHORTLIFFE, 1981).

As the consequences and mistakes of these systems become more significant, it becomes more important to have visibility and transparency on the inner workings of models decision making. This transparency then can improve the means of accountability and detection of anomalies, which can be used by engineers to improve the technology seeing what went wrong. This becomes a serious question when the predictions start to fail. When this happens there is no right answer to what is happening. This could be because of some biases in the training or testing data, or to peculiarities inherent of the algorithm. However, without digging deeper and uncovering the real reasons there is no way to know. For that, the addition of explanations to learning models is needed.

As these applications that leverage on machine learning models are going into a production environment, the complexity to explain decisions is becoming more intricate. There are many supervised and unsupervised ways to train a neural network, and these systems are guided with approaches that the human or the system judges as correct or incorrect. However, once the system is "in the wild", many of the times, there is no supervisor looking at it anymore. Therefore, we do not know how the AI system is truly operating. However, the growing problem is that with the rise of deep neural networks the complexity to reach this explainability of decision making is becoming harder.

Another thing to take into account is that there are many levels of need for an explanation. There is much less impact in distinguishing a food as a hotdog or "not hotdog"[1], than an autonomous car taking a right turn where it was not supposed to, or telling a person that there is an urgent need for a surgery. Therefore, it also needed to introduce explanations in systems while they are making decisions as well as after the fact, so that is possible to audit the decision making process to see what went wrong or right.

[1] <https://itunes.apple.com/us/app/not-hotdog/id1212457521?mt=812>. Accessed in 05/11/2018

This is a rising topic on artificial intelligence, that aims to address the process of decision making of black-boxes in AI systems. However, it has been around for a long time. Since 80‘s decade, researchers were concerned with decision-making software (mainly expert systems at the time) (CLANCEY, 1983; CLANCEY; SHORTLIFFE, 1984; CHANDRASEKARAN; TANNER; JOSEPHSON, 1989). They addressed this topic with the same discourse that is being raised nowadays, that a system that deals with sensitive decisions must be able to answer “Why” questions. However, there are still discussions on how to answer such questions, how to answer “What is a good explanation to this question?”.

3.1 Concepts

If used in the colloquial sense, any information that clarifies a decision can be used as an explanation, thus, with an explanation, we can call a system interpretable. However, this explanation can be made in the same sense in which gravitation and baking a cake are explained, with a defined set of rules that the system follows without a reference to any specific example. When an explanation is asked for we generally ask for the reasons or the justification for a particular decision, rather than the description or the followed rules in the decision-making process.

Explainability and interpretability are terms that are used with different claims to what they mean (LIPTON, 2016). Furthermore, in this work, we will use the definition brought by Miller (2017), which does not define a distinction between explainability and interpretability, although, there is a distinction for explanations.

Explanations can be defined as an answer to a *why* question, to address a single inquiry (DENNETT, 1989; OVERTON, 2011). Moreover, they are processes that happen after the fact of interpretability (MILLER, 2017). In turn, interpretability is the degree to which a human can understand the cause of a decision (in this case, a prediction) (MILLER, 2017). So, the higher the interpretability of a model, the easier it is for a person to comprehend the why‘s of a certain decision.

Furthermore, Biran and Cotton (2017) state that a system is interpretable if a human can understand its operations, either through introspection or a produced explanation. Therefore, to make a machine learning interpretable we can, but are not obligated to, provide a human-style explanation of a decision. Moreover, they also define what is a justification, which is what explains why a decision may be good or bad but does not inform about the decision-making process (BIRAN; COTTON, 2017).

Therefore, explainable AI, or XAI, refers to the agent that is responsible to explain its decision-making process or another agent decision. Furthermore, although we can use more models to understand and explain what an already deployed model in production,

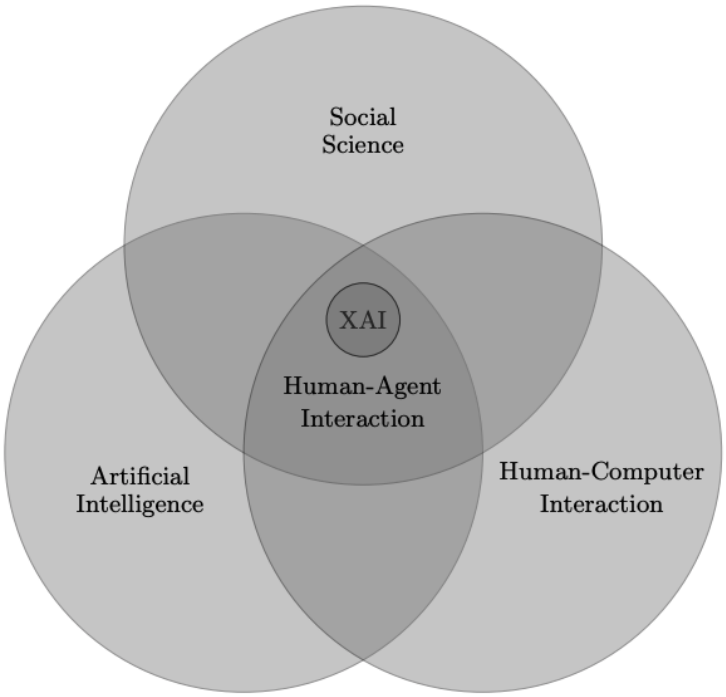


Figure 15 – Scope of explainable artificial intelligence.

Source – Miller (2017)

usually the solution it is not more models. XAI involves many fields of study such as philosophy, psychology, cognitive science, and human-computer interaction. Each one of these fields contributes to the meaning and methods of how an explanation should be approached. Miller (2017) also provides a new scope to XAI as a human-agent interaction problem, this can be viewed in Figure 15.

3.2 Importance

As mentioned above, AI systems are becoming ubiquitous, entering every branch of society and making important decisions. Furthermore, these systems are created, deployed and used by people, people that interact daily with explanations in their relationships. Whom may not be satisfied with only the information predicted (the *what*), but may also need to know the process in which the information was created (the *why* and *how*). This necessity for explainability comes from a fundamental flaw in the problem formalization (DOSHI-VELEZ; KIM, 2017) since a correct answer only solves part of the original problem.

It has passed the time when all we needed to trust a system was its high accuracy. As Doshi-Velez and Kim (2017) state in their work, just a classification accuracy is an incomplete description of most real-world tasks, since a single metric cannot give all information needed to assist in these tasks. Furthermore, over time the performance of a model may change due to many factors, one of them being concept drift (TSYMBAL,

2004). Therefore, the importance of XAI comes to place when we talk about technical advances, product adoption, and law compliance. The latter being a topic of discussion on governments around the world.

3.2.1 Technical advances

Engineers and data scientists often do not know what is happening under the hood of deep learning models. That usually raises concerns to whether the model is learning what it was supposed to do or it is learning any biases present in the data. The machine learning model might even be breaking some laws unintentionally, when it uses race as a primary feature to decide whether a person should be released from prison or not (ANGWIN et al., 2016). Interpretability is a useful tool to detect bias in machine learning models, as it serves as a tool to understand what is causing the decisions.

Another example of bias is when a model perceives adjacent features as main features (not what prior belief would dictate) for a given class. A case of this phenomenon is the "Husky vs Wolf" test, where miss-classified Huskies where due to the background snow present in the picture (RIBEIRO; SINGH; GUESTRIN, 2016b). Therefore, an explanation revealing the snow as being a predominant feature to classify Wolves may drive professionals to adjust and review the system.

Furthermore, this same fact may help specialists understand the domain that they are studying and gain insights. An example of this can be the use of an autonomous driving car, that detects a cyclist in the road correctly. Looking into an explanation, the scientists might see that the two wheels are the major features that identify the cyclist. However, seeing a bicycle that uses side bags that covers the rear wheel may cause the detection to malfunction and cause an accident. This can be prevented if these flaws are spotted early on development.

Moreover, explanations may function as to assist researchers in their endeavors. Some explanations may lead to insights on how a regular model or algorithm interprets the data and gives it a meaning. And usually, when the findings are not prosperous with deep learning, researchers are mostly blind to what are the root causes. Using an interface to help to access these explanations may help people understand decisions better (HOLZINGER et al., 2017). Thus, the insights gained may improve future implementations, driving technology further.

Furthermore, the availability of decision‘s explanations make easier to evaluate other traits such as fairness, privacy, reliability, causality, and trust (DOSHI-VELEZ; KIM, 2017). This, in turn, pushes the field of artificial intelligence to a more humane and more robust, driving users adoption.

3.2.2 Product adoption

As a product embeds explanations in its core, it may be easier to have market adoption. Handing out an explanation to the user can establish a relationship of trust that leads the user to be more inclined to use the model‘s decisions (LOMBROZO, 2006). Furthermore, the use of explanations can facilitate learning (WILLIAMS; LOMBROZO; REHDER, 2013), which may lead doctors to better understand what is going on with the current and next patients.

Moreover, Lombrozo (2006) states that the use of explanations has several other benefits, such as persuasion. This can be of interest to an AI health-care system, since persuading the explainee on taking a positive action on a life-threatening issue may save his life. Persuasion that a decision is correct also generates trust among the users. Therefore, an explanation that is more persuasive can be favored, instead of a more descriptive one, if the goal is to have the user behave in a positive way.

3.2.3 Law compliance

Currently, the legal bodies from around the world are concerned with the rising of intelligent systems automating every aspect of our lives. This brings a troubling question: “Who will be responsible if it fails?” (DOSHI-VELEZ et al., 2017). Therefore, the need to accountability inside these systems has risen. Furthermore, the legal bodies initiated a move to build regulations around data and systems that leverage data to make decisions (GOODMAN; FLAXMAN, 2016; BRASIL, 2018; The Economist; Intelligence Unit, 2018).

Moreover, the General Data Protection Regulation (GDPR), that the European Parliament adopted in 2016 and is applicable since 25 May 2018, brings within itself a “right to explanation”. Although it may be to very limited contexts[2] it is expected that questions around explanations and AI systems to be important in future regulation and systems (GOODMAN; FLAXMAN, 2016). Therefore, explanations are important and necessary for the future of AI.

3.2.4 When it is not necessary?

Although interpretability may be a good thing to have when dealing with AI systems, there are some scenarios that it may not be necessary. These scenarios usually are the ones that do not have a significant impact on people‘s lives (e.g. Ads services), the problem at hand is a well-known one (e.g. optical character recognition), or the explainability might bring advantage to those with malicious ambitions (e.g. priority in social services). In all these cases the addition of interpretability should be thought carefully.

[2] The regulation states that the individuals have the right to question and object decisions made about them solely on the basis that it was used automated processing, but only when those decisions have significant/legal consequences (GDPR Art. 22).

3.3 The role of trust

Often when explainability is mentioned trust comes in a second plane, may it be as a consequence or as the goal (KIM, 2015; DOSHI-VELEZ; KIM, 2017; RIDGEWAY et al., 1998; RIBEIRO; SINGH; GUESTRIN, 2016b). However, as Lipton (2016) states, trust is a volatile concept that needs a definition of how it is treated. Furthermore, trust is an ever-changing state that is grown and cultivated (HOFFMAN et al., 2013). Therefore, we need to understand what is trust to people and then start thinking how to build trustworthy machines.

Trust is an interpersonal trait that is based on the willingness of a person (trustor) to be perceived as vulnerable by a second person (trustee) that may perform important actions to the first person (trustor) (MAYER; DAVIS; SCHOORMAN, 1995). Furthermore, this relationship is created and based on a few factors, that depending on the situation may have different importance. These factors mainly are understandability, perceived competence, benevolence or malevolence, and directability[3] (BRADSHAW et al., 2005). Therefore, the role of trust is to build relationships between people, establish an influence on the trustee and exchange pieces of information relevant to both parties.

3.3.1 Trustworthy machines

Seeing the relationship around trust, trustworthy machines are the ones who utilize concepts from the interpersonal relationships of establishing a connection. Therefore, the process to build trustworthy systems has much to do with how humans interact with one another, although it is quite distinct the factors that bring importance to the building trust. When dealing with trustable machines, these factors mainly are reliability, utility, robustness, false-alarm rate, and validity (MUIR; MORAY, 1996; BRADSHAW et al., 2005). However, when intelligent systems come into play it becomes more complicated to analyze which factors may help in building trust since the AI system is perceived as a mixture of both human and person (VISSER et al., 2012).

Seeing that, an intelligent system may run into some problems when building trust. One aspect of this is that humans tend to be much more unforgiving when dealing with failures from machines. Thus, it is much easier to break trust when a machine fails then when a human does (VISSER et al., 2012). Therefore, focusing on robustness is an important thing for a system that deals with sensitive decisions and is supposed to never fail.

Moreover, the mistrust on automation systems is inherent in people (HOFFMAN et al., 2013). They tend to develop a negative trust towards technology, thinking that systems are bound to fail and have bugs, that will inevitably slow down work process

[3] The degree to which the interlocutor has asserted influence onto the listener.

(KOOPMAN; HOFFMAN, 2003; HOFFMAN et al., 2009). So, an intelligent system that performs actions that sometimes may be obscure to some users should also focus on explanations, moreover, in good explanations, that transpires validity and engage with users.

3.4 What is a good explanation?

This section is devoted to unraveling what an explanation is and which traits a good explanation should have.

3.4.1 What is an explanation?

Taking the concept mentioned in section 3.1, a *why* question is composed of an inner whether question (with a yes or no answer), followed by a presupposition (BROMBERGER, 1966). An example of this is the *why* question "Why did the chicken cross the road?", that have the inner question "Did the chicken cross the road?", followed by the presupposition "The chicken crossed the road". However, why questions are intrinsically more complicated than that – they can be contrastive (why event a happened instead of event b) –, and this definition only exposes a fine perspective (OVERTON, 2011; MILLER, 2017). Therefore, this work will focus only on the definition that an explanation is an answer to a *why* question.

This simpler definition can constitute every-day explanations (or local explanations) that are explanations for particular facts, and answer why specific events occurred. So, the more scientific explanations, such as "Why rain falls down?" or "Why humankind have never filmed a UFO with a good camera?", are not the scope of this work, but are important to advance with the field.

Furthermore, Aristotle has defined a model that can be used as the basis to provide answers to why questions. This model was then called *Modes of Explanation* or Aristotle's *Four Causes* (HANKINSON, 2001). It states that a why-question may be asked with either a material, formal, efficient or a final purpose. Material purpose is when it is asked about the substance which the object is made; Formal is when the questioner wants to know the properties or form the subject; Efficient, or mechanistic, is when the near mechanisms of the subject cause something to change (e.g. the chef is an efficient cause for the food); Final, is the goal or objective of something.

Moreover, Overton (2012) created a structure for scientific explanations that may be used to structure explanation in general. In this structure, he defines five categories that can be explained by science (theories, models, kinds, entities, and data). The structure can be viewed in Figure 16. In his work, he states that for an explanation to be given at any level it must relate to any other, and every level in between.

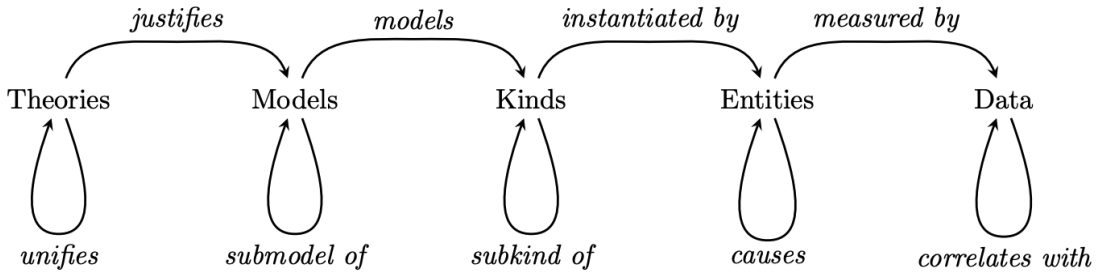


Figure 16 – Structure for scientific explanation containing five categories.

Source – Overton (2012).

To Overton (2012) theories is a set of principles that is the building block to models, that in turn are an abstract description of the relationship between kinds and their qualities. Kinds are an abstract universal class of entities, that are a concrete thing or process which is measured by a statement, or data. In this aspect, we can model that an SCC (entity) has a tone of red and scaly plaque (data), and every entity of this kind are non-melanomas, which agrees to the model of our theory of lesions.

3.4.2 A good explanation

Usually, when a system is developed the programmers are in charge of the software design decision, which leads to poor user design and experience. When dealing with explainable AI the products face the same risk that the needs of the user are neglected (MILLER; HOWE; SONENBERG, 2017). Furthermore, there is a need to put the users first, on how the explanations are perceived in their context and which aspects and traits of an explanation are more important. Therefore, we need to tailor explanations to be more human-friendly.

Miller (2017) in his work summarizes what traits are expected from a good explanation, based on the social sciences expertise. Furthermore, in this work, we try to summarize this list exposing what are the consequences to machine learning as a whole. Therefore, to achieve a good explanation it should be:

- **Constrastive**: People usually tend to understand decisions when there is an understanding of what could be changed to affect the outcome of the decision. An example is "How would the prediction be if the lesion were not scaly?". Constrastive (or counterfactual) explanations are easier to understand than complete explanations (LIPTON, 2016). Moreover, humans do not want complete explanations, but rather a comparison with other decisions. However, this approach is context dependent, since there is a need for points of data that serves as a reference to generate explanations. One good solution to achieve this trait is to build interactive tools, that let the user play and tweak with the data to see perturbations on the decision

(PAIR; Google AI, 2018; OLAH et al., 2018). This can be a much simpler solution that let the explainee understand the workings of the decision model.

- **Selected**: When humans try to pass down news or an information, they usually select few facts that are the core of understanding that piece of information. An example of this is news channels that when reporting a story they choose very few facts to build up an explanation that covers the cause of the event, although it might be thousands of possible explanations. This is called the "Rashomon effect", that describes an event may have multiple correct explanations to the same event (ANDERSON, 2016). Therefore, for an AI application, just a few reasons can be more impactful than the whole explanation of a complex world.

