UNIVERSIDADE FEDERAL FLUMINENSE

ADRIANO LIMA E SOUZA

CLASSIFICATION OF SKIN LESIONS IN DARK SKIN USING VGG AND ISIC2020 DATA SET

NITERÓI

2024

ADRIANO LIMA E SOUZA

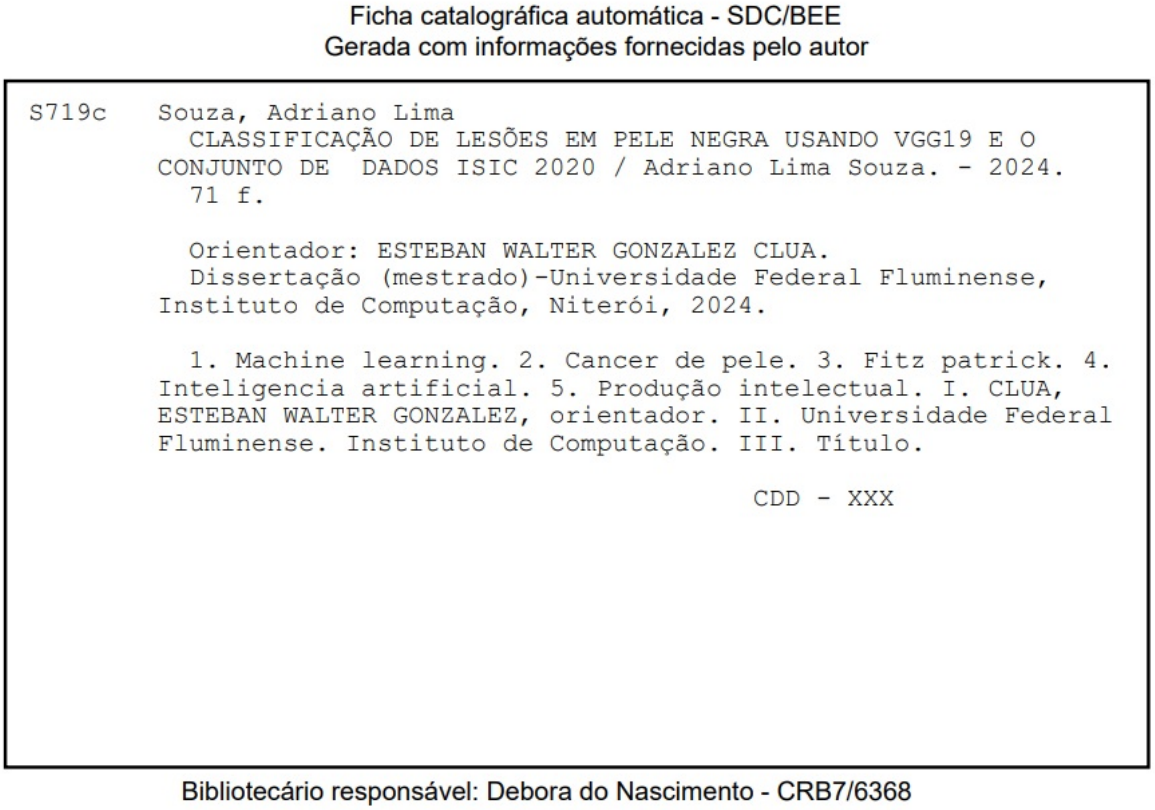
CLASSIFICATION OF SKIN LESIONS IN DARK SKIN USING VGG AND ISIC2020 DATA SET

Master's Dissertation submitted to the Graduate Program in Computing of the Fluminense Federal University as a partial requirement for obtaining the degree of Master in Computing. Area of concentration: COMPUTER SCIENCE

Orientador:
ESTEBAN WALTER GONZALEZ CLUA

NITERÓI

2024



Ficha catalográfica automática - SDC/BEE
Gerada com informações fornecidas pelo autor

S719c Souza, Adriano Lima
CLASSIFICAÇÃO DE LESÕES EM PELE NEGRA USANDO VGG19 E O CONJUNTO DE DADOS ISIC 2020 / Adriano Lima Souza. - 2024.
71 f.

Orientador: ESTEBAN WALTER GONZALEZ CLUA.
Dissertação (mestrado)-Universidade Federal Fluminense, Instituto de Computação, Niterói, 2024.