- **Social**: Explanations are an inherent part of a conversation or a social interaction, therefore the social context determines the content and ways the explanation is built. Furthermore, to explain why a lesion was diagnosed as melanoma to a dermatologist would be different from explaining to a lay patient. Moreover, applications must be designed thinking on the targeted audience, with nuances to fit their world models (LOVEJOY, 2018). A good method to transpire this is using conversational explanations, since "[...] the verb to explain is a three-place predicate: Someone explains something to someone. Causal explanation takes the form of conversation and is thus subject to the rules of conversation." (HILTON, 1990). This can also be used in text explanations when they are user specific (GALE et al., 2018). Therefore, AI applications need to be designed with a human-centered mentality.

- **Focused on abnormality**: Similar to the contrastive trait, humans tend to focus on abnormal causes to explain events (KAHNEMAN; TVERSKY, 1981) and consider them good causes (ŠTRUMBELJ; KONONENKO, 2011). Abnormal causes have very little chances of happening, but happened nevertheless (can be counted as a counterfactual explanation). An abnormal feature in a lesion would be a secretion, that might diagnose a lesion as a melanoma, although all other features resemble to a mole (WAJAPEYEE et al., 2008). Therefore, would be best to use this cause as an explanation to this case.

- **Truthful/Reliable**: Good explanations are proven to be true in other similar situations. However, reliability is less important than selectiveness, even when it omits parts of the truth (MILLER, 2017). An example would be that for a given case the truth is that there are hundreds of causes to explain why a lesion is diagnosed as a melanoma (e.g. radiation exposure, no use of sunscreen, genetic proneness, ...), however only a few factors may explain it.

- **Coherent with explainee beliefs**: Nickerson (1998) has proven that people tend to disregard information that is contrary to prior personal beliefs. This is a hard

trait to achieve since prior beliefs may be contrary to the truth sometimes and this can affect the performance of AI systems. An example is that a prior belief that if a lesion is bleeding it must be malignant can be refuted when a *pyogenic granuloma* is diagnosed. Therefore, this feature should be approached with caution.

- **General and probable**: People often prefer a general explanation that can explain many events (LOMBROZO, 2007). Moreover, a probable explanation is one that is expected that since event A happened it must be because of B. This can be easily measured an evaluated in a system by the number of instances which a given explanation applies over the total number of instances.

These criteria are important to any work wanting to establish an explainable AI. Giving explanations that are simpler increase the chances that the explainee both understand and accepts it. Additionally, it may be more useful to establish trust, if used with this purpose. However, to achieve this one should be able to measure the performance and evaluate its goals.

3.5 How to measure an explanation

Taking into account the properties listed in subsection 3.4.2, it is not clear how to measure them correctly, so one of the current challenges is to formalize how they could be calculated and demonstrated in mathematical formulae.

Doshi-Velez and Kim (2017) proposes a new view on how to evaluate explainability. They list three levels that can be used as validation phases or publics to different AI systems context.

A taxonomy of evaluation approaches is proposed by Doshi-Velez and Kim (2017) that is analogous to approaches that are already proven an used in machine learning field. Therefore, they defined the levels in application-ground, human-ground, and functionally-ground, being the latter less costly and less specific (see Figure 17).

- **Application level**: This level is thought with respect to how we evaluate an explanation in the context of its end task. This takes into account how the explanations will perform when put to test with real end-users. In the context of this work, it would be to have dermatologists use the explanations in their daily practices. This is a costly experiment as the is a need to evaluate and benchmark how well a *human-produced* explanation helps a professional in their practice. This evaluation usually also require the need for a viable product to be tested. This requires a well structured experimental setup. As Antunes et al. (2012) states, this kind of evaluation is

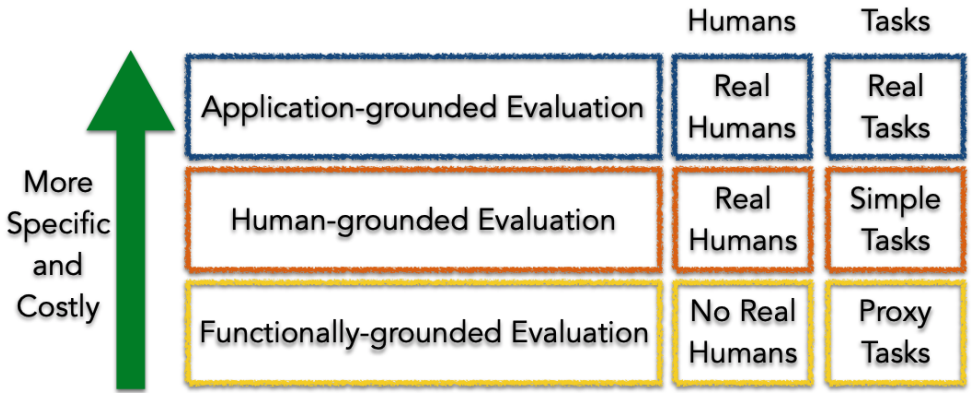


Figure 17 – Taxonomy of evaluation approaches for explainability.

Source – Doshi-Velez and Kim (2017).

not an easy task. However, it is the most complete that test the main objective and thus evaluate the true performance of the explanation.

- **Human level**: This is similar to application level evaluation, however in human-ground testing is done with lay humans instead of experts. This can reduce drastically the costs of an experiment and increase the subject pool. Moreover, this evaluation is ideal when the experiment wants to test more general ideas of the explanation quality. Ideally, the evaluation approach only depends on the quality of the explanation provided. This can be tested in a mode that lay humans are presented with a pair of explanations and asked to judge which one is the best.

- **Function level**: This evaluation does not require human experiments, instead, it uses a formal definition of explainability to test an explanation quality. This formal definition serves as a proxy. Therefore, it is useful when there is a class of models that have already been tested and validated via human-grounded experiments. This brings the challenge to determine what kind of proxy to use. An example would be to use decisions trees to evaluate explanations since it is known that they are interpretable in many situations (FREITAS, 2014).

3.6 Interpretability methods

Seeing the definitions, importance, features, and evaluation of interpretability, now is important to review some possible methods to build interpretability in AI systems. First of all, interpretability can be seen on two levels of implementation, global and local. This work has focused mainly on using local interpretability concept examples to explain what interpretability is.

Moreover, local interpretability can still be subdivided for single predictions and for group predictions. The first is when we take a single decision, or instance, and examine it

singling it out to explain the outcome. This can be more accurate to explain the decision since there is a good chance that the local distribution of the target behaves linearly. Furthermore, the group predictions analysis is simply an aggregated list of single instances.

Additionally, global interpretability is more complex due to the fact that we need to comprehend the whole model at once (LIPTON, 2016). This refers to the trained model, the knowledge about the algorithm, and the data. Furthermore, to explain the whole model and how it takes its decisions, it is paramount to have a holistic view of the weights, parameters, features, and structures. However, any model – even linear regressions (that are interpretable by nature) – that have more than three features and thus exceeds the three-dimensional space, are fundamentally hard to be imagined by humans.

Therefore, when talking about interpretable models, like linear regression, we do not expect to have global interpretability as described above. However, the term "interpretable models" come from the ability that these models have to facilitate the explanations of single weights and distribution of the features. Consequently, the weights only make sense in the context that the other features are inserted. Examples of interpretable models would be linear regression; linear models (e.g. general additive models, generalized linear models); decision trees; if-then-else rules.

These models, often called shallow models, are known to be less powerful than deep models in many tasks (MURPHY, 2012). So, there is a trade-off here, the researcher should choose between accuracy or model interpretability (JOHANSSON et al., 2011). However, for all the other models other than those that are inherently interpretable, there are model-agnostic methods to be used in more complex models such as convolutional neural networks.

3.6.1 Model-agnostic methods

Agnostic methods are the ones that do not care about what is being applied to, they have the flexibility to adhere to many contexts. Thereafter, these methods are by definition separated from the models they are being applied to, as is a representation that is built on top of it. Furthermore, doing this separability have some benefits such as model, explanation and representation flexibilities (RIBEIRO; SINGH; GUESTRIN, 2016a).

Model flexibility is straight-forward, one can use as many models as it wants without having to worry about the interpretability method changing. This benefit is very useful when we are dealing with innovation and research tasks that require many experiments. Explanation flexibility in regard to the possibility that one model may be used by several different methods, each one outputting different types of explanations. With this targeting different audiences that need different explanations is possible. Finally, representation

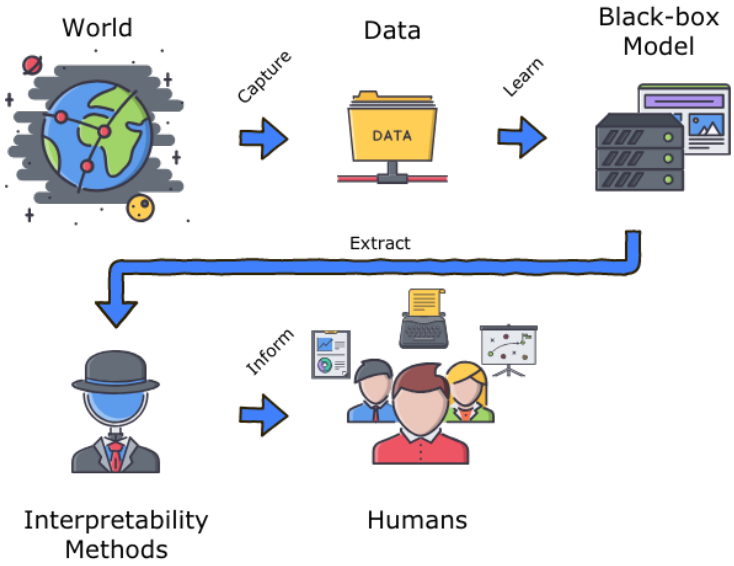


Figure 18 – The big picture of explainable AI. The path that the world features has to go through until it reaches the human as explanations.

Source – Adapted from Molnar (2018).

flexibility is the possibility that one has to shape how to present the explanations.

Therefore, we can represent the general path that the information travels until it reaches the humans in the form of explanations. First, the information is captured from the world, which can be anything that we want to understand and interact. Then this information, or data, is put in a digital form to be processed by computers that will learn and produce a model, or black-box model, that abstracts the patterns and learn from data. From this model, we can use interpretability methods to extract explanations from the opaque model. Finally, the explanations can be processed and represented in different formats to be provided to humans. Figure 18 summarizes this general path.

Furthermore, the model-agnostic explanation methods can be subdivided into two categories. The first are methods based on gradient approaches, which are focused on neural networks explainability, and the second are approaches that use input-perturbations to generate explanations (ROBNIK-ŠIKONJA; BOHANEC, 2018). Gradient-based approaches use the computation of gradients outputted by neurons with respect to the input. Input-perturbations, as the name suggests, use the input with small perturbances to test whether the output of the model is changed.

3.6.1.1 Perturbation-based approaches

Perturbation-based approaches, perturb the input and evaluate the consequences in the output of the model. These perturbations remove small pieces of information from specific regions of the input by applying noise (e.g. blur (FONG; VEDALDI, 2017)). Due to this nature, these methods are inherently more computationally costly than gradient-

based approach.

Furthermore, it is difficult to choose a perturbance that removes information without adding any new information. The simplest form is replacing the region with a gray square (RIBEIRO; SINGH; GUESTRIN, 2016a). This method inserts some problems, that inserting a grey square on an image may add the chances of the model to output a high confidence for wrong classes. An example of this would be a classifier that has a concrete-road or elephant class, thus a gray square would increase these classes confidence.

Moreover, if the perturbations applied to the input are too small then there is a chance that uninterpretable and arbitrary image regions of the image are highlighted. To mitigate this problem some methods use larger regions of the image, thus becoming less precise (FONG; VEDALDI, 2017).

3.6.1.1.1 LIME

LIME (Local Interpretable Model-agnostic Explanations) is a method introduced by Ribeiro, Singh and Guestrin (2016a), which implements a concrete local-surrogate model that explains individual predictions. LIME use perturbation-based approaches to train a new interpretable model (e.g. Lasso model) that is weighted by the proximity of the perturbed image samples to the instance of interest. This learned model should have a good local interpretable representation that is locally faithful.

LIME can be expressed as the mathematical equation 3.1, where x is the instance to be explained, g is the interpretable model (e.g. linear regression), L is the loss function, or fidelity function, (e.g. mean squared error), that computes the explanation proximity to the prediction of the original model f (e.g. CNN), while the model complexity $\Omega(g)$ is low – favoring fewer features. Moreover, G is the family of potential interpretable models, such as decision trees. The proximity measure π_x defines the neighborhood size around the original instance x. In practice, LIME only optimizes the L function since the user has to define the complexity needed.

$$\text{explanation}(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{3.1}$$

Finally, the general idea behind the algorithm is that chosen an instance of interest, use perturbations in the dataset to get predictions from the black-box model, store the predictions and weight the new samples by their proximity to the original sample. Sequentially, fit a weighted, interpretable model on the new dataset and, finally, explain the prediction by interpreting the local model.

An example of this process can be viewed in Figure 19, where the blue/pink background is the black-box model fitted function. The big red cross is the instance of interest

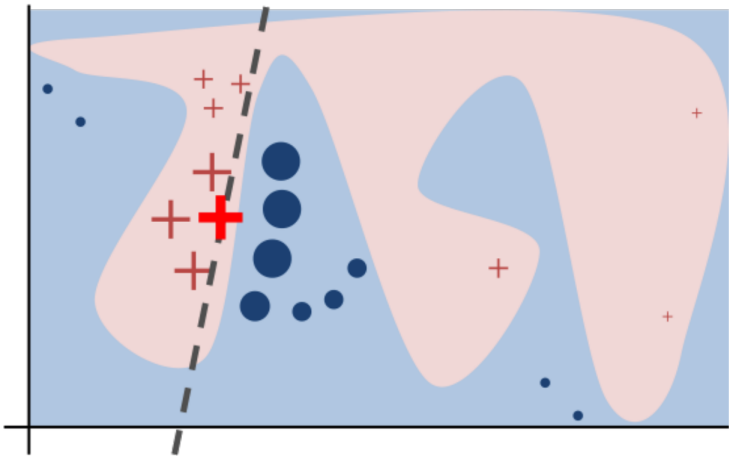


Figure 19 – Example to represent the method implemented by LIME.

Source – Ribeiro, Singh and Guestrin (2016a).

x that is being explained. The perturbed images are distributed around the original instance and weighted accordingly to their proximity, represented by size. Then a model (dashed line) is fitted to the predictions of the original model (red crosses and blue circles). Note that the explanation is locally faithful, however, it does not explain the model globally (see the small red cross on far-right).

3.6.1.2 Gradient-based approaches

Gradient-based, or saliency methods, are the most popular local explanation methods for image classification (ERHAN et al., 2009; SMILKOV et al., 2017; SUNDARARAJAN; TALY; YAN, 2017). In these techniques, an image is generated based on the importance of each and single pixel in the input image for the decision outputted. These methods are good in explaining single samples outcomes, however when dealing with knowledge-based applications to obtain global insights on single classes, it difficult to do so automatically.

Moreover, saliency methods detect which parts of a given image are more relevant to the outcome. This can be obtained through a technique that computes iteratively the smallest parts whose occlusion affect most the decision score (which would be a perturbance-based approach) (DABKOWSKI; GAL, 2017). However, this can be time and computationally consuming. Other approaches would be to train a model to predict the regions (DABKOWSKI; GAL, 2017) or compute the regions using mathematical methods (SELVARAJU et al., 2016; SMILKOV et al., 2017).

This work will focus on mathematical methods for gradient-based approaches. This decision was made based on (1) applications of traits listed in subsection 3.4.2; (2) how much time-consuming is the method. The reason for the latter is due to the applicability of this method in a real-world system, in which doctors and patients should be able to

get a swift feedback.

3.6.1.2.1 GradCAM

Gradient-weighted Class Activation Mapping or GradCAM, proposed by Selvaraju et al. (2016) is a generalization of CAM (class activation maps), given by Zhou et al. (2015), agnostic to the architecture of a CNN. Moreover, GradCAM makes possible to obtain the localization map of any target class in a model. The general idea of this method is to try to directly use the activation maps of the final convolution layer to infer the relevance map of the input pixels. Therefore, the final heat-map is calculated from the feature maps (or filters/kernels) in the final convolutional layer.

Furthermore, in order to calculate the class localization map the method computes the gradients of the score y^c for class c with respect to feature maps $A^1, \dots A^k$ produced by a convolutional layer. Then we take these gradients and apply a global-average-pooling function to obtain the weights α_k^c.

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}} \tag{3.2}$$

Furthermore, the score y^c does not need to be the predicted class but can be any of the available classes. Moreover, with the computed weights it is then possible to do a weighted combination of the feature maps A, similar to the CAM algorithm. However, for GradCAM a ReLU activation is applied to normalize the output to only display positive values.

$$L^c_{\text{Grad-CAM}} = ReLU \underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}} \tag{3.3}$$

The process can be seen in Figure 20, where is displayed the computations to produce a heat-map. Additionally, we are only interested in the bottom part of the diagram since the top part is an additional computation that adds the values calculated with guided backpropagation to then modulate the input features in a fine-grained image.

3.6.2 Conversation as an explanation

Although model-agnostic methods may be good to disseminate explainability as a whole, some representations of explanation may not be the best fit with we take into account the traits for a good explanation. Furthermore, to use a good explanation is often needed to tailor a specific approach for each application.

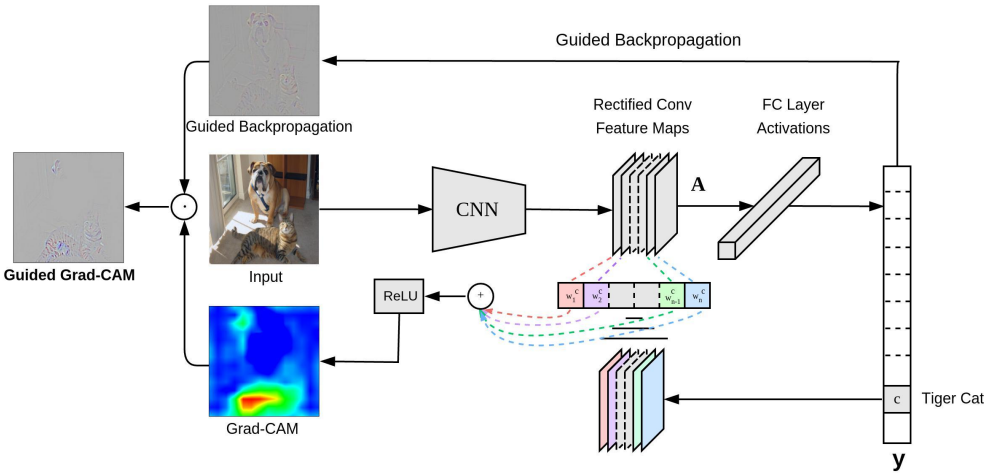


Figure 20 – GradCAM overview.