1. Machine learning. 2. Cancer de pele. 3. Fitz patrick. 4. Inteligencia artificial. 5. Produção intelectual. I. CLUA, ESTEBAN WALTER GONZALEZ, orientador. II. Universidade Federal Fluminense. Instituto de Computação. III. Título.

CDD - XXX

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

Figura 1: Location of the catalog card

ADRIANO LIMA E SOUZA

CLASSIFICATION OF SKIN LESIONS IN DARK SKIN USING VGG AND ISIC2020 DATA SET

Master's Dissertation submitted to the Graduate Program in Computing of the Fluminense Federal University as a partial requirement for obtaining the degree of Master in Computing. Area of concentration: COMPUTER SCIENCE

Expected approval in AUGUST 2024.

EXAMINING COMMITTEE

Prof. ESTEBAN WALTER GONZALEZ CLUA – Advisor, UFF

Prof. FLAVIO LUIZ SEIXAS, UFF

Prof. LUIZ MARCOS GARCIA GONÇALVES, UFRN

Niterói
2024

To God, source of wisdom and constant inspiration, To my beloved family, for their unconditional love and unwavering support, To my dedicated advisor, whose guidance was essential for this work, To the esteemed professors, for their generosity in sharing knowledge and experience.

Acknowledgments

I thank God for my achievements.

To my advisor, for showing me the paths to follow and for the trust placed in me.

To my family, for their unconditional support and for always believing in me.

Abstract

The classification of skin lesion images is a crucial task in the early detection of skin cancer, especially melanoma, which is the most severe form. This paper proposes the application of the VGG19 convolutional neural network architecture to classify seven types of skin lesions using the ISIC 2020 dataset. Furthermore, the study investigates the application of the Fitzpatrick Scale to assess how different skin tones may influence classification accuracy. It also discusses how the use of computer vision to identify and classify skin according to the Fitzpatrick Scale can improve diagnostic accuracy. The aim is to demonstrate how feature extraction through VGG19 can significantly enhance diagnostic precision, benefiting both patients and physicians. Different performance metrics are applied and discussed to evaluate the effectiveness of the model.

Keywords: Image classification, Skin lesions, melanoma detection, VGG19, ISIC 2020 dataset, Fitzpatrick Scale, computer vision, diagnostic accuracy, convolutional neural network.

Abstract

The classification of skin lesion images is a crucial task in the early detection of skin cancer, particularly melanoma, which is the most severe form. This paper explores the application of the convolutional neural network architecture VGG19 to classify seven types of skin lesions using the ISIC 2020 dataset. Additionally, the study investigates the application of the Fitzpatrick Scale to assess how different skin tones can influence classification accuracy, as well as the use of computer vision to identify skin and classify it according to the Fitzpatrick Scale to improve accuracy and precision. The objective is to demonstrate how feature extraction through VGG19 can significantly improve diagnostic accuracy, benefiting both patients and healthcare professionals. Various performance metrics are applied and discussed to evaluate the model's effectiveness.

Keywords: Image classification, skin lesions, melanoma detection, VGG19, ISIC 2020 dataset, Fitzpatrick Scale, computer vision, diagnostic accuracy, convolutional neural networks.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introduction

Skin cancer is one of the most common types of cancer worldwide. Among its various forms, melanoma is the most aggressive and potentially fatal when not detected in the early stages. Accurate identification of skin lesions is therefore essential for effective treatment. Different convolutional neural networks (CNNs), such as VGG19, have shown potential in the task of classifying medical images due to their ability to extract relevant features from complex images [21].

The term "melanocytic lesion"refers to a group of different types of skin lesions that arise from the proliferation of melanocytes, which can be either benign or malignant. Among the malignant lesions, melanoma is the most aggressive cutaneous neoplasm and often exhibits rapid progression and metastatic spread. Based on their clinical and histological characteristics, melanomas can be subdivided into four main categories: nodular melanoma, lentigo maligna melanoma, acral melanoma, and superficial spreading melanoma. The application of the Fitzpatrick Scale is crucial to understanding how different skin tones can influence the detection and classification of these skin lesions.