Source – Selvaraju et al. (2016).

When building an application for people is only fair to put them in the first plane of importance (LOVEJOY, 2018). In doing that we, as developers and researchers, have to abide by the rules of causal explanations, that are followed in human-to-human interactions. Furthermore, causal explanations usually occur in the form of interactive conversations, this can be seen in a piece of Hilton‘s work.

> “Causal explanation is first and foremost a form of social interaction. One speaks of giving causal explanations, but not attributions, perceptions, comprehensions, categorizations, or memories. The verb to explain is a three-place predicate: Someone explains something to someone. Causal explanation takes the form of conversation and is thus subject to the rules of conversation.” — (HILTON, 1990).

It is important to remember that as conversations are guided with wide accepted rules. Grice (1975) rules of conversation states that people should only say what they believe, stating only the necessary, when it is relevant, and say it all in a nice way. And seeing explanations as conversations (HILTON, 1990), they should follow these maxims.

For experts in human image analysis is often needed to explain their findings and assessment of an image to other experts in the field, patients and colleagues. In general, these explanations come in the form of conversations, that experts are asked to express, first, what part of the image they are referring to, second, what features they are seeing in the image, third, to synthesize a conclusion (GALE et al., 2018). Figure 21 shows how this method can be integrated with visual explanations.

When bringing this issue to AI systems this can be attained with some methods that create captions/descriptions to images (KARPATHY; FEI-FEI, 2015; VINYALS et

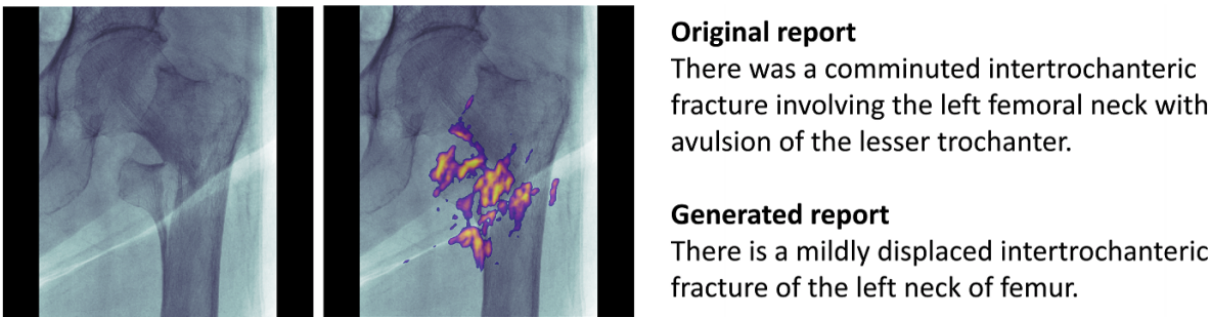


Figure 21 – Example of image captioning in the radiology field.

Source – Gale et al. (2018).

al., 2015; WANG et al., 2018b). This can be a short, but descriptive information that helps users to understand what is being shown in the picture.

These methods usually use a recurrent neural network to learn a text embedding that is then capable of generating new texts. Furthermore, the text generation model is trained on report input that accompanies the images. So, there is a need for datasets that not only contain images but short descriptions of each. Therefore, although the great level of usability, this method of explainability generation is the most time and resource consuming.

Despite the fact that this method seems to best to apply in this work, we lack the description of each image an although we could annotate each one by hand, it is not feasible with the given time. Therefore, we will not focus on this possibility and leave it for future work.

4 Results

Have seen all the techniques and information gathered around the approaches and resources, now we reserve this chapter to discuss the achieved results.

4.1 Dataset

As seen in section 2.3, the datasets of clinical images for skin lesions publicly available are just three. Therefore, the experiments performed on this work leverage on those datasets for its results. Therefore, we gathered all three datasets and merged them into a single database. Moreover, the dataset was divided into three separated directories following the division of 20%, 10%, and 70% for testing, evaluation, and training, respectively.

However, by a mistake made in the time of the experiment, only the training and testing directories were used for the experiment. This left 533 evaluation images unused, using only 3,797 images for training and 956 for testing. Moreover, the training dataset was augmented, using the transformations cited in subsection 2.4.2.1, by a factor of 29 times. The testing dataset did not suffer any transformations.

With this oversight of the evaluation images directory, the training dataset was further divided with a proportion of 80/20 between training and evaluation datasets. Furthermore, the final numbers for the datasets can be seen in table 4.

Table 4 – Number of images used in the dataset for the final experiment.

	Number of images		
Lesion Type	***Train***	***Validation***	***Test***
Actinic Keratosis	742	186	8
Basal Cell Carcinoma	30,067	7,517	324
Dermatofibroma	1,067	267	12
Hemangioma	1,601	400	18
Intraepithelial Carcinoma	1,299	325	14
Lentigo	1,137	284	13
Malignant Melanoma	6,218	1,554	68
Melanocytic Nevus (mole)	17,632	4,408	191
Pyogenic Granuloma	371	93	5
Seborrheic Keratosis	19,256	4,814	208
Squamous Cell Carcinoma	1,462	365	16
Wart	7,238	1,810	79
TOTAL	88,090	22,023	956

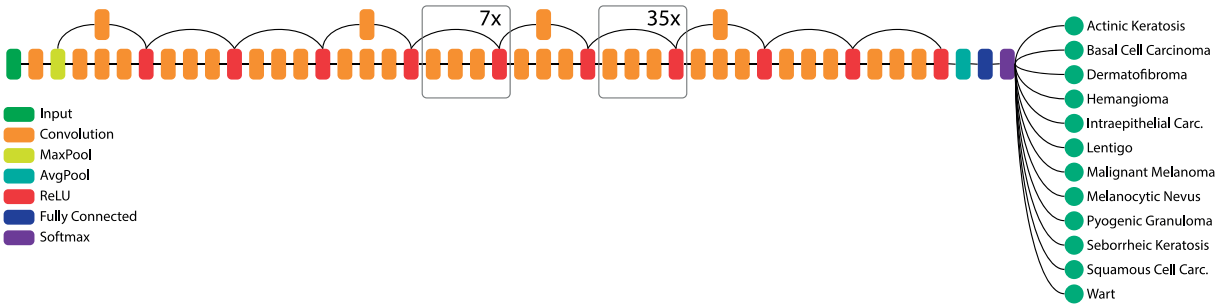


Figure 22 – ResNet-152 architecture used.

Source – Author.

4.2 Infrastructure

All the experiments were conducted under the same environment, that consisted of Antergos 18.3 (Linux kernel 4.16) running BVLC Caffe (JIA et al., 2014b) with support for an NVIDIA GTX 1070 GPU (Cuda 9.1 and cuDNN 7.1).

4.3 Classification

After knowing the possible architectures to use and what are the characteristics of the datasets it was time to gather all this knowledge and implement some solutions. The best results achieved on this work were with the use of the ResNet-152 architecture, trained over an augmented dataset with a mixture of MED-NODE, Edinburgh and Atlas datasets. The augmentation made was of 30x for each class, leaving the classes unbalanced.

Furthermore, the ResNet architecture had to be modified to accommodate the needs of the problem at hand. So, the last layer of the architecture was changed from 1,000 classes to 12 classes, the 12 skin lesions. Therefore, the final architecture produced followed the same schema seen in Figure 22.

4.3.1 Training process

Moreover, the technique of transfer learning was applied to generate the best results more rapidly. For that, the hyperparameters of the network had to be tuned and carefully set, for that same purpose.

4.3.1.1 Hyperparameters

The process of choosing the hyperparameters passed through arduous steps of research and trial and error. This was mainly due to the fact that “there is no such

thing as free lunch" in machine learning. So, this process started in the definitions of the hyperparameters used by Seog Han et al. (2018) in their work.

They stated that the learning rate used was a low learning rate since it was expected to keep the middle layer's weights from altering too much since they had been trained extensively on the ImageNet dataset and judged to be well defined. However, for this work the only hyperparameters that stayed equal between these works were: $weight\_decay = 0.00001$, $momentum = 0.9$, and $gamma = 0.1$.

Furthermore, all the hyperparameters were defined in a separated configuration file called "Solver". This was needed since the experiments were leveraging on the Caffe framework to train the DNNs, and it made necessary to define a *.prototxt* file with free parameters. This file can be seen on Appendix C.

Therefore, for the iteration finding process the final modified hyperparameters were:

- ***batch_size***: This hyperparameter sets the number of examples that will be saved in memory to be processed by the neural network in the same forward pass. Thus the limiter of this value is mainly hardware (GPU memory or Random Access Memory capacity). For the infrastructure used in this work the limit of parallel images load, alongside the weights of the ResNet-152, was 5 images.

- ***max_iter***: This refers to the maximum iteration cycles to be done. This defines the number of *epochs* to be run in the training. This can be calculated with the formula: $max\_iter = \frac{N}{batch\_size} \times epoch$; where N is the total number of examples in the training dataset. Thus with a train size of 88,090, *batch_size* of 5, and a training time of 10 *epochs*, the *max_iter* was equal 176,180.

- ***test_iter***: The *test_iter* parameter is responsible to set the number of iterations done in the test dataset, used in the test phase. Thus, a number of 22,023 iterations was set, since the test phase has a *batch_size* of 1 and only runs for 1 *epoch*.

- ***test_interval***: It is responsible to set the frequency of the test cycles done in training time. This value is configured in iterations, so it is calculated on top of the *max_iter* hyperparameter. To check the number of test cycles done in training the division of the maximum number of iteration by the test interval. For the experiment, an interval of 2,000 was chosen. It is good to be mindful that the more frequent the test cycle is the longer the training phase will take.

- ***stepsize***: This refers to the frequency that the learning rate will decay, by a degree equals to the *weight_decay*[1]. It was used a *stepsize* of 17,618, that means that after every *epoch* the learning rate was updated.

[1] Using the 'step' learning rate policy.

- ***iter_size***: The *iter_size* is responsible to alleviate the pressure on the low-end hardware. This is done in a way that this parameter holds the update of the gradient by n, the *iter_size*. Therefore, with this, hardware with low memory can mimic a larger batch. The final batch used to update the gradient is given by the multiplication $batch\_size \times iter\_size$. The number chosen for this hyperparameter was 12, since it was what approximated the final *batch_size* to 64. Although using this hyperparameter may affect the batch normalization layers used in the architecture, the final results did not show this effect.

- ***base_lr***: This is the start learning rate set for the network, used by the optimizers to update the learnable weights. For the experiment was used a start learning rate of 0.01.

The learning rate was chosen to be higher than the used in the related works, for two major factors. First of all, the past experiments showed that with a low learning rate, a plateau on the very start of the training was found. Thus, the network did not have the power to learn the features of the skin lesions. Secondly, it was found that increasing the learning rate often aids to reduce underfitting (SMITH, 2018). Therefore, with the past experiments, the number of 0.01 was found. However, to the counterpart, the high learning rate, the *stepsize* was decreased to one *epoch*, and a frequent test phase was used to monitor the network learning.

The batch size was not chosen by experiments, but rather by the hardware limitations. However, the *iter_size* was chosen to achieve a batch size close to 64, that was a size found to accelerate the training process towards the final results, but not deteriorate the stability of the network (MASTERS; LUSCHI, 2018).

4.3.1.2 Model Training

For the training process, it was used the technique of transfer learning with the approach of fine-tuning the network. Thus, it was necessary to gather the ResNet-152 pre-trained weights for the ImageNet dataset[2] first, and then modify the network for the purpose of this work.

Moreover, seeing the work done by Seog Han et al. (2018) and the method used to increase the fully connected layer learning rate. The final dense layer has a 10 times factor of multiplication for the learning rate, compared to the other layers of the network. However, different from the process of freezing the early layers, used in the same research, this work approximates more to the approach implemented in Esteva et al. (2017), that fine-tuned all the layers of the network.

[2] Available at <https://github.com/KaimingHe/deep-residual-networks>. Last accessed on June 26th, 2018.

This was done with the premise in mind, that although the ImageNet dataset is far diverse and comprehends many different objects, it does not have classes that approximate in characteristics and problems encountered in this dataset of skin lesions. Furthermore, it the weights in the early layers may not be properly trained to extract fine features such as the ones found within the problem that is faced in this work. Therefore, it was needed to fine-tune the learnable parameters since the early layers and learn the final classifier from scratch.

Another fact that it is worth mentioning is that, although the number of maximum iterations used was a 10 *epochs* iteration size, the training was not concluded after the full completion of these iterations. The training was early stopped, since, around the iteration number 30,000 the loss function, both in training and validation, did not alter significantly. Therefore, it was judged that the horizontal part of the validation loss was achieved, thus a good convergence of the network (SMITH, 2018).

Finally, the model used in the testing phase was the product of the iteration number 38,000. This training phase took an uninterrupted total time of 35 hours (approximately 167 seconds for every 50 iterations).

4.3.2 Results

For this phase, the model generated in training was submitted to analysis with the testing dataset. Furthermore, the metrics defined in subsection 2.1.4 were used to analyze the predictions of the model.

With the confusion matrix generated for the predictions in the testing dataset, was found that for all the 11 lesions, with exception of the *Actinic Keratosis*, achieved a accuracy higher than 80%, using the formula shown in equation 2.7 (seen on Figure 29 in Appendix D), thus accounting for a 78% total accuracy for the model. However, this metric has a bias attached to it, since the distribution of the classes is not even and therefore can cause misleading in the analysis of this metric.

One takeaway from this matrix is the trouble that the model has to predict some class. When in the same row two classes have high color values, it means that there is a high error rate in the row, caused by a confusion between these classes. Thus, it may mean that the two classes have some shared characteristics that cause this phenomenon.

Furthermore, the classification report was calculated, this gives us the recall, precision and f-1 score for the individual classes as well as the total average. This values can be seen on table 6 in Appendix D.

Finally, the AUC and cut-off values for each ROC curve have been calculated. This metric is common among many kinds of research that deals with classification of diagnostics. Moreover, this metric has been used in the researches used as guidelines to

Table 5 – Comparative between AUC metrics.

Lesion	Esteva et al. (2017)	Seog Han et al. (2018)	This work
Actinic Keratosis	-	0.83	**0.96**
Basal cell carcinoma	-	0.90	**0.91**
Dermatofibroma	-	0.90	0.90
Hemangioma	-	0.83	**0.99**
Intraepithelial carcinoma	-	0.83	**0.99**
Lentigo	-	0.95[3]	0.95
Malignant Melanoma	**0.96**	0.88	**0.96**
Melanocytic nevus	-	0.94	**0.95**
Pyogenic granuloma	-	0.97	**0.99**
Seborrheic keratosis	-	0.89	**0.90**
Squamous cell carcinoma	-	0.91	**0.95**
Wart	-	**0.94**[3]	0.89

quantify the quality of the trained models. Therefore, the table 5 shows a comparison between the results in these three works.

The work done by Esteva et al. (2017) was faced with a final binary classification on benign and malignant lesions, thus the metric that was fair to compare was only the AUC metric for Melanoma lesions. However, the work done by Seog Han et al. (2018) used a trained model for the exact same lesions, thus it was fair the comparison. Additionally, the results of this work seen on table 5 can be further examined in graphics on Appendix D.

Furthermore, for the plots of the AUC for each class, please refer to Appendix D.

4.4 Interpretability

Additionally, to results in the classification task, this work achieved some interesting results in the model interpretability task. The results are shown and discussed below.

4.4.1 Results

For the results, we bring three kinds of discussions, first a discussion based on the raw predictions of the models. This is the easiest task to achieve since it only leverages on the predictions of the already trained model. Second, discussion on visual explanations using the GradCAM technique. Finally, we bring some examples which the visual explanations brought some insights into what may be happening on the network.

[3] Metric calculated with an Asian dataset, thus may not serve as a comparative in a *stricto sensu*.

4.4.1.1 Prediction analysis

An analysis of the model predictions was made, as a way to discover which were the examples in the validation dataset, that the model most got right, wrong and was undecided about. Furthermore, it was seen that the images that were mostly right, were sharp clear images, that centered the lesion in a good way, and did not have hair or other objects causing occlusions of projected shadows.

For the most wrong predictions, it was found that the causes may fall under 3 factors: the lesion analyzed indeed caused confusion between the lesions (Figures 34a and 34c); the image had some other features that added noise to the input (limbs, facial characteristics, nails, etc) (Figures 34d and 34f); the lesion was either occluded by some other object or was far away and not centered in the image (Figures 34b and 34e). These pictures can be seen on Appendix E.

The undecided lesions brought more information to what the model is struggling with, that is necessary to further dedicate more time to the analysis of what it might be causing the undecided cases such as Figures 35d and 35a.

4.4.1.2 Visual explanations

Furthermore, the GradCAM method was applied for this work, however, for that, it was necessary to implement the code to utilize it in python alongside with Caffe, since the official repository[4] utilizes LUA language for its solution. This implementation can be seen on Appendix subsection E.4.

Furthermore, it was found that the model generalizes well for the examples that it was shown, correctly activating the regions that contained the lesion, even in images that the pose made challenging the localization of the lesion. Moreover, the Figure 23 shows an example of such pose, that the nose in an acute angle distances the lesion from focus, in spite of this, the model is able to detect not only the class but the localization of the lesion.

Also, it was done an experiment in which we analyzed two groups of malignant melanoma. The first consisted of the most false-negative malignant melanomas, to see whether the model was looking for the right features in the image. Second, it was analyzed the most true-positive malignant melanoma, that were used to confirm the generalization of the model. All images that were tested here with the GradCAM method were taken from the testing set.

For the most false-negative predictions, it was found that the causes may fall under 2 factors: the lesion analyzed indeed caused confusion between the lesions (Figure 24a); the model did not generalize well and was struggling to extract predominant features in

[4] Available ai <https://github.com/ramprs/grad-cam/>. Last accessed in 28/06/2018.

Figure 23 – GradCAM appplied to a *Basal Cell Carcinoma* lesion.