1.0.1 Innovation Potential

The innovation potential of this study lies in its multifaceted approach to the classification of skin lesions. By addressing ethnic representativeness and employing advanced deep learning techniques, the developed model can be adapted to a variety of clinical contexts, improving diagnostic accuracy across different populations. This can lead to increased equity in healthcare, a critical area that has traditionally been overlooked in computer vision studies.

1.0.2 Challenges in Collaboration

Interdisciplinary collaboration is essential for the success of projects involving the application of CNNs in medical diagnostics, as it integrates knowledge from areas such as

computing, medicine, and ethics, resulting in more comprehensive and effective solutions. Studies such as that of Turing [40] emphasize the importance of multidisciplinary teams in developing AI systems that are both technically sound and clinically relevant. The relevance of this study to computing lies in the practical application of deep learning techniques to solve public health problems, such as the early detection of melanoma, which can save lives. The creation of a robust automated system assists dermatologists in the screening and diagnosis of skin lesions, and the combination of CNNs with data augmentation techniques and ethnic variability analysis promotes significant advances in building fairer and more accurate models.

Despite these benefits, interdisciplinary collaboration faces significant challenges, including differences in terminology and methodologies used by different disciplines, as well as the need for effective communication among team members. Studies by Woolliscroft and Howell [41] suggest that education and training programs that promote mutual understanding and effective communication can help overcome these challenges.

2 Literature Review

The literature review on skin lesion classification and early melanoma detection includes several studies that investigate different approaches and technologies aimed at improving diagnostic accuracy and treatment. Such research ranges from the application of conventional techniques to the implementation of advanced neural networks, seeking to develop more effective solutions to identify and treat potentially dangerous lesions, such as melanoma.

The use of convolutional neural networks (CNNs) in medical image classification, especially for the early detection of melanoma, has attracted growing interest in recent years. This particular study adopts architectures such as EFFICIENTNET, DENSENET, MOBILENET, RESNET50, VGG16, VGG19, and many others to classify skin lesions, demonstrating how feature extraction significantly improves diagnostic accuracy. The importance of this research can be contextualized by its comparison with other recent studies, reinforcing its impact in the fields of computing and computer vision.

2.0.1 Comparison with Recent Studies

The work of Esteva et al. [12] used deep learning for skin cancer classification, achieving accuracy levels comparable to those of dermatologists. The authors trained a model based on the Inception v3 architecture with a large dermatological dataset, showing that the neural network could identify melanoma with significantly high accuracy [12]. Our study's approach, using VGG19, presents similar results and complements this line of research by incorporating the Fitzpatrick Scale to assess the influence of skin tones on model accuracy.

Another relevant study is that of Haenssle et al. [17], which compared the performance of a CNN-based model with that of dermatologists in identifying melanomas through dermatoscopic images. The CNN outperformed most dermatologists in diagnostic accuracy, highlighting the potential of CNNs as support tools in clinical diagnosis [16]. Our study contributes additionally by focusing on the ethnic diversity of patients, something not

deeply explored in these previous studies, as shown in Table 2.1.

CAPES Research 2019–2024	**Number of Results**
melanoma AND deep learning	3658
skin cancer AND deep learning	2618
skin cancer AND machine learning	1134
melanoma AND VGG16	98
black skin AND melanoma AND deep learning	45
black skin AND melanoma AND machine learning	24
black skin AND melanoma AND VGG16	4
black skin AND melanoma AND RESNET	3
black skin AND melanoma AND VGG19	1
skin cancer AND fitzpatrick	945
skin cancer AND fitzpatrick AND machine learning	8
skin cancer AND fitzpatrick AND deep learning	20

Tabela 2.1: CAPES Research Results 2019–2024

2.0.2 Innovation and Impact on Computing

This study not only validates the effectiveness of CNNs in classifying skin lesions, but also introduces a detailed analysis of ethnic variability using the Fitzpatrick Scale. This aspect is innovative as it addresses a significant gap in the representativeness of dermatological datasets. Most computer vision models are trained primarily with images of light-skinned individuals, which can lead to diagnostic bias.

The employed methodology, which includes data augmentation tailored to different skin tones, regularization with dropout and L2 (techniques used to prevent overfitting, which occurs when a model fits too closely to the training data and loses the ability to generalize well to new data), and continuous evaluation with rigorous metrics, contributes to the development of a more robust and generalizable model.

2.1 Regularization with Dropout and L2

In the context of neural networks and deep learning, regularization with *dropout* and *L2* refers to techniques used to prevent *overfitting*, which occurs when a model fits too closely to the training data, losing its generalization capability.

2.1.1 L2 Regularization

L2 regularization, also known as *Ridge Regularization*, adds a penalty term to the model's loss function proportional to the square of the network weights. The goal of this technique is to encourage the network to keep weights small, preventing the model from becoming overly complex or sensitive to the training data. The loss function with *L2* regularization can be expressed as follows:

$$Loss = Loss_{original} + \lambda \sum w^2$$

where λ is the hyperparameter that controls the strength of the penalty and w are the neural network weights. This helps simplify the model and make it more robust.

2.1.2 Dropout

Dropout is a stochastic regularization technique used during the training of neural networks, in which some units (*neurons*) in the network are randomly "turned off" at each training iteration, meaning their contributions are temporarily ignored. This forces the network to learn more robust patterns instead of relying too heavily on specific neurons. During the inference phase, all units are used, but their weights are adjusted to compensate for the absence of *dropout* during training. This technique is widely effective for deep and dense networks, helping to prevent *overfitting*.

Adding more layers to the model also enhances the network's feature extraction capability, allowing for more detailed and accurate detection.

2.1.3 Comparison with Other Articles in the Field

The study by Tschandl et al. (2020) investigated the use of different CNN architectures for skin lesion classification, including ResNet and DenseNet, as well as Inception v4. The results showed that deeper and more complex architectures, such as DenseNet, provided a marginal improvement in accuracy, but with a significantly higher computational cost [39]. In comparison, our use of VGG19 offers a balance between complexity and performance, being more efficient in terms of training and inference time while still maintaining high accuracy.

In addition, the work by Brinker et al. (2019) focused on transfer learning, using pre-trained models on large general image datasets before fine-tuning with specific derma-

tological images. This approach proved effective in improving model accuracy, especially when domain-specific data is scarce [6]. Our study follows a similar approach, using a pre-trained VGG19, but introduces an innovative method by integrating analysis based on Fitzpatrick skin phototypes.

3 Classification and Causes of Cancer Types

Cancers can be classified according to the origin of the malignant cells and their location in the body. Among the most common types are carcinoma, sarcoma, lymphoma, leukemia, and melanoma. Each of these cancer types has distinct characteristics and behaviors, which directly influence the diagnosis, treatment, and prognosis of patients.

- **Carcinoma**: Carcinoma is the most prevalent type of cancer and originates in epithelial cells, which line the skin and the internal tissues of organs. Examples include basal cell carcinoma, squamous cell carcinoma, and adenocarcinoma.
- **Sarcoma**: This type of cancer arises in the cells of connective tissues, such as bones, muscles, cartilage, and fat. Osteosarcoma (bone) and liposarcoma (adipose tissue) are typical examples of sarcomas.
- **Lymphoma**: Lymphoma affects the cells of the lymphatic system, which is part of the immune system. It is classified into two main types: Hodgkin's lymphoma and non-Hodgkin's lymphoma.
- **Leukemia**: Leukemia is characterized by the abnormal proliferation of blood cells, especially leukocytes. It begins in the bone marrow and can spread to the bloodstream, interfering with the normal production of blood cells.
- **Melanoma**: Melanoma is an aggressive type of skin cancer that develops from melanocytes, the cells that produce pigment in the skin. It is the most dangerous form of skin cancer due to its ability to metastasize rapidly to other parts of the body [37].

3.1 Risks and Causes of Cancer

The risk factors and causes of the different types of cancer vary widely, and each type is associated with different genetic predispositions, lifestyle habits, and environmental exposures. Table **??** below summarizes some of the main risk factors and causes associated with the cancer types mentioned.

3.2 Skin Cancer in Brazil

According to the article by Luz et al. (2023) [26], a detailed analysis of variations in the number of melanoma cases and other skin diseases in Brazil is presented. The research covers a seven-year period and examines regional differences in the incidence of these conditions. Using epidemiological data, the study aims to identify trends and patterns that may shed light on the distribution of these conditions across different regions of the country. The introduction highlights the relevance of skin cancers as a significant public health concern and the need for continuous monitoring to formulate effective health policies.