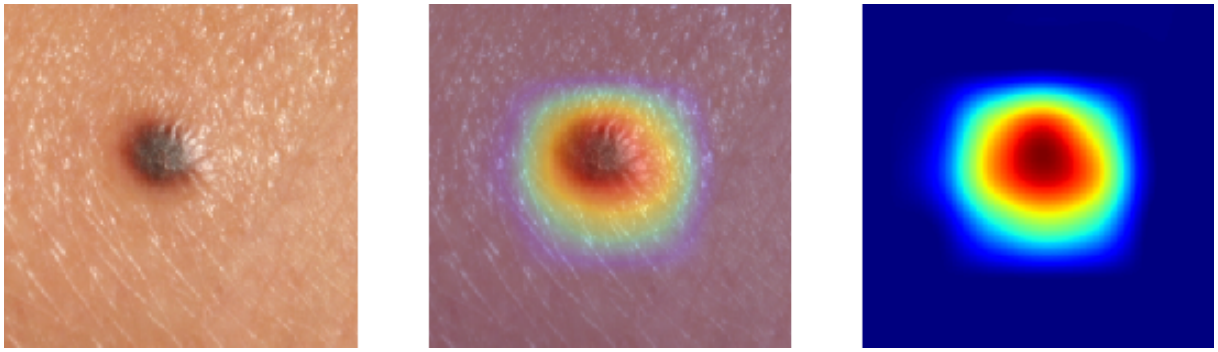
Source – Author.

some of the images, thus giving more importance to areas that were not relevant, from a practical perspective (Figures 24b and 24c).

The most true-positive lesions brought more information on what the model was already good at, and how this translates in a perspective of image features. For example, in Figures 25a and 25b, the model found out that the regions that mostly identify the lesions as Malignant Melanomas are indeed the ones that would bring more relevance to the doctor‘s decision-making process. However, there are still some examples, such as Figure 25c, that are more emblematic and need an expert‘s eye to shed a light on it. Moreover, we can speculate that the model took advantage of the geometric and color asymmetry in the lesion to make an accurate decision. This trait can also be seen in Figure 26 that we took, and thus is not from any of the datasets, further demonstrating that the model indeed generalized well.

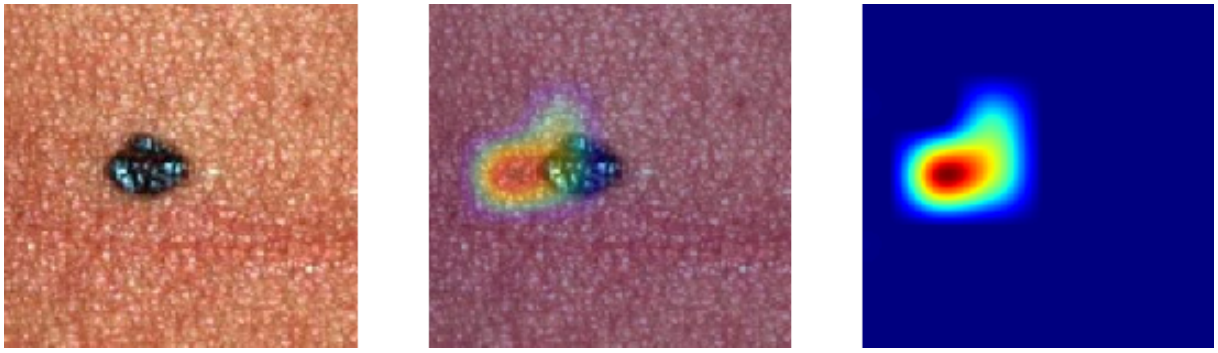
Furthermore, in this work, we observed interesting features learned by the model that were effects from the data collection process made for Atlas dataset. In this dataset, we can observe some pictures that present unique poses and viewpoints for lesion images. This inserted on the model unique features that were only observed in these pictures since they were unique and no other image shared the same features. One example of this is the long distance shot of a human back, that happens to present a lesion. Similar to this pose there is only another image, taken on the same physical location and pose, only changing the human in the picture. However, both subjects presented the same lesion, malignant melanoma. Figure 27 shows the images discussed, being 27a from the test dataset, and 27b from the training dataset.

This unexpected event made the CNN memorize the predictions for this kind of image taken in the same pose (overfitted for these images). This was done in such a way that the most distinguishable feature presented in the image was the one that the activations were picking-up to output a 100% certainty prediction for both of them.



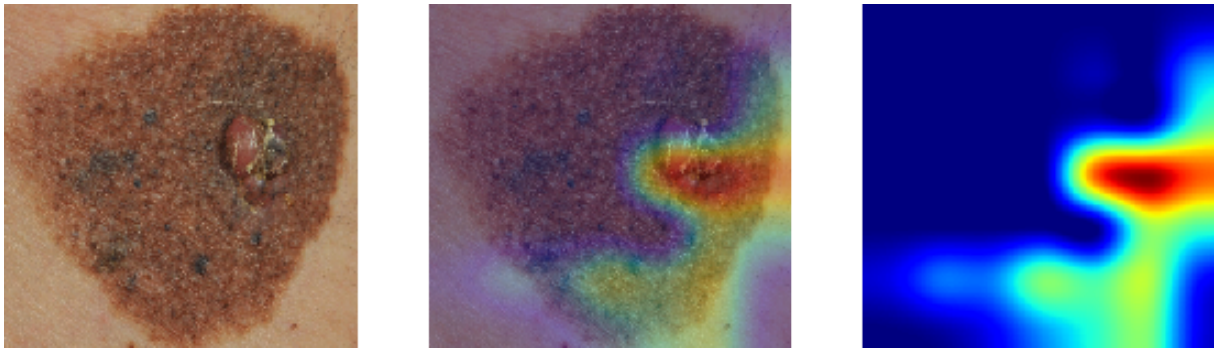
(a) Predicted as Haemangioma with 100% confidence.

Source Edinburgh dataset.



(b) Predicted as Basal Cell Carcinoma with 100% confidence.

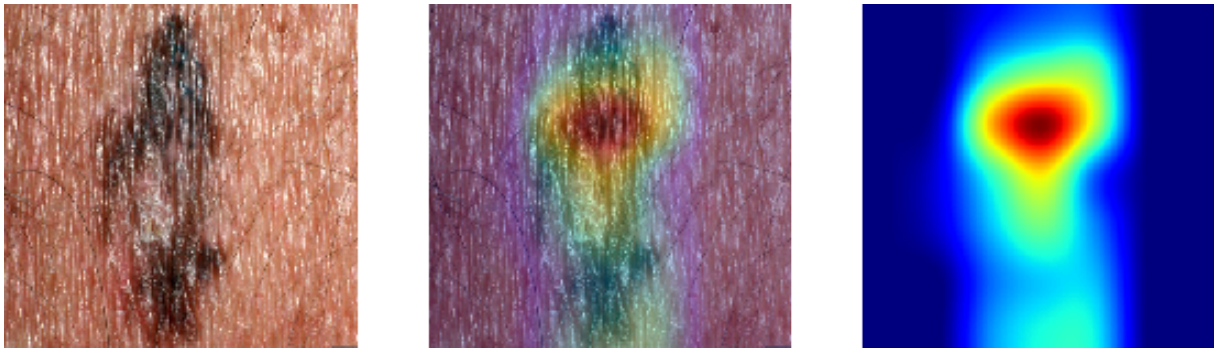
Source Edinburgh dataset.



(c) Predicted as Melanocytic Nevus with 97% confidence.

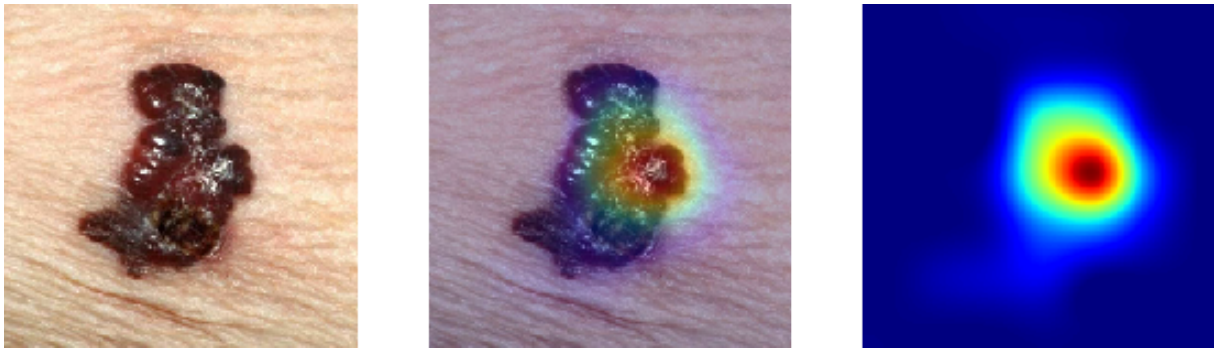
Source MED-NODE dataset.

Figure 24 – Most false-negative predictions for Malignant Melanoma. Columns from left to right: original image; original image fused with heat-map; heat-map produced by GradCAM.



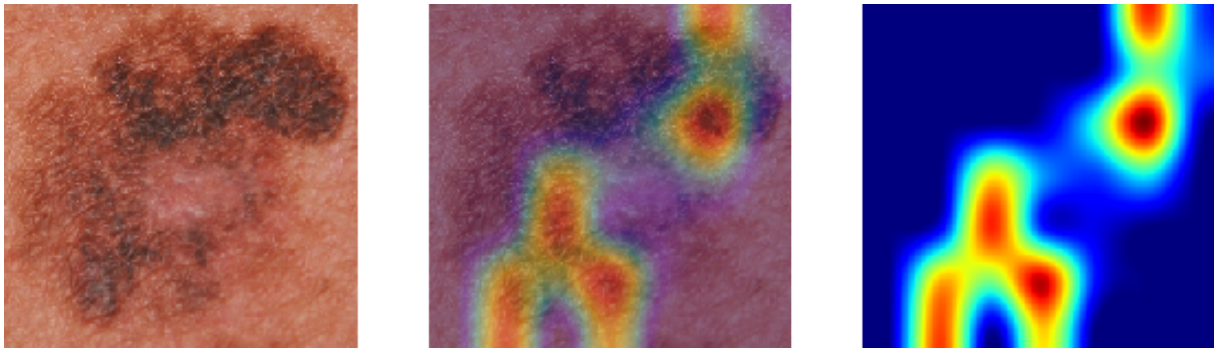
(a) Melanoma with 100% confidence.

Source Edinburgh dataset.



(b) Melanoma with 100% confidence.

Source Edinburgh dataset.



(c) Melanoma with 100% confidence.

Source MED-NODE dataset.

Figure 25 – Most true-positive predictions for Malignant Melanoma. Columns from left to right: original image; original image fused with heat-map; heat-map produced by GradCAM.

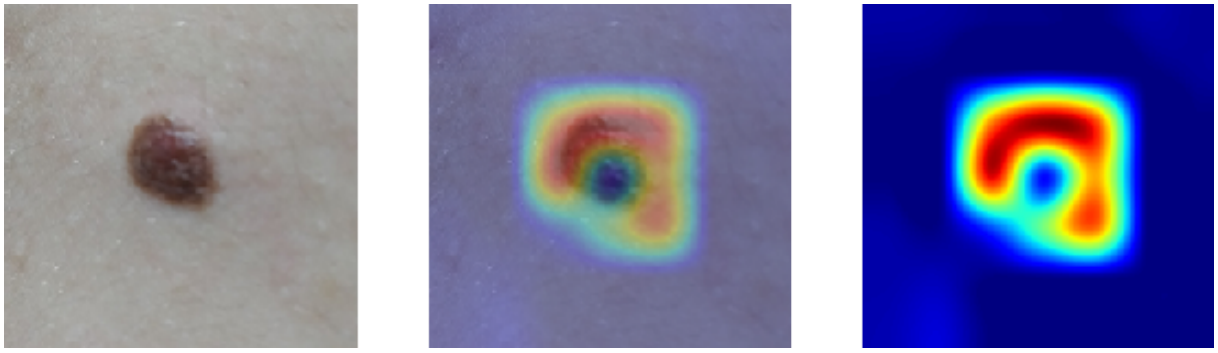


Figure 26 – Visual explanation applied to skin lesion.

Source – Author.

(a) Torso pose of subject one.

Source Atlas dataset.

(b) Torso pose of subject two.

Source Atlas dataset.

Figure 27 – Unexpected poses for clinical images of skin lesion.

Therefore, as seen in Figure 28 the light-blue background wall was the most distinct feature that described both images.

These images show us how powerful a neural network can be. Only with two samples that resembled the same features, it was capable of learning (too well) and predicting correctly. Also, it is possible to see that the network also managed to pick some activations on the single lesion in the center of Figure 28a. This tells us that it is also not ignoring the single point that – smaller that it is – approximates from the rest of the dataset. From this insight, we can investigate and learn new methods to build a more robust network or refine our methods of data pre-processing.

This technique makes especially easy to detect whether the model has generalized well over the problem or if it is detecting biases in the dataset. Nonetheless, this kind of interpretation is important not only in a developer vision but also for the possible doctors that were to receive a simple prediction and to take action on another human life based on it. With these other tools, the doctor has much more to base the next decision that is necessary to take on the patient. Therefore, making the models as interpretable

(a) Visual explanation of torso pose of subject one.

Source Atlas dataset.

(b) Visual explanation of torso pose of subject one.

Source Atlas dataset.

Figure 28 – Unexpected explanation for images.

as accurate turns a tool in a good counselor.

5 Conclusion

Skin cancer is the most common kind of cancer in society today, representing a tremendous health and economic problem. Furthermore, the dermatology field has approached the problem always in similar ways, by examining each patient individually either with the naked eye or the aid of some magnifying tool. Seeing the advances in the field of machine learning some opportunities arise from this situation in the form of skin lesion classifiers.

These opportunities show themselves as solutions to aid the early diagnosis of skin lesions, where patients and doctors can be benefited from it. They can take the form of smartphones applications, websites, and stations in hospitals. Therefore, this solution may help many lives early diagnosing malignant lesions, helping in decision making, reducing costs of diagnostic and reducing money spent in treatment.

However, it is not a simple task to apply machine learning techniques in the medical field. The data scarcity problem poses a major obstacle towards good and reliable models, especially when the training models are deep neural networks. Moreover, the problem domain and dataset encountered in this field is not something to take lightly, particularly when talking about clinical images.

The method presented in this work was a deep convolutional neural network that used clinical images of skin lesions to distinguish 12 different conditions based on the image. This proposal was based on state-of-the-art works (ESTEVA et al., 2017; Seog Han et al., 2018), that leverages on DNNs trained with techniques such as transfer learning and data augmentation. Therefore, it was shown a final model trained with 88,090 images of 12 different skin lesions, that achieved results comparable to the state-of-the-art.

Furthermore, we presented useful explanations using a gradient-based method called GradCAM. The visual explanations generated were capable of showing the good generalization of the model, as well as, some biases that the model learned from outlier images. Moreover, these insights empower researchers and field experts to have a look inside in the inner workings of the black-box model and turn predictions in counseling.

Finally, we conclude that this work brought good results to the research and practice communities. Furthermore, this work may serve as the stepping stone to build an application that may help innumerous patients and unaware people to fight skin cancer, and thus save and improve many lives.

5.1 Future Work

For the continuity of this work and investigations described here, it is necessary to collect more images to build a more robust model. A model that deals with other diseases, more importantly, adversarial examples such as healthy skin, fingers, hair, nose, eyes, background objects, *etc.* It is expected that this addition will make the model generalize well what are the features that really constitute a given lesion, ignoring adjacent features. Only then the model will be prepared to be deployed in the wild (real-world) and go to as many hands as possible.

Therefore, another task that would bring the adoption of this model, is gathering written reports of lesion observations, both in technical and non-technical languages. With these in hand, it will be possible to develop a model to generate captions for images, serving as an explanation that may convey a clear message of what is on the image and why that is important for the decision taken.

Bibliography

ABDEL-HAMID, O. et al. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, IEEE, v. 22, n. 10, p. 1533–1545, 2014. Cited on page 37.

AERTS, H. J. et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*, Nature Publishing Group, v. 5, p. 4006, 2014. Cited on page 21.

American Cancer Society. *About Basal and Squamous Cell Skin Cancer*. 2016. Available at: <https://www.cancer.org/content/dam/CRC/PDF/Public/8818.00.pdf>. Accessed in 29/04/2018. Cited 4 times on pages 20, 44, 46, and 49.

American Cancer Society. *Cancer facts & figures 2017*. [S.l.]: Atlanta, American Cancer Society, 2017. Available at: <https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2017.html>. Accessed in 29/04/2018. Cited 3 times on pages 18, 20, and 44.

American Cancer Society. *Cancer facts & figures 2018*. [S.l.]: Atlanta, American Cancer Society, 2018. Available at: <https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2018.html>. Accessed in 19/05/2018. Cited 3 times on pages 15, 47, and 49.

ANDERSON, R. The rashomon effect and communication. *Canadian Journal of Communication*, v. 41, n. 2, 2016. Cited on page 65.

ANGWIN, J. et al. Machine bias risk assessments in criminal sentencing. *ProPublica https://www. propublica. org*, 2016. Cited on page 60.

ANTUNES, P. et al. Structuring dimensions for collaborative systems evaluation. *ACM Computing Surveys (CSUR)*, ACM, v. 44, n. 2, p. 8, 2012. Cited on page 66.

Association of American Medical Colleges (AAMC). *Specialties in medicine*. [S.l.]: Carrers in Medicine, 2018. Available at: <https://www.aamc.org/cim/specialty>. Accessed in 22/05/2018. Cited on page 17.

BALLERINI, L. et al. A color and texture based hierarchical k-nn approach to the classification of non-melanoma skin lesions. In: *Color Medical Image Analysis*. [S.l.]: Springer, 2013. p. 63–86. Cited on page 22.

BENGIO, Y.; BOULANGER-LEWANDOWSKI, N.; PASCANU, R. Advances in optimizing recurrent networks. *CoRR*, abs/1212.0901, 2012. Available at: <http://arxiv.org/abs/1212.0901>. Cited on page 35.

BICKERS, D. R. et al. The burden of skin diseases: 2004: A joint project of the american academy of dermatology association and the society for investigative dermatology. *Journal of the American Academy of Dermatology*, Elsevier, v. 55, n. 3, p. 490–500, 2006. Cited on page 17.

BIRAN, O.; COTTON, C. Explanation and justification in machine learning: A survey. In: *IJCAI-17 Workshop on Explainable AI (XAI)*. [S.l.: s.n.], 2017. p. 8. Cited on page 58.

BLOICE, M. D.; STOCKER, C.; HOLZINGER, A. Augmentor: An image augmentation library for machine learning. *CoRR*, abs/1708.04680, 2017. Available at: <http://arxiv.org/abs/1708.04680>. Cited on page 55.

BÖER-AUER, A.; JONES, M.; LYASNICHAYA, O. V. Cytokeratin 10-negative nested pattern enables sure distinction of clonal seborrheic keratosis from pagetoid bowen's disease. *Journal of cutaneous pathology*, Wiley Online Library, v. 39, n. 2, p. 225–233, 2012. Cited on page 47.

BOWEN, J. T. Precancerous dermatoses: a study of two cases of chronic atypical epithelial proliferation. *Archives of dermatology*, American Medical Association, v. 119, n. 3, p. 243–260, 1983. Cited on page 46.

BRADSHAW, J. M. et al. Toward trustworthy adjustable autonomy in kaos. In: *Trusting Agents for Trusting Electronic Societies*. [S.l.]: Springer, 2005. p. 18–42. Cited on page 62.

BRASIL. *Lei no 13.709, de 14 de Agosto de 2018*. 2018. Accessed in 10/10/2018. Available at: <http://www.planalto.gov.br/CCIVil_03/_Ato2015-2018/2018/Lei/L13709.htm>. Cited on page 61.

BRASIL; Ministério da Saúde. *Sistema de informações sobre mortalidade*. [S.l.]: Brasília, Departamento de Informática do SUS, 2017. Available at: <http://www.datasus.gov.br>. Cited on page 46.

BROMBERGER, S. Why-questions. *R. G. Colodny (ed.), Mind and cosmos: Essays in contemporary science and philosophy*, University of Pittsburgh Press Pittsburgh, p. 68–111, 1966. Cited on page 63.

CANZIANI, A.; PASZKE, A.; CULURCIELLO, E. An analysis of deep neural network models for practical applications. *CoRR*, abs/1605.07678, 2016. Available at: <http://arxiv.org/abs/1605.07678>. Cited on page 38.

CARUANA, R. et al. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In: ACM. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [S.l.], 2015. p. 1721–1730. Cited on page 57.

CELEBI, M. E.; STOECKER, W. V.; MOSS, R. H. Advances in skin cancer image analysis. *Computerized Medical Imaging and Graphics*, v. 2, n. 35, p. 83–84, 2011. Cited on page 19.

CHANDRASEKARAN, B.; TANNER, M. C.; JOSEPHSON, J. R. Explaining control strategies in problem solving. *IEEE Intelligent Systems*, IEEE, n. 1, p. 9–15, 1989. Cited on page 58.

CHAUDHURI, S. et al. Detection of blood vessels in retinal images using two-dimensional matched filters. *IEEE Transactions on medical imaging*, IEEE, v. 8, n. 3, p. 263–269, 1989. Cited on page 21.

CHU, H. Mdb: A memory-mapped database and backend for openldap. In: *Proceedings of the 3rd International Conference on LDAP, Heidelberg, Germany*. [S.l.: s.n.], 2011. p. 35. Cited on page 56.

CIREŞAN, D. et al. Multi-column deep neural network for traffic sign classification. *Neural networks*, Elsevier, v. 32, p. 333–338, 2012. Cited on page 37.

CLANCEY, W. J. The epistemology of a rule-based expert system—a framework for explanation. *Artificial intelligence*, Elsevier, v. 20, n. 3, p. 215–251, 1983. Cited on page 58.

CLANCEY, W. J.; SHORTLIFFE, E. H. *Readings in medical artificial intelligence: the first decade*. [S.l.]: Addison-Wesley Longman Publishing Co., Inc., 1984. 382-398 p. Cited on page 58.

CUA, A.; WILHELM, K.-P.; MAIBACH, H. Elastic properties of human skin: relation to age, sex, and anatomical region. *Archives of Dermatological Research*, Springer, v. 282, n. 5, p. 283–288, 1990. Cited on page 52.

CUBUK, E. D. et al. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018. Cited on page 55.

DABKOWSKI, P.; GAL, Y. Real time image saliency for black box classifiers. In: GUYON, I. et al. (Ed.). *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017. p. 6967–6976. Available at: <http://papers.nips.cc/paper/7272-real-time-image-saliency-for-black-box-classifiers.pdf>. Cited on page 71.

DENGEL, L. T. et al. Total body photography for skin cancer screening. *International journal of dermatology*, Wiley Online Library, v. 54, n. 11, p. 1250–1254, 2015. Cited on page 52.

DENNETT, D. C. *The intentional stance*. [S.l.]: MIT press, 1989. Cited on page 58.

DERANCOURT, C. et al. Multiple large solar lentigos on the upper back as clinical markers of past severe sunburn: a case-control study. *Dermatology*, Karger Publishers, v. 214, n. 1, p. 25–31, 2007. Cited on page 47.

DONNÉ, A. *Cours de microscopie complémentaire des études médicales: Atlas executé d'aprés nature au microscope-daguerreotype*. [S.l.]: J.-B. Bailliére, 1845. Cited on page 17.

DOSHI-VELEZ, F.; KIM, B. *Towards a rigorous science of interpretable machine learning.(2017)*. 2017. Available at: <https://arxiv.org/abs/1702.08608>. Cited 5 times on pages 59, 60, 62, 66, and 67.

DOSHI-VELEZ, F. et al. Accountability of AI under the law: The role of explanation. *CoRR*, abs/1711.01134, 2017. Available at: <http://arxiv.org/abs/1711.01134>. Cited on page 61.

DUCHI, J.; HAZAN, E.; SINGER, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, v. 12, n. Jul, p. 2121–2159, 2011. Cited on page 35.

DYK, D. A. V.; MENG, X.-L. The art of data augmentation. *Journal of Computational and Graphical Statistics*, Taylor & Francis, v. 10, n. 1, p. 1–50, 2001. Cited on page 25.

ERHAN, D. et al. Visualizing higher-layer features of a deep network. *University of Montreal*, v. 1341, n. 3, p. 1, 2009. Cited on page 71.

ESTEVA, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 2017. ISSN 14764687. Cited 9 times on pages 21, 24, 25, 26, 38, 39, 78, 80, and 87.

EVANS, J.; CARMAN, C. C.; THORNDIKE, A. S. Solar anomaly and planetary displays in the antikythera mechanism. *Journal for the History of Astronomy*, SAGE Publications Sage UK: London, England, v. 41, n. 1, p. 1–39, 2010. Cited on page 22.

EVANS, J. et al. Dermatofibromas and arthropod bites: is there any evidence to link the two? *The Lancet*, Elsevier, v. 334, n. 8653, p. 36–37, 1989. Cited on page 46.

FERLAY, J. et al. *GLOBOCAN 2012 v1. 0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11. Lyon, France: International Agency for Research on Cancer*. 2012. Available at: <http://globocan.iarc.fr>. Accessed in 30/04/2018. Cited on page 20.

FONG, R. C.; VEDALDI, A. Interpretable explanations of black boxes by meaningful perturbation. *arXiv preprint arXiv:1704.03296*, 2017. Cited 2 times on pages 69 and 70.

FREITAS, A. A. Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter*, ACM, v. 15, n. 1, p. 1–10, 2014. Cited on page 67.

GALE, W. et al. Producing radiologist-quality reports for interpretable artificial intelligence. *arXiv preprint arXiv:1806.00340*, 2018. Cited 3 times on pages 65, 73, and 74.

GILMORE, S.; HOFMANN-WELLENHOF, R.; SOYER, H. P. A support vector machine for decision support in melanoma recognition. *Experimental dermatology*, Wiley Online Library, v. 19, n. 9, p. 830–835, 2010. Cited on page 22.

GIOTIS, I. et al. Med-node: a computer-assisted melanoma diagnosis system using non-dermoscopic images. *Expert systems with applications*, Elsevier, v. 42, n. 19, p. 6578–6585, 2015. Cited on page 41.

GOLDBAUM, M. et al. Automated diagnosis and image understanding with object extraction, object classification, and inferencing in retinal images. In: IEEE. *Image Processing, 1996. Proceedings., International Conference on*. [S.l.], 1996. v. 3, p. 695–698. Cited on page 21.

GOODFELLOW, I. et al. *Deep learning*. [S.l.]: MIT press Cambridge, 2016. Cited 3 times on pages 22, 29, and 34.

GOODMAN, B.; FLAXMAN, S. European union regulations on algorithmic decision-making and a" right to explanation". *arXiv preprint arXiv:1606.08813*, 2016. Cited on page 61.

GOOROCHURN, R. et al. Differential morphological and functional features of fibroblasts explanted from solar lentigo. *British Journal of Dermatology*, Wiley Online Library, v. 177, n. 4, 2017. Cited on page 47.

GRICE, H. P. Logic and conversation. *Speech Arts*, New York: Academic Press, p. 41–58, 1975. Cited on page 73.

GUPTA, S. et al. Bowen disease over photoprotected site in an indian male. *Dermatology online journal*, v. 15, n. 10, 2009. Cited on page 46.

HAFNER, C.; VOGT, T. Seborrheic keratosis. *JDDG: Journal der Deutschen Dermatologischen Gesellschaft*, Wiley Online Library, v. 6, n. 8, p. 664–677, 2008. Cited on page 49.

HAJDU, S. I. A note from history: landmarks in history of cancer, part 1. *Cancer*, Wiley Online Library, v. 117, n. 5, p. 1097–1102, 2011. Cited on page 20.

HAJIAN-TILAKI, K. Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine*, Babol University of Medical Sciences, v. 4, n. 2, p. 627, 2013. Cited on page 37.

HAMBLIN, M. R.; AVCI, P.; GUPTA, G. K. *Imaging in Dermatology*. [S.l.]: Academic Press, 2016. Cited 2 times on pages 15 and 18.

HANKINSON, R. J. *Cause and explanation in ancient Greek thought*. [S.l.]: Oxford University Press, 2001. Cited on page 63.

HAY, R. J. et al. The global burden of skin disease in 2010: an analysis of the prevalence and impact of skin conditions. *Journal of Investigative Dermatology*, Elsevier, v. 134, n. 6, p. 1527–1534, 2014. Cited on page 17.

HAYKIN, S. *Neural Networks: A comprehensive foundation*. 2nd. ed. [S.l.]: Pearson Education, Prentice Hall, 1999. Cited 3 times on pages 15, 28, and 33.

HE, K. et al. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. Available at: <http://arxiv.org/abs/1512.03385>. Cited on page 40.

HILTON, D. J. Conversational processes and causal explanation. *Psychological Bulletin*, American Psychological Association, v. 107, n. 1, p. 65, 1990. Cited 2 times on pages 65 and 73.

HOFFMAN, R. R. et al. Trust in automation. *IEEE Intelligent Systems*, IEEE, v. 28, n. 1, p. 84–88, 2013. Cited on page 62.

HOFFMAN, R. R. et al. The dynamics of trust in cyberdomains. *IEEE Intelligent Systems*, IEEE, n. 6, p. 5–11, 2009. Cited on page 63.

HOLZINGER, A. et al. What do we need to build explainable AI systems for the medical domain? *CoRR*, abs/1712.09923, 2017. Available at: <http://arxiv.org/abs/1712.09923>. Cited on page 60.

HOWELL, J. Y.; RAMSEY, M. L. Cancer, squamous cell, skin. StatPearls Publishing, 2017. Accessed in 19/05/2018. Cited on page 49.

HUBEL, D. H.; WIESEL, T. N. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, Wiley Online Library, v. 195, n. 1, p. 215–243, 1968. Cited on page 28.

Instituto Nacional de Câncer José Alencar Gomes da Silva. *Estimativa 2018 – Incidência de Câncer no Brasil*. 2018. Available at: <http://www1.inca.gov.br/inca/Arquivos/estimativa-2018.pdf>. Accessed in 29/04/2018. Cited 2 times on pages 20 and 47.

International Society for Digital Imaging of the Skin. *International Skin Imaging Collaboration Project*. 2018. Available at: <https://isic-archive.com>. Accessed in 20/03/2018. Cited on page 43.

IOFFE, S.; SZEGEDY, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015. Available at: <http://arxiv.org/abs/1502.03167>. Cited on page 41.

JIA, Y. et al. *Caffe: Convolutional Architecture for Fast Feature Embedding*. 2014. Available at: <https://raw.githubusercontent.com/BVLC/caffe/master/examples/images/cat.jpg>. Accessed in 27/08/2018. Cited on page 28.

JIA, Y. et al. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. Cited on page 76.

JOHANSSON, U. et al. Trade-off between accuracy and interpretability for predictive in silico modeling. *Future medicinal chemistry*, Future Science, v. 3, n. 6, p. 647–663, 2011. Cited on page 68.

KAHNEMAN, D.; TVERSKY, A. *The simulation heuristic*. [S.l.], 1981. Cited on page 65.

KARPATHY, A. *CS231n Convolutional Neural Networks for Visual Recognition*. 2018. Available at: <http://cs231n.github.io>. Accessed in 20/05/2018. Cited 4 times on pages 30, 32, 33, and 35.

KARPATHY, A. *Neural Networks Part 1: Setting up the Architecture*. 2018. Available at: <http://cs231n.github.io/neural-networks-1>. Accessed in 01/06/2018. Cited on page 32.

KARPATHY, A.; FEI-FEI, L. Deep visual-semantic alignments for generating image descriptions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2015. p. 3128–3137. Cited 2 times on pages 73 and 74.

KAVAK, A. et al. Preliminary study among truck drivers in turkey: effects of ultraviolet light on some skin entities. *The Journal of dermatology*, Wiley Online Library, v. 35, n. 3, p. 146–150, 2008. Cited on page 47.

KERMANY, D. S. et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, 2018. ISSN 10974172. Cited on page 21.

KIM, B. *Interactive and interpretable machine learning models for human machine collaboration*. Phd Thesis (PhD Thesis) — Massachusetts Institute of Technology, 2015. Cited on page 62.

KOOPMAN, P.; HOFFMAN, R. R. Work-arounds, make-work, and kludges. *IEEE Intelligent Systems*, IEEE, v. 18, n. 6, p. 70–75, 2003. Cited on page 63.

KORNBLITH, S.; SHLENS, J.; LE, Q. V. Do better imagenet models transfer better? *arXiv preprint arXiv:1805.08974*, 2018. Cited on page 54.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2012. p. 1097–1105. Cited 4 times on pages 15, 25, 32, and 55.

KUMAR, R.; INDRAYAN, A. Receiver operating characteristic (roc) curve for medical researchers. *Indian pediatrics*, Springer, v. 48, n. 4, p. 277–287, 2011. Cited on page 37.

KWASIGROCH, A.; MIKOŁAJCZYK, A.; GROCHOWSKI, M. Deep neural networks approach to skin lesions classification – a comparative analysis. 2017. Cited on page 23.

LARSSON, S.; BRANE, J. Medical photography. In: *The Focal Encyclopedia of Photography (Fourth Edition)*. [S.l.]: Elsevier, 2007. p. 569–572. Cited on page 17.

LECUN, Y. et al. Generalization and network design strategies. *Connectionism in perspective*, Citeseer, p. 143–155, 1989. Cited on page 28.

LEE, H. et al. Fully automated deep learning system for bone age assessment. *Journal of digital imaging*, Springer, v. 30, n. 4, p. 427–441, 2017. Cited on page 21.

LIAO, Q.; POGGIO, T. A. Bridging the gaps between residual learning, recurrent neural networks and visual cortex. *CoRR*, abs/1604.03640, 2016. Available at: <http://arxiv.org/abs/1604.03640>. Cited on page 41.

LIN, M.; CHEN, Q.; YAN, S. Network in network. *CoRR*, abs/1312.4400, 2013. Available at: <http://arxiv.org/abs/1312.4400>. Cited 2 times on pages 38 and 39.

LIPTON, Z. C. The mythos of model interpretability. *CoRR*, abs/1606.03490, 2016. Available at: <http://arxiv.org/abs/1606.03490>. Cited 4 times on pages 58, 62, 64, and 68.

LOBATO-BEREZO, A. et al. Dermatofibroma arising within a black tattoo. *Case reports in dermatological medicine*, Hindawi, v. 2014, 2014. Cited on page 46.

LOMBROZO, T. The structure and function of explanations. *Trends in cognitive sciences*, Elsevier, v. 10, n. 10, p. 464–470, 2006. Cited on page 61.

LOMBROZO, T. Simplicity and probability in causal explanation. *Cognitive psychology*, Elsevier, v. 55, n. 3, p. 232–257, 2007. Cited on page 66.

LOO, S. K.-f.; TANG, W. Y.-m. Warts (non-genital). BJM clinical evidence, 2014. Accessed in 20/05/2018. Cited on page 50.

LOVEJOY, J. The ux of ai. Google Design, 2018. Accessed in 30/11/2018. Available at: <https://design.google/library/ux-ai/>. Cited 2 times on pages 65 and 73.

MASTERS, D.; LUSCHI, C. Revisiting small batch training for deep neural networks. *CoRR*, abs/1804.07612, 2018. Available at: <http://arxiv.org/abs/1804.07612>. Cited on page 78.

MATSUNAGA, K. et al. Image Classification of Melanoma, Nevus and Seborrheic Keratosis by Deep Neural Network Ensemble. 2017. Cited 2 times on pages 15 and 22.

MAYER, R. C.; DAVIS, J. H.; SCHOORMAN, F. D. An integrative model of organizational trust. *Academy of management review*, Academy of Management Briarcliff Manor, NY 10510, v. 20, n. 3, p. 709–734, 1995. Cited on page 62.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, v. 5, n. 4, p. 115–133, 1943. ISSN 1522-9602. Available at: <http://dx.doi.org/10.1007/BF02478259>. Cited 2 times on pages 15 and 27.

MENEGOLA, A. et al. Knowledge transfer for melanoma screening with deep learning. In: *Proceedings - International Symposium on Biomedical Imaging*. [S.l.: s.n.], 2017. ISBN 9781509011711. ISSN 19458452. Cited on page 23.