The results presented in the article show substantial variations among Brazilian regions, suggesting that environmental and socioeconomic factors play a crucial role in the incidence of these diseases. The study's conclusion emphasizes the importance of considering these variables when planning public health interventions and awareness campaigns. Furthermore, the work suggests the need for further research to better understand the underlying factors contributing to these variations, proposing strategies to improve the prevention and treatment of skin cancer in Brazil. Thus, the article by Luz et al. (2023) [26] provides a valuable foundation for future investigations and for the development of more effective public health policies to combat skin malignancies.

3.3 Melanoma Variants and Clinical Characteristics

Melanoma is the most aggressive cutaneous neoplasm, often exhibiting rapid progression and metastatic spread. Cabrera and Recule [8] reviewed the clinical and histological characteristics of unusual clinical presentations of melanoma, emphasizing the importance of dermatoscopic findings. Magro et al. [27] discussed uncommon variants of melanoma, highlighting the need for accurate recognition and diagnosis of these variants to improve patient outcomes.

Cancer Type	**Risk Factors**	**Causes**
Carcinoma	Advanced age, sun exposure, smoking, viral infections (HPV)	Genetic mutation, exposure to carcinogens, UV radiation, chronic inflammation
Sarcoma	Exposure to chemicals, radiotherapy, family history	Genetic alterations, ionizing radiation, hereditary genetic syndromes
Lymphoma	Immunosuppression, viral infections (EBV, HIV), advanced age	DNA alterations in lymphocytes, chronic infections, autoimmune diseases
Leukemia	Radiation exposure, chemicals, smoking, Down syndrome	Mutation in hematopoietic stem cells, genetic factors, exposure to benzene
Melanoma	Sun exposure, history of sunburn, fair skin, multiple nevi	DNA damage caused by UV radiation [20], genetic predisposition, exposure to chemicals [22]

Tabela 3.1: Risk factors and causes of different types of skin cancer

An analysis of the 2018 global cancer statistics, as discussed by Bray et al. [5], provides crucial data on the global burden of skin cancer. The study reveals that skin cancer, excluding melanoma, contributed significantly to the total incidence and mortality figures in 2018, underscoring the ongoing importance of surveillance and preventive intervention. Furthermore, melanoma skin cancer, while not the most common among all skin cancers in Brazil, ranks as one of the most lethal forms, highlighting the need for effective early detection and treatment strategies on a global scale.

The data also highlight the influence of geographic, economic, and lifestyle factors on the incidence and mortality of skin cancer. The substantial geographic variability observed in the study reflects significant disparities in the distribution of skin cancer worldwide, with areas of higher sun exposure and lower economic development often showing higher incidence rates. These findings underscore the need for tailored and culturally sensitive approaches to skin cancer management and prevention, aiming to mitigate its global burden and improve public health outcomes.

Black individuals have a lower incidence of melanoma due to the natural protection provided by melanin against harmful UV rays, with a significantly lower rate of disease development compared to non-Hispanic White individuals. However, when diagnosed with melanoma, Black patients face substantial challenges in long-term survival. According to a study by the United States Centers for Disease Control and Prevention, the five-year survival rate for Black patients was only 66%, compared to 90% for non-Hispanic White patients during the period from 2011 to 2015.

Early diagnosis is crucial, but for Black patients, melanoma often presents differently, such as acral lentiginous melanoma (ALM), located on areas like the palms of the hands, soles of the feet, or under the nails, where sun exposure is minimal. This contrasts with the typical presentation in lighter skin, where melanoma is more often found in sun-exposed areas. Research indicates that dermatologists may face challenges in accurately identifying skin conditions in darker skin tones, being less likely to diagnose malignancies and more likely to recommend biopsies for benign conditions compared to lighter skin. These discrepancies highlight the critical need to raise awareness and improve clinical competence to enhance health outcomes for patients of different ethnicities and skin tones.

The lack of representation of diverse skin tones in medical education materials and diagnostic tools has significant negative implications for public health, particularly regarding skin cancer. The scarcity of dark skin images in dermatology atlases compromises the ability to correctly identify skin lesions in Black and Brown individuals, resulting in delays in diagnosis and treatment. Such delays can have serious consequences, such as worsening of health conditions and reduced chances of successful treatment, especially in cases like melanoma, which require early diagnosis for better prognosis [19].