MESNIL, G. et al. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, IEEE, v. 23, n. 3, p. 530–539, 2015. Cited on page 37.

MIKHAIL, G. R.; MEHREGAN, A. H. Basal cell carcinoma in seborrheic keratosis. *Journal of the American Academy of Dermatology*, Elsevier, v. 6, n. 4, p. 500–506, 1982. Cited on page 49.

MILLER, T. Explanation in artificial intelligence: Insights from the social sciences. *CoRR*, abs/1706.07269, 2017. Available at: <http://arxiv.org/abs/1706.07269>. Cited 5 times on pages 58, 59, 63, 64, and 65.

MILLER, T.; HOWE, P.; SONENBERG, L. Explainable ai: Beware of inmates running the asylum. In: *IJCAI-17 Workshop on Explainable AI (XAI)*. [S.l.: s.n.], 2017. v. 36. Cited on page 64.

MILLS, S. E.; COOPER, P. H.; FECHNER, R. E. Lobular capillary hemangioma: the underlying lesion of pyogenic granuloma. a study of 73 cases from the oral and nasal mucous membranes. *The American journal of surgical pathology*, v. 4, n. 5, p. 470–479, 1980. Cited on page 48.

MOLNAR, C. *Interpretable machine learning. a guide for making black box models explainable*. 2018. Available at: <https://christophm.github.io/interpretable-ml-book/>. Accessed in 28/06/2018. Cited on page 69.

MORTON, C.; BIRNIE, A.; EEDY, D. British association of dermatologists' guidelines for the management of squamous cell carcinoma in situ (bowen's disease) 2014. *British Journal of Dermatology*, Wiley Online Library, v. 170, n. 2, p. 245–260, 2014. Cited on page 46.

MOY, R. L. Clinical presentation of actinic keratoses and squamous cell carcinoma. *Journal of the American Academy of Dermatology*, Elsevier, v. 42, n. 1, p. S8–S10, 2000. Cited 2 times on pages 44 and 49.

MUIR, B. M.; MORAY, N. Trust in automation. part ii. experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, Taylor & Francis, v. 39, n. 3, p. 429–460, 1996. Cited on page 62.

MURPHY, K. P. *Machine Learning: A Probabilistic Perspective. Adaptive Computation and Machine Learning*. [S.l.]: MIT press, 2012. Cited 3 times on pages 15, 21, and 68.

NACHBAR, F. et al. The abcd rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions. *Journal of the American Academy of Dermatology*, Elsevier, v. 30, n. 4, p. 551–559, 1994. Cited on page 19.

NASR-ESFAHANI, E. et al. Melanoma detection by analysis of clinical images using convolutional neural network. In: *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. [S.l.: s.n.], 2016. ISBN 978-1-4577-0220-4. ISSN 1557-170X. Cited on page 23.

National Cancer Institute. *Cancer Stat Facts: Melanoma of the Skin*. 2018. Available at: <https://seer.cancer.gov/statfacts/html/melan.html>. Accessed in 19/05/2018. Cited on page 48.

National Cancer Institute. *Common Moles, Displastic Nevi, and Risk of Melanoma*. 2018. Available at: <https://www.cancer.gov/types/skin/moles-fact-sheet>. Accessed in 19/05/2018. Cited on page 48.

National Cancer Institute. *Melanoma Treatment (PDQ®)*. 2018. Available at: <https://www.cancer.gov/types/skin/hp/melanoma-treatment-pdq>. Accessed in 19/05/2018. Cited 2 times on pages 47 and 48.

NAVERSEN, D. N. et al. Painful tumors of the skin:"lend an egg". *Journal of the American Academy of Dermatology*, Elsevier, v. 28, n. 2, p. 298–300, 1993. Cited on page 46.

NEUSE, W. H. et al. The history of photography in dermatology: Milestones from the roots to the 20th century. *Archives of dermatology*, American Medical Association, v. 132, n. 12, p. 1492–1498, 1996. Cited on page 17.

NG, A. *Artificial Intelligence is the New Electricity*. 2017. Available at: <https://www.youtube.com/watch?v=21EiKfQYZXc>. Stanford MSx Future Forum. Accessed in 20/05/2018. Cited on page 21.

NICKERSON, R. S. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, Educational Publishing Foundation, v. 2, n. 2, p. 175, 1998. Cited on page 65.

NORTHERNCEDAR. *A battery powered double polarized dry dermatoscope made by 3GEN called the DERMLITE*. 2009. Available at: <https://commons.wikimedia.org/wiki/File:Dermatoscope.jpg>. Accessed in 27/08/2018. Cited on page 19.

ODOM, R.; JAMES, W.; BERGER, T. Dermal and subcutaneous tumors: cherry angiomas. *Andrew's Diseases of the Skin: Clinical Dermatology*, 2000. Cited on page 46.

OLAH, C. et al. The building blocks of interpretability. *Distill*, 2018. Https://distill.pub/2018/building-blocks. Cited on page 65.

ORTH, G.; FAVRE, M.; CROISSANT, O. Characterization of a new type of human papillomavirus that causes skin warts. *Journal of virology*, Am Soc Microbiol, v. 24, n. 1, p. 108–120, 1977. Cited on page 49.

OVERTON, J. Scientific explanation and computation. In: *ExaCt*. [S.l.: s.n.], 2011. p. 41–50. Cited 2 times on pages 58 and 63.

OVERTON, J. A. *Explanation in Science*. Phd Thesis (PhD Thesis) — Citeseer, 2012. Cited 2 times on pages 63 and 64.

PAIR; Google AI. *What-if Tool*. Google, 2018. Accessed in 30/11/2018. Available at: <https://pair-code.github.io/what-if-tool/index.html>. Cited on page 65.

PAN, S. J.; YANG, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, IEEE, v. 22, n. 10, p. 1345–1359, 2010. Cited on page 25.

PATTERSON, J. W. *Weedon's Skin Pathology E-Book*. [S.l.]: Elsevier Health Sciences, 2014. Cited on page 44.

PEREZ, L.; WANG, J. The effectiveness of data augmentation in image classification using deep learning. *CoRR*, abs/1712.04621, 2017. Available at: <http://arxiv.org/abs/1712.04621>. Cited on page 55.

PRAJAPATI, V.; BARANKIN, B. Answer: Can you identify this condition? *Canadian Family Physician*, The College of Family Physicians of Canada, v. 54, n. 5, p. 699–699, 2008. ISSN 0008-350X. Available at: <http://www.cfp.ca/content/54/5/699>. Cited on page 44.

PRIME, J. *Des accidents toxiques produits par l'éosinate de sodium*. Phd Thesis (PhD Thesis), 1900. Cited on page 18.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016. Cited 3 times on pages 68, 70, and 71.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016. Available at: <http://arxiv.org/abs/1602.04938>. Cited 2 times on pages 60 and 62.

RIDGEWAY, G. et al. Interpretable boosted naïve bayes classification. In: *KDD*. [S.l.: s.n.], 1998. p. 101–104. Cited on page 62.

ROBNIK-ŠIKONJA, M.; BOHANEC, M. Perturbation-based explanations of prediction models. In: *Human and Machine Learning*. [S.l.]: Springer, 2018. p. 159–175. Cited on page 69.

ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, American Psychological Association, v. 65, n. 6, p. 386, 1958. Cited on page 26.

RUSSAKOVSKY, O. et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, v. 115, n. 3, p. 211–252, 2015. Cited 2 times on pages 38 and 39.

SANTURKAR, S. et al. How does batch normalization help optimization?(no, it is not about internal covariate shift). *arXiv preprint arXiv:1805.11604*, 2018. Cited on page 41.

SELVARAJU, R. R. et al. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016. Available at: <http://arxiv.org/abs/1610.02391>. Cited 3 times on pages 71, 72, and 73.

Seog Han, S. et al. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *The Journal of Investigative Dermatology*, 2018. Cited 12 times on pages 24, 25, 26, 38, 41, 42, 50, 52, 77, 78, 80, and 87.

SIMARD, P. Y. et al. Best practices for convolutional neural networks applied to visual document analysis. In: *ICDAR*. [S.l.: s.n.], 2003. v. 3, p. 958–962. Cited on page 25.

SIRAISI, N. G. History, antiquarianism, and medicine: The case of girolamo mercuriale. *Journal of the History of Ideas*, University of Pennsylvania Press, v. 64, n. 2, p. 231–251, 2003. Cited on page 17.

SMILKOV, D. et al. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. Cited on page 71.

SMITH, L. N. A disciplined approach to neural network hyper-parameters: Part 1 - learning rate, batch size, momentum, and weight decay. *CoRR*, abs/1803.09820, 2018. Available at: <http://arxiv.org/abs/1803.09820>. Cited 2 times on pages 78 and 79.

SOUZA, R. J. S. A. P. de et al. Estimativa do custo do tratamento do câncer de pele tipo não-melanoma no estado de são paulo - Brasil. *Anais Brasileiros de Dermatologia*, 2011. ISSN 03650596. Cited 2 times on pages 15 and 20.

SRIVASTAVA, R. K.; GREFF, K.; SCHMIDHUBER, J. Highway networks. *CoRR*, abs/1505.00387, 2015. Available at: <http://arxiv.org/abs/1505.00387>. Cited on page 40.

STEWART, B.; WILD, C. P. et al. World cancer report 2014. *Health*, 2014. Cited on page 20.

ŠTRUMBELJ, E.; KONONENKO, I. A general method for visualizing and explaining black-box regression models. In: SPRINGER. *International Conference on Adaptive and Natural Computing Algorithms*. [S.l.], 2011. p. 21–30. Cited on page 65.

SUNDARARAJAN, M.; TALY, A.; YAN, Q. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*, 2017. Cited on page 71.

SWETS, J. A. Indices of discrimination or diagnostic accuracy: their rocs and implied models. *Psychological bulletin*, American Psychological Association, v. 99, n. 1, p. 100, 1986. Cited on page 37.

SZEGEDY, C.; IOFFE, S.; VANHOUCKE, V. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261, 2016. Available at: <http://arxiv.org/abs/1602.07261>. Cited on page 39.

SZEGEDY, C. et al. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014. Available at: <http://arxiv.org/abs/1409.4842>. Cited on page 39.

SZEGEDY, C. et al. Rethinking the Inception Architecture for Computer Vision. 2015. ISSN 08866236. Cited on page 39.

TEACH, R. L.; SHORTLIFFE, E. H. An analysis of physician attitudes regarding computer-based clinical consultation systems. In: *Use and impact of computers in clinical medicine*. [S.l.]: Springer, 1981. p. 68–85. Cited on page 57.

The Economist; Intelligence Unit. *Saudi Arabia plans for an AI future*. 2018. Accessed in 29/11/2018. Available at: <http://country.eiu.com/article.aspx?articleid=1786940762&Country=Saudi%20Arabia&topic=Economy>. Cited on page 61.

TPCN, P. Development of lentigines in german and japanese women correlates with variants in the slc45a2 gene. *Journal of Investigative Dermatology*, 2011. Cited on page 47.

TSYMBAL, A. The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin*, Citeseer, v. 106, n. 2, 2004. Cited on page 60.

UNAL, I. Defining an optimal cut-point value in roc analysis: An alternative approach. *Computational and mathematical methods in medicine*, Hindawi, v. 2017, 2017. Cited on page 37.

VANDENBERGHE, M. E. et al. Relevance of deep learning to facilitate the diagnosis of her2 status in breast cancer. *Scientific reports*, Nature Publishing Group, v. 7, p. 45938, 2017. Cited on page 21.

VEIT, A.; WILBER, M. J.; BELONGIE, S. J. Residual networks are exponential ensembles of relatively shallow networks. *CoRR*, abs/1605.06431, 2016. Available at: <http://arxiv.org/abs/1605.06431>. Cited on page 41.

VINYALS, O. et al. Show and tell: A neural image caption generator. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2015. p. 3156–3164. Cited 2 times on pages 73 and 74.

VISSER, E. J. de et al. The world is not enough: Trust in cognitive agents. In: SAGE PUBLICATIONS SAGE CA: LOS ANGELES, CA. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. [S.l.], 2012. v. 56, n. 1, p. 263–267. Cited on page 62.

WAJAPEYEE, N. et al. Natural secretion marks difference between mole and melanoma. EurekAlert!, 2008. Accessed in 30/11/2018. Available at: <https://www.eurekalert.org/pub_releases/2008-02/cp-nsm020408.php>. Cited on page 65.

WANG, W. et al. Towards cooperation in sequential prisoner's dilemmas: a deep multiagent reinforcement learning approach. *CoRR*, abs/1803.00162, 2018. Available at: <http://arxiv.org/abs/1803.00162>. Cited on page 37.

WANG, X. et al. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2018. p. 9049–9058. Cited 2 times on pages 73 and 74.

WERBOS, P. J. Applications of advances in nonlinear sensitivity analysis. In: *System modeling and optimization*. [S.l.]: Springer, 1982. p. 762–770. Cited on page 27.

WILLIAMS, J. J.; LOMBROZO, T.; REHDER, B. The hazards of explanation: Overgeneralization in the face of exceptions. *Journal of Experimental Psychology: General*, American Psychological Association, v. 142, n. 4, p. 1006, 2013. Cited on page 61.

YOSINSKI, J. et al. How transferable are features in deep neural networks? *CoRR*, abs/1411.1792, 2014. Available at: <http://arxiv.org/abs/1411.1792>. Cited 2 times on pages 25 and 55.

ZHANG, X.; LECUN, Y. Text understanding from scratch. *CoRR*, abs/1502.01710, 2015. Available at: <http://arxiv.org/abs/1502.01710>. Cited on page 28.

ZHOU, B. et al. Learning deep features for discriminative localization. *CoRR*, abs/1512.04150, 2015. Available at: <http://arxiv.org/abs/1512.04150>. Cited on page 72.

Appendix

APPENDIX A – Data Augmentation

Source Code A.1 – Data augmentation script using Augmentor library.

```
# coding: utf-8
# Originally a jupyter notebook
# Dataset Augmentation Script

import Augmentor as aug # Library that handles the augmentation methods
import glob
import os
import numpy as np
import cv2 # OpenCV library
import PIL
from Augmentor.Operations import Operation

'''
Light Transformation Implementation

This transformation tries to add more light variance to the images in the dataset.
For that, a low and upper bound of intensity is set so that the image is not too bright nor too dark.
On top of these thresholds, a random number is picked to then apply the lightning transformation.
'''
class Lightning(Operation):
    def __init__(self, probability, intensity_low=0.7, intensity_high=1.2):
        Operation.__init__(self, probability)
        self.intensity_low = intensity_low
        self.intensity_high = intensity_high

    def perform_operation(self, images):
        for i, image in enumerate(images):
            image = np.array(image.convert('RGB'))
            row, col, _ = image.shape
```

```
30              light_intensity = np.random.randint(
31                  int(self.intensity_low * 100),
32                  int(self.intensity_high * 100)
33              )
34              light_intensity /= 100
35
36              gaussian = 100 * np.random.random((row, col, 1))
37              gaussian = np.array(gaussian, dtype=np.uint8)
38              gaussian = np.concatenate((gaussian, gaussian, gaussian),
                ↪  axis=2)
39              image = cv2.addWeighted(image, light_intensity, gaussian,
                ↪  0.25, 0)
40
41              image = PIL.Image.fromarray(image)
42              images[i] = image
43          return images
44
45
46  # Multiplier used to set the final augmented images number
47  MULTIPLIER = 29
48  directory = '../../../data/edited/Med_atlas_edin/train/*'
49
50  folders = []
51  for f in glob.glob(directory):
52      if os.path.isdir(f):
53          folders.append(os.path.abspath(f))
54
55  print('Classes found {}'.format([os.path.split(x)[1] for x in folders]))
56
57
58  pipelines = {}
59  # Create the augmentation pipeline for each lesion found
60  for folder in folders:
61      pipelines[os.path.split(folder)[1]] = \
62              (aug.Pipeline(source_directory=folder,
63                            output_directory='resnet_augmented',
64                            save_format='jpg'
65                           )
66                  )
```

```
67
68 classes_count = []
69 for p in pipelines.values():
70     print("Class '{}' has {} samples".format(
71         p.augmentor_images[0].class_label,
72         len(p.augmentor_images)
73     ))
74     classes_count.append(len(p.augmentor_images))
75
76
77 lightning = Lightning(probability=0.5)
78
79 for p in pipelines.values():
80     p.rotate(probability=0.5, max_left_rotation=10,
       ↪  max_right_rotation=10)
81     p.zoom_random(probability=0.4, percentage_area=0.9)
82     p.flip_left_right(probability=0.7)
83     p.flip_top_bottom(probability=0.5)
84     p.random_distortion(probability=0.8, grid_width=5, grid_height=5,
       ↪  magnitude=15)
85     p.add_operation(lightning)
86     p.resize(probability=1.0, width=224, height=224)
87
88 # If a equal sampling of the lesions is needed
89 # Mind that the final MULTIPLIER can scale many times if True
90 SAME_SAMPLING = False
91 for p in pipelines.values():
92     if SAME_SAMPLING:
93         diff = max(classes_count) - len(p.augmentor_images)
94         p.sample((len(p.augmentor_images) + diff)*MULTIPLIER + diff)
95     else:
96         p.sample(len(p.augmentor_images)*MULTIPLIER)
```

APPENDIX B – Dataset Preparation

Source Code B.1 – Data preparation script.

```
import warnings
warnings.simplefilter(action='ignore', category=FutureWarning)

import os
import glob
import argparse
import random
import numpy as np
from time import time
from sklearn.model_selection import train_test_split
import cv2
import pickle

import caffe
from caffe.proto import caffe_pb2
import lmdb

ap = argparse.ArgumentParser()
ap.add_argument('-p', '--path', required=True, help='Path to dataset
    directory')
args = vars(ap.parse_args())

# Size of images
IMAGE_WIDTH = 224
IMAGE_HEIGHT = 224


def transform_img(img, img_width=IMAGE_WIDTH, img_height=IMAGE_HEIGHT,
    equalize=False):
    """Function that resize an image and equalize it if necessary."""
    if equalize:
        # Histogram Equalization
        img[:, :, 0] = cv2.equalizeHist(img[:, :, 0])
```

```
32          img[:, :, 1] = cv2.equalizeHist(img[:, :, 1])
33          img[:, :, 2] = cv2.equalizeHist(img[:, :, 2])
34
35      # Image Resizing
36      img = cv2.resize(img, (img_width, img_height), interpolation =
        ↪  cv2.INTER_CUBIC)
37
38      return img
39
40
41  def make_datum(img, label):
42      # Image is numpy.ndarray format. BGR instead of RGB
43      return caffe_pb2.Datum(
44          channels=3,
45          width=IMAGE_WIDTH,
46          height=IMAGE_HEIGHT,
47          label=label,
48          data=np.rollaxis(img, 2).tostring())
49
50  path = args['path']
51  parent_path = os.path.sep.join(path.split(os.path.sep)[:-1])
52  sibling_path = path.split(os.path.sep)[-1] + '_lmdb'
53  sibling_path = os.path.sep.join([parent_path, sibling_path])
54  train_lmdb = os.path.sep.join([sibling_path, 'train'])
55  validation_lmdb = os.path.sep.join([sibling_path, 'validation'])
56
57  if not os.path.exists(sibling_path):
58      os.makedirs(sibling_path)
59
60  os.system('rm -rf  ' + train_lmdb)
61  os.system('rm -rf  ' + validation_lmdb)
62
63  dataset = []
64  for r, dirs, files in os.walk(path):
65      if len(dirs) > 0:
66          labels = dirs
67          continue  # use only leaf folders
68      files_full_path = ['{}/{}'.format(r, f) for f in files]
69      directory_name = r.split(os.path.sep)[-1]
```

```
70      dataset.append((files_full_path, directory_name))
71
72  label_dict = [(l, i) for i, l in enumerate(labels)]
73  label_dict = dict(label_dict)
74
75  '''
76  Save dictionary in the form of:
77
78  label_dict = {
79      'basalcellcarcinoma': 0,
80      'lentigo': 1,
81      'malignantmelanoma': 2,
82      'pigmentednevus': 3,
83      'seborrheickeratosis': 4,
84      'wart': 5,
85      ...
86  }
87  '''
88
89  with open('label_dict.pkl', 'wb') as f:
90      pickle.dump(label_dict, f)
91      f.close()
92
93  X = [(img, label) for ndataset, label in dataset for img in ndataset]
94  y = [label_dict[label] for _, label in X]
95
96  # Shuffle dataset
97  random.shuffle(X)
98
99  train_data, test_data, _, _ = train_test_split(X, y, train_size=0.8,
    ↪  stratify=y)
100
101 print('Creating train_lmdb...')
102
103 train_time = time()
104 in_db = lmdb.open(train_lmdb, map_size=int(1e12))
105 with in_db.begin(write=True) as in_txn:
106     for in_idx, (img_path, label) in enumerate(train_data):
107         if in_idx % 100 == 0:
```

```
108             print('Processed {}/{}'.format(in_idx, len(train_data)),
                ↪  end='\r')
109
110         img = cv2.imread(img_path, cv2.IMREAD_COLOR)
111         img = transform_img(img, img_width=IMAGE_WIDTH,
            ↪  img_height=IMAGE_HEIGHT)
112
113         num_label = label_dict[label]
114         datum = make_datum(img, num_label)
115
116         key = '{:0>6d}'.format(in_idx)
117         in_txn.put(key.encode(), datum.SerializeToString())
118
119
120 in_db.close()
121 print('Finished {} train_lmdb in {:.2f} sec'.format(len(train_data),
    ↪  (time() - train_time)))
122
123
124 print('\nCreating validation_lmdb...')
125
126 test_time = time()
127 in_db = lmdb.open(validation_lmdb, map_size=int(1e12))
128 with in_db.begin(write=True) as in_txn:
129     old_t = time()
130     for in_idx, (img_path, label) in enumerate(test_data):
131         if in_idx % 100 == 0:
132             print('Processed {}/{}'.format(in_idx, len(test_data)),
                ↪  end='\r\r')
133
134             old_t = time()
135
136         img = cv2.imread(img_path, cv2.IMREAD_COLOR)
137         img = transform_img(img, img_width=IMAGE_WIDTH,
            ↪  img_height=IMAGE_HEIGHT)
138
139         num_label = label_dict[label]
140         datum = make_datum(img, num_label)
141
```

```
142         key = '{:0>6d}'.format(in_idx)
143         in_txn.put(key.encode(), datum.SerializeToString())
144
145
146 in_db.close()
147 print('Finished {} test_lmdb in {:.2f} sec'.format(len(test_data),
    ↪  (time() - test_time)))
148
149 print('\nFinished processing all images in {:.2f}'.format(time() -
    ↪  train_time))
```

APPENDIX C – Configuration file for CNN training

Source Code C.1 – Caffe solver *.prototxt* configuration file.

```
net:
  ↪ "./experiments/architectures/resnet152/train/ResNet_152_train.prototxt"
iter_size: 12
test_iter: 22023
test_interval: 2000
test_initialization: false
display: 50
base_lr: 0.01
lr_policy: "step"
stepsize: 17618
gamma: 0.1
momentum: 0.9
weight_decay: 0.00001
max_iter: 176180
snapshot: 2000
snapshot_prefix:
  ↪ "./experiments/architectures/resnet152/train/med_atlas_edin_2"
solver_mode: GPU
```

APPENDIX D – Metrics results of evaluation dataset for best experiment

Table 6 – Classification report for predictions on evaluation dataset.

Lesion	**precision**	**recall**	**f1-score**	**support**
Lentigo	0.38	0.62	0.47	13
Haemangioma	0.81	0.72	0.76	18
Seborrhoeic Keratosis	0.84	0.73	0.78	208
Actinic Keratosis	0.75	0.38	0.50	8
Wart	0.94	0.75	0.83	79
Basal Cell Carcinoma	0.90	0.83	0.86	324
Malignant Melanoma	0.76	0.71	0.73	68
Dermatofibroma	0.71	0.83	0.77	12
Pyogenic Granuloma	0.83	1.00	0.91	5
Melanocytic Nevus	0.75	0.83	0.79	191
Intrapithelial Carcinoma	0.71	0.71	0.71	14
Squamous Cell Carcinoma	0.21	0.81	0.33	16
avg / total	0.82	0.78	0.80	956

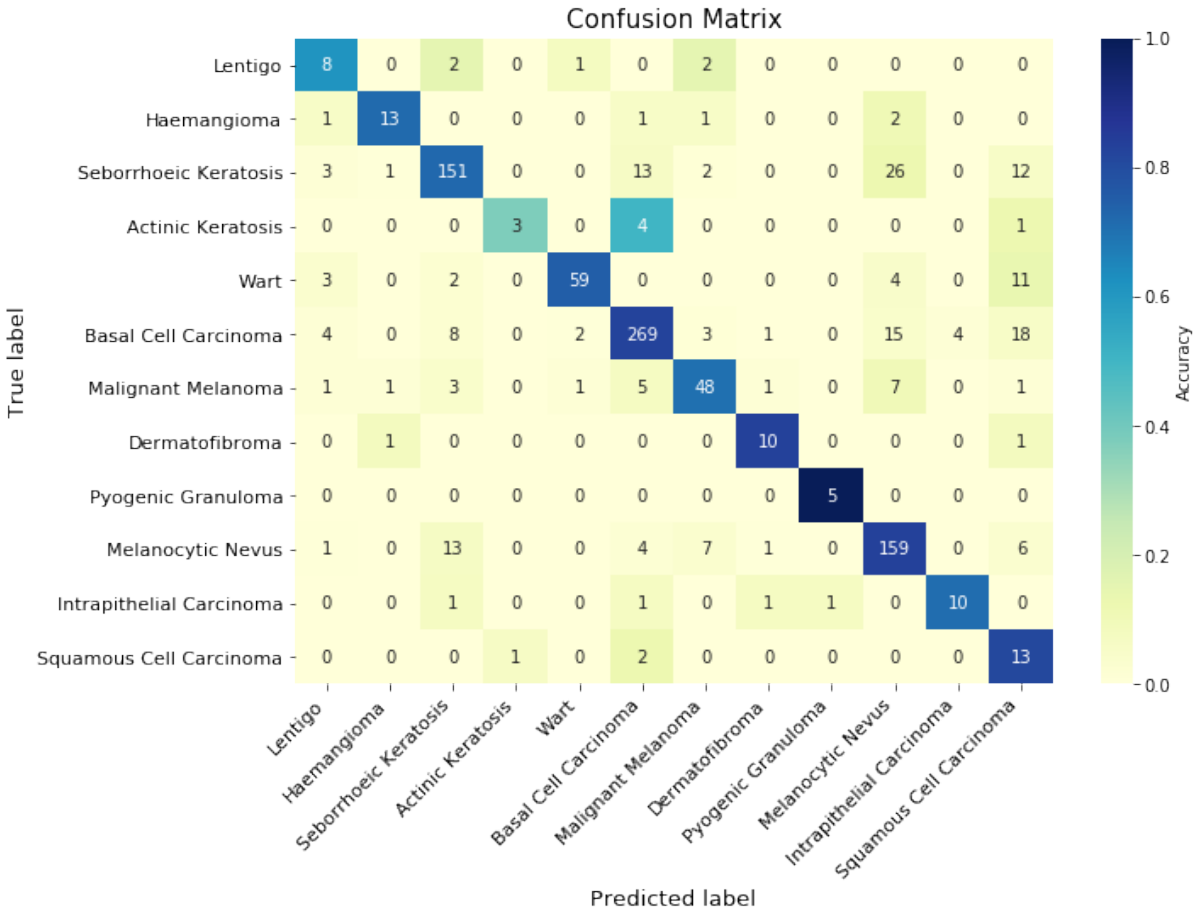
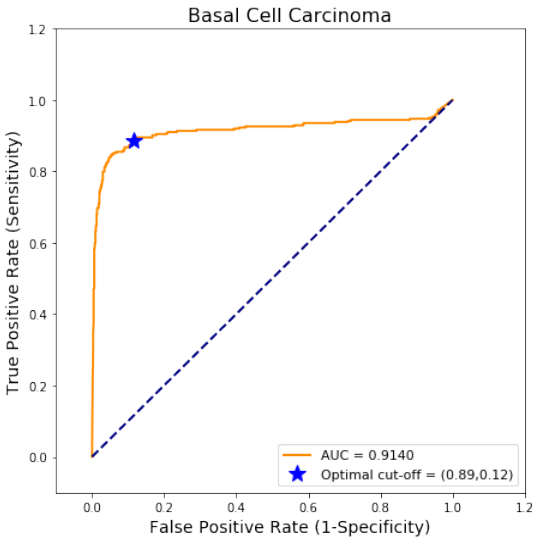
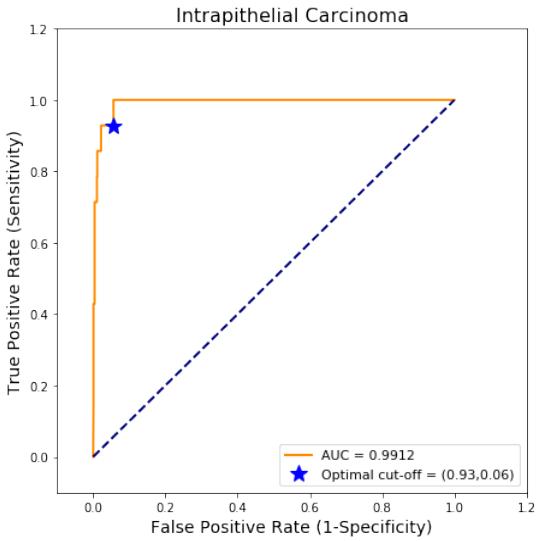


Figure 29 – Confusion matrix for the 12 skin lesions.

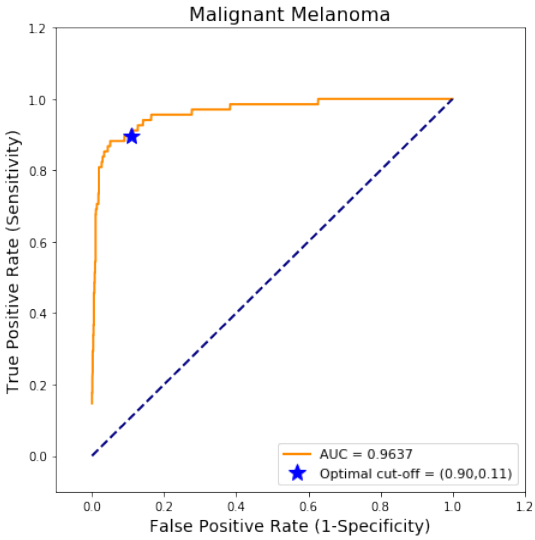
Source – Author.



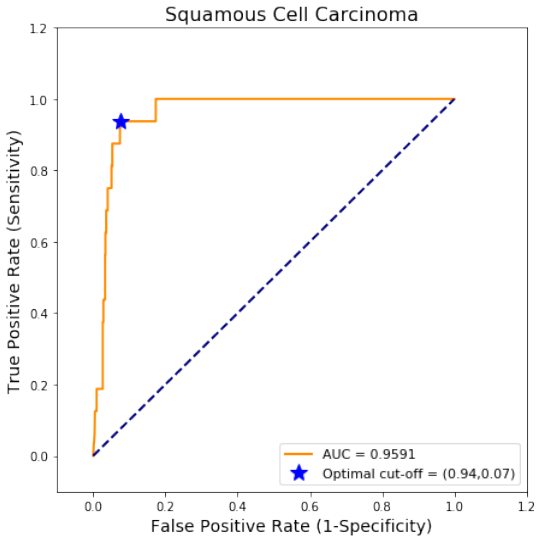
(a) Basal Cell Carcinoma ROC curve



(b) Intrapithelial Carcinoma ROC curve



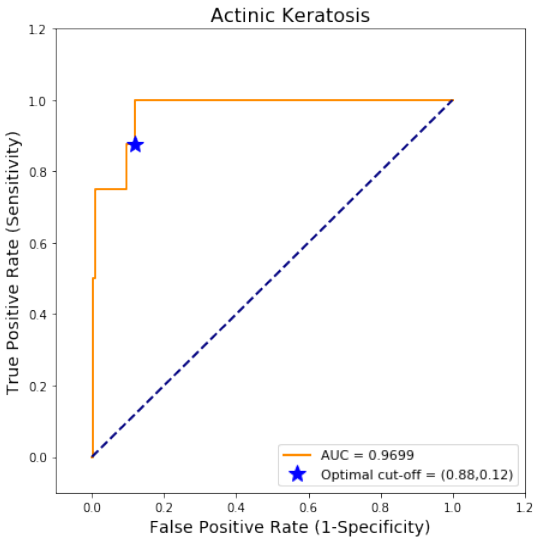
(c) Malignant Melanoma ROC curve



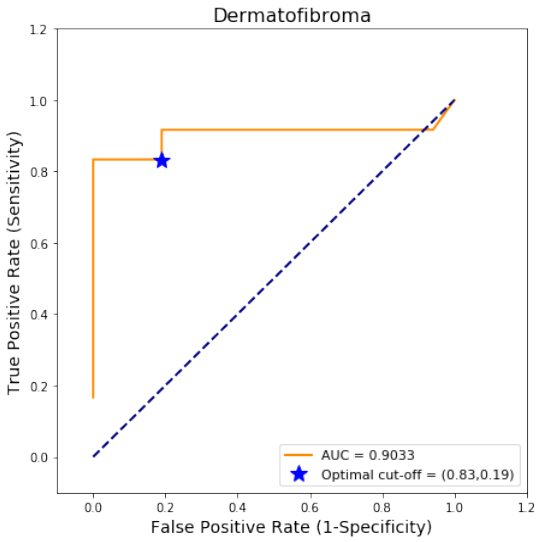
(d) Squamous Cell Carcinoma ROC curve

Figure 30 – ROC curve of the skin lesions.

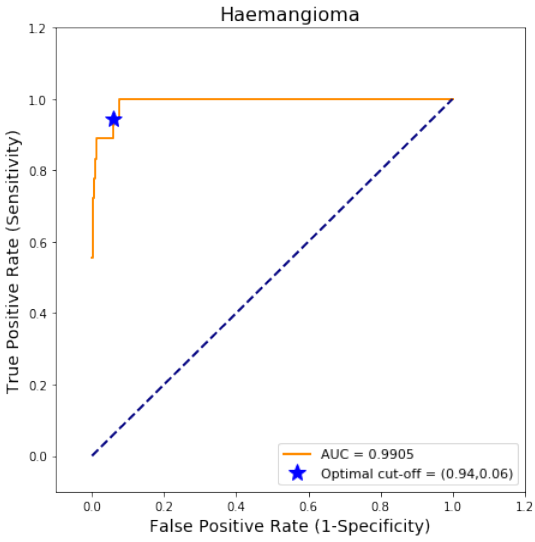
Source – Author.