Non-Hispanic Black Americans have a lower incidence of melanoma than non-Hispanic White Americans. However, they are often diagnosed with melanoma at a more advanced stage and have lower survival rates [9]. This is likely due to a combination of factors, including lower awareness of melanoma risk factors, reduced access to healthcare, and differences in melanoma biology in non-Hispanic Black Americans. The most common type of melanoma in non-Hispanic Black Americans is acral lentiginous melanoma, which occurs on the palms, soles, or nail beds. This type of melanoma is often difficult to diagnose because it can be mistaken for a benign nevus. Better education of healthcare professionals and the public about melanoma in non-Hispanic Black Americans is needed to improve early diagnosis and survival rates.

The lack of representation of skin diversity [2] in medical education materials and diagnostic tools negatively impacts public health, especially in the context of skin cancer. The low representation of dark skin tones in dermatology atlases, for example, makes it harder to recognize lesions in Black and Brown individuals, leading to delays in diagnosis and treatment [19], with serious consequences for patients' health and well-being.

Studies show that skin pigmentation can mask the warning signs of skin cancer [19], such as asymmetries, irregular borders, color variations, and lesion growth. The absence of images with darker skin tones in manuals and diagnostic tools hinders the perception

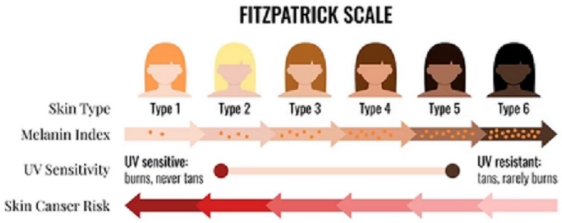


Figura 3.1: FitzPatrick Palete [31]

of these features in Black and Brown patients by healthcare professionals. This racial disparity in access to accurate and timely diagnosis contributes to the lower skin cancer survival rates among the Black population compared to the White population [19].

3.4 Convolutional Neural Networks and Medical Image Classification

Convolutional neural networks (CNNs), such as VGG19, have shown great potential in the task of medical image classification due to their ability to extract relevant features from complex images. In the work of Simonyan and Zisserman [36], the architecture of VGG19 is detailed, which is widely used in computer vision applications for large image datasets. Esteva et al. [13] demonstrated that skin cancer classification with deep neural networks can achieve accuracy levels comparable to those of dermatologists.

3.5 Fitzpatrick Scale and Skin Lesion Detection

The Fitzpatrick Scale established a standard for classifying skin tone (similar to a color palette).

This definition is crucial for understanding how different skin tones may influence the detection and classification of skin lesions. Studies such as that of Goon et al. [15] question whether the scale provides a false sense of security for skin types IV–VI, highlighting the need for more inclusive models. Kinyanjui et al. [24] investigated the estimation of skin tone and its effects on classification performance in dermatology datasets, emphasizing the importance of including diverse skin tones in studies.

3.6 Legal Approaches and Challenges in Dermatology

Troxel [38] and Moshell et al. [30] discussed medicolegal aspects related to errors in pathology and professional liability claims against dermatologists, respectively, highlighting the importance of accurate diagnoses and proper documentation to avoid litigation. Rayess et al. [34] analyzed medical malpractice lawsuits related to melanoma, suggesting the need for more comprehensive biopsies to prevent misdiagnoses.

3.7 Applications of Neural Networks in Dark Skin

Barros et al. [3] evaluated the generalization of deep neural network-based models for skin lesions in dark skin. They found that models trained predominantly on light skin tones do not perform well on dark skin, highlighting the need to create diverse datasets and specialized models to ensure inclusive diagnoses.

3.8 Research Standards and Guidelines

Xavier [42] provides detailed guidelines on how to use ABNT standards for citations, which is essential for the standardization and quality of academic work. Da Silva and Menezes [10] offer a comprehensive methodology for research and dissertation preparation, while Keele [23] and Kitchenham et al. [25] discuss conducting systematic literature reviews in software engineering, emphasizing the importance of rigorous methodologies to ensure the reliability of results.

4 Methodology

This section describes the methods and materials used for classifying skin lesions using the VGG19 convolutional neural network architecture with the ISIC 2020 dataset.

4.1 ISIC 2020 Dataset

The ISIC