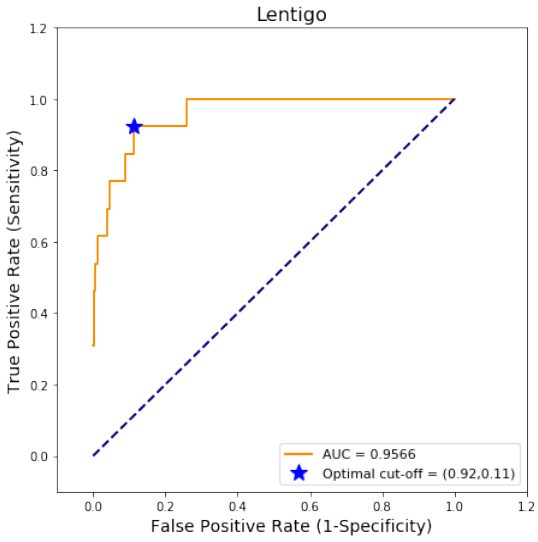
(a) Actinic Keratosis ROC curve



(b) Dermatofibroma ROC curve



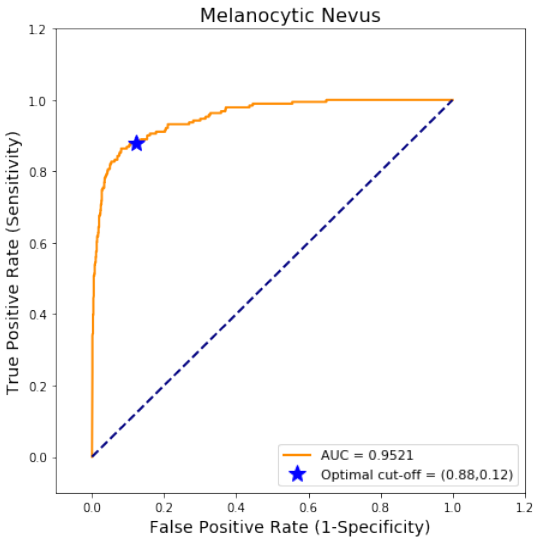
(c) Hemangioma ROC curve



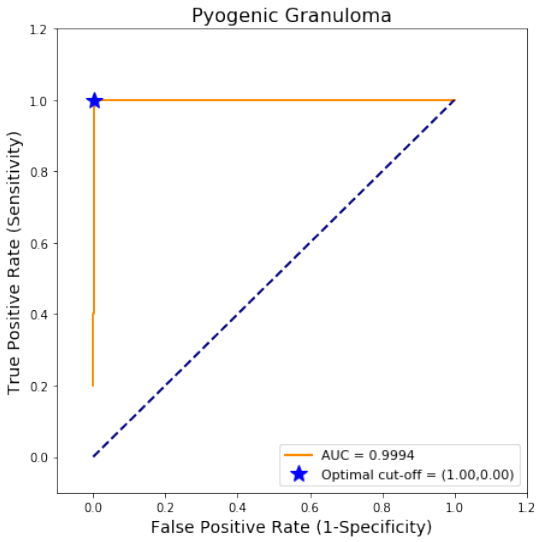
(d) Lentigo ROC curve

Figure 31 – ROC curve of the skin lesions. Continued 2/3.

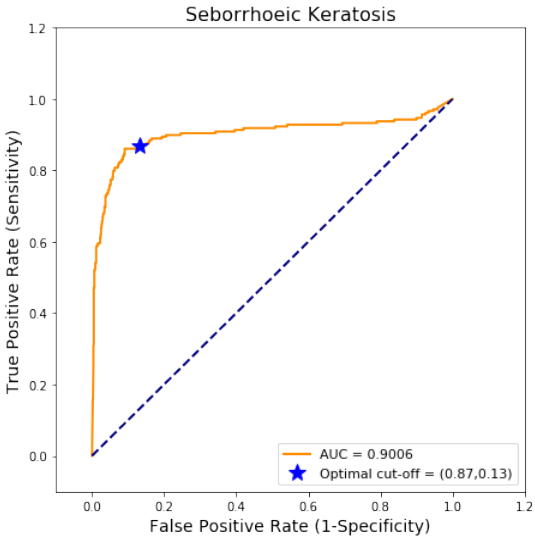
Source – Author.



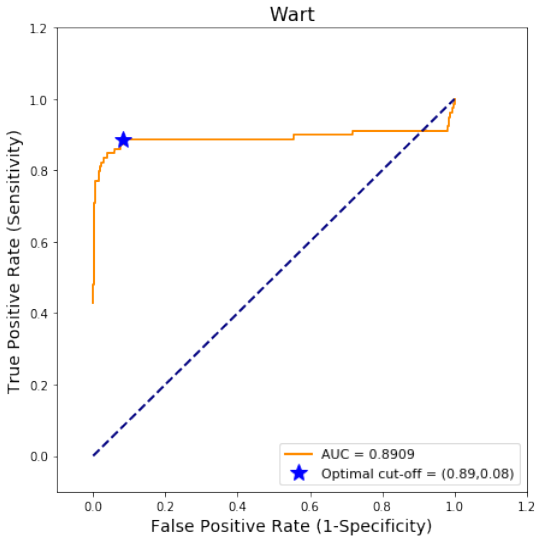
(a) Melanocytic nevus ROC curve



(b) Pyogenic Granuloma ROC curve



(c) Seborrheic Keratosis ROC curve



(d) Wart ROC curve

Figure 32 – ROC curve of the skin lesions. Continued 3/3.

Source – Author.

APPENDIX E – Interpretability graphics

E.1 Most Correct Results

(a) Basal Cell Carcinoma (b) Intrapithelial Carcinoma (c) Malignant Melanoma

(d) Squamous Cell Carcinoma (e) Actinic Keratosis (f) Dermatofibroma

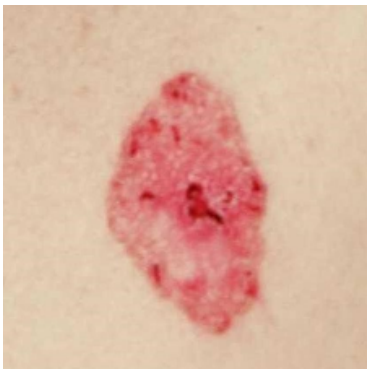
(g) Hemangioma (h) Lentigo (i) Melanocytic nevus

(j) Pyogenic Granuloma (k) Seborrheic Keratosis (l) Wart

Figure 33 – Most correctly predicted lesions in the dataset. All were predicted with a probability of 1.00.

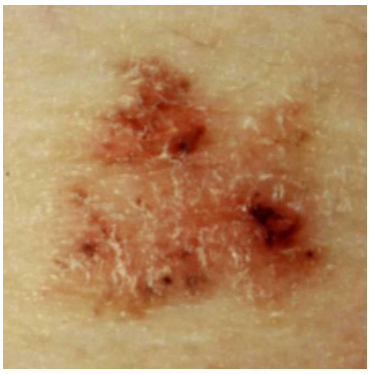
Source – Edinburgh & Atlas datasets.

E.2 Most Wrong Results



(a) Basal Cell Carcinoma that was predicted as Intrapithelial Carcinoma.

(b) Seborrheic Keratosis that was predicted as Melanocytic Nevus.



(c) Seborrheic Keratosis that was predicted as BCC.

(d) Seborrheic Keratosis that was predicted as SCC.

(e) Wart that was predicted SCC.

(f) Wart that was predicted SCC.

Figure 34 – Most wrong predictions in the dataset. All were predicted with a probability of 0.0 *vs* 1.0.

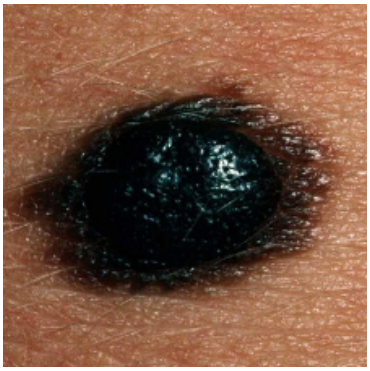
Source – Edinburgh & Atlas datasets.

E.3 Most Undecided Results

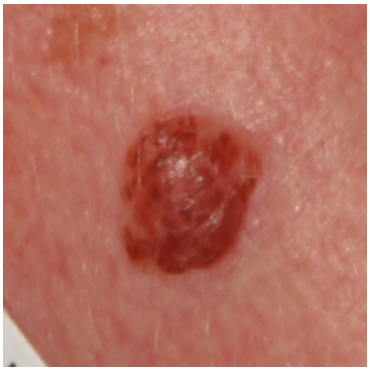
(a) Basal Cell Carcinoma that was predicted as Melanocytic Nevus, with second guess as Actinic Keratosis (t:0 *vs* p:0.5 *vs* s:0.38).

(b) Basal Cell Carcinoma that was predicted as Malignant Melanoma Nevus, with second guess as BCC (t:0.43 *vs* p:0.56).

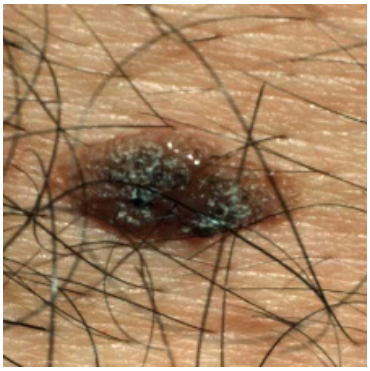
(c) Basal Cell Carcinoma that was predicted as Melanocytic Nevus, with second guess as BCC (t:0.46 *vs* p:0.54).



(d) Melanocytic Nevus that was predicted as Malignant Melanoma, with second guess as Malignant Melanoma (t:0.44 *vs* p:0.56).



(e) Melanocytic Nevus that was predicted as Malignant Melanoma (t:0.44 *vs* p:0.56).



(f) Seborrheic Keratosis that was predicted as Melanocytic Nevus (t:0.49 *vs* p:0.51).

Figure 35 – Undecided predictions defined with a delta of 0.15. t: true, p: predicted, s: second top prediction.

Source – Edinburgh & Atlas datasets.

E.4 GradCAM Implementation

1 Gradient-weighted Class Activation Mapping (GradCAM)

In this notebook we will implement the technique proposed by (RIBEIRO; SINGH; GUESTRIN, 2016).

```
In [1]: import caffe # pycaffe library
        import numpy as np
        import matplotlib.pyplot as plt
        %matplotlib inline

        # set display defaults
        plt.rcParams['figure.figsize'] = (10, 10)        # large images
        # don't interpolate: show square pixels
        plt.rcParams['image.interpolation'] = 'nearest'
        # use grayscale output rather than a color heatmap
        plt.rcParams['image.cmap'] = 'gray'
```

1.0.1 2. Load the architecture and weights and define the pre-process step

Set Caffe to GPU mode and load the net from disk.

```
In [2]: caffe.set_mode_gpu()

        # Used a copy of the original model file, to edit it
        model_def = './experiments/architectures/resnet152/train' + \
            '/ResNet_152_deploy.skin.prototxt'

        if not 'force_backward: true' in open(model_def).read():
            with open(model_def, 'a') as f:
                f.write('force_backward: true\n')

        model_weights = './experiments/architectures/resnet152/train' + \
            '/med_atlas_edin_2/ResNet-152-dan-solver_1_iter_38000.caffemodel'

        net = caffe.Net(model_def,      # defines the structure of the model
                        model_weights,  # contains the trained weights
                        caffe.TEST)     # use test mode (e.g., don't perform drop-
                                        # out)
```

2.1 Set up the pre-processing step For the pre-processing, we will use the mean image generated for the dataset. This is used to subtract the mean values for each channel (in RGB order) to normalize the input data.

Furthermore, for this process we will use Caffe's `caffe.io.Transformer`.

```
In [3]: from caffe.proto import caffe_pb2

        mean_blob = caffe_pb2.BlobProto()
        with open(
```

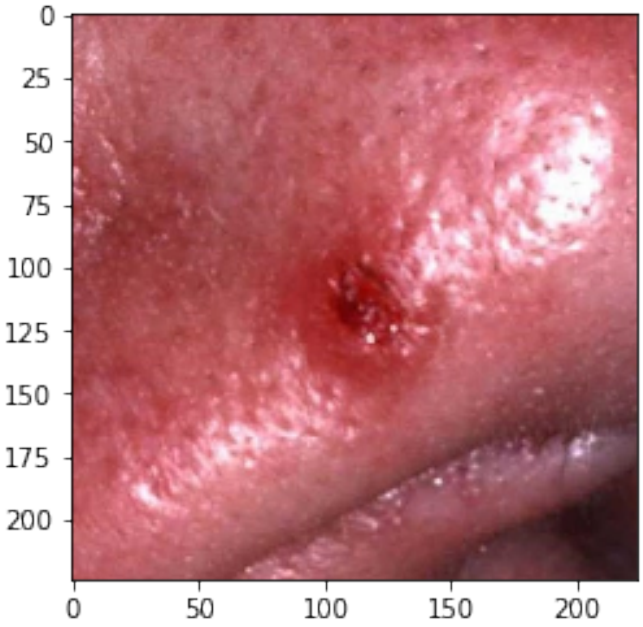
```
            './data/edited/Med_atlas_edin_29x_lmdb/input/mean.binaryproto',
            'rb'
        ) as f:

        mean_blob.ParseFromString(f.read())
    mean_array = np.asarray(mean_blob.data, dtype=np.float32).reshape(
        (mean_blob.channels, mean_blob.height, mean_blob.width)
    )

    transformer = caffe.io.Transformer({'data': net.blobs['data'].data.shape})
    transformer.set_mean('data', mean_array)
    transformer.set_transpose('data', (2,0,1))
```

1.0.2 3. Classification

Define the function to handle the resize of the images.
Load the chosen image and perform the set up pre-processing step.

```
In [4]: IMAGE_WIDTH = 224
        IMAGE_HEIGHT = 224

        def transform_img(
                img, img_width=IMAGE_WIDTH,
                img_height=IMAGE_HEIGHT, equalize=False
            ):

            if equalize:
                # Histogram Equalization
                img[:, :, 0] = cv2.equalizeHist(img[:, :, 0])
                img[:, :, 1] = cv2.equalizeHist(img[:, :, 1])
                img[:, :, 2] = cv2.equalizeHist(img[:, :, 2])

            # Image Resizing
            img = cv2.resize(img,
                             (img_width, img_height),
                             interpolation=cv2.INTER_CUBIC
                            )

            return img

In [5]: import cv2

        test_img = './data/validation/Med_atlas_edin/' + \
            'basal_cell_carcinoma/402_003072HB.JPG'
        image = caffe.io.load_image(test_img)

        image = cv2.imread(test_img, cv2.IMREAD_COLOR)
        image = transform_img(image,
```

```
                                img_width=IMAGE_WIDTH,
                                img_height=IMAGE_HEIGHT
                               )

transformed_image = transformer.preprocess('data', image)
visualization_img = cv2.cvtColor(image, cv2.COLOR_BGR2RGB)
plt.imshow(visualization_img)
plt.show()
```



Perform the foward pass with the image to collect the predictions.

```
In [6]: # copy the image data into the memory allocated for the net
        net.blobs['data'].data[...] = transformed_image

        ### perform classification
        output = net.forward()
        print(output['prob'].shape)

        # the output probability vector for the first image in the batch
        output_prob = output['prob'][0]
        print(output_prob.shape)

        print('predicted class is:', output_prob.argmax())
```

```
(1, 12)
(12,)
predicted class is: 5
```

With that we can transform the integer prediction to the lesion name, using the cheat sheet saved in the setup of the dataset.

```
In [7]: import pickle
        with open(
                './experiments/architectures/resnet152/train/' + \
                'med_atlas_edin_2/label_dict.pkl',
                'rb'
            ) as f:

            label_dict = pickle.load(f)
            f.close()

        labels = dict(zip(list(label_dict.values()), list(label_dict.keys())))
        labels[output_prob.argmax()]

Out[7]: 'basal_cell_carcinoma'
```

Here we see that the prediction was right about the lesion class. However, for the sake of it, let's examine the full probabilities array.

But just for analysis purposes, let's sort the top five predictions to see if the other predictions are sensible.

```
In [8]: top_inds = output_prob.argsort()[::-1][:5]

        list(
            zip(output_prob[top_inds],
            [l for i, l in enumerate(list(labels.values())) if i in top_inds ])
        )

Out[8]: [(1.0, 'seborrhoeic_keratosis'),
         (3.4086275e-13, 'wart'),
         (9.648443e-14, 'basal_cell_carcinoma'),
         (2.0363428e-14, 'pyogenic_granuloma'),
         (1.633301e-14, 'squamous_cell_carcinoma')]
```

Therefore, it was a 100% sure classification.

1.0.3 4. GradCAM Visualization

With the label at hands, now we can use the GradCAM visualization technique to see what are the regions of the most interest for this label.

```
In [9]: from PIL import Image
        import matplotlib.cm as cm

        net.blobs['data'].data[...] = transformed_image
        output = net.forward()
```

```
output_prob = output['prob'][0]
print('predicted class is: {} = {}'.format(
        labels[output_prob.argmax()],
        output_prob.argmax())
     )

final_layer = 'fc1000-skin' # dense layer
vis_layer = 'res5c_branch2c' # visualization layer (Last convolution layer)
image_size = (224,224) # input image size
filter_shape = (7, 7) # size of the receptive field in the vis_layer
category_index = output['prob'].argmax()

# create prediction a array with 100% for the
# predicted class and 0% for the rest.
label_index = output_prob.argmax()
caffeLabel = np.zeros(net.blobs[final_layer].shape)
caffeLabel[0, label_index] = 1;

# hook for the gradients in the vis_layer, computed in the backward pass
grads = net.backward(
        diffs=['data', vis_layer],
        **{net.outputs[0]: caffeLabel}
    )

vis_grad = grads[vis_layer] # gradients for the predicted class

# Compute the mean gradients
vis_grad = vis_grad / (np.sqrt(np.mean(np.square(vis_grad))) + 1e-5)
vis_grad = vis_grad[0,:,:,:]

# Get the mean weights
weights = np.mean(vis_grad, axis=(1, 2))

# Get the activations of the neurons in the vis_layer to
# then compute the visualization map with the weights
vis = np.zeros(filter_shape, dtype=np.float32)
activations = net.blobs[vis_layer].data[0, :, :, :]
for i, w in enumerate(weights):
    vis += w * activations[i, :, :]

# We select only those activation which has positively
# contributed in prediction of given class
vis = np.maximum(vis, 0) # ReLU activation
vis_img = Image.fromarray(vis, None)
vis_img = vis_img.resize((224,224),Image.BICUBIC)
vis_img = vis_img / np.max(vis_img)
vis_img = Image.fromarray(np.uint8(cm.jet(vis_img) * 255))
vis_img = vis_img.convert('RGB') # dropping alpha channel
```

```
        input_image = Image.open(test_img)
        input_image = input_image.resize((224,224))
        input_image = input_image.convert('RGB')

        heat_map = Image.blend(input_image, vis_img, 0.3)
        plt.imshow(heat_map)
        plt.axis('off')
        plt.show()
        plt.imshow(vis_img)
        plt.axis('off')

        fig = plt.figure(figsize=(10, 10))
        columns = 3
        rows = 1
        images = [input_image, heat_map, vis_img]
        for i in range(1, columns*rows +1):
            img = np.random.randint(10, size=(10, 10))
            fig.add_subplot(rows, columns, i)
            plt.imshow(images[i-1])
            plt.axis('off')
        plt.savefig('grad_cam_test.png')
        plt.show()
```

```
predicted class is: basal_cell_carcinoma = 5
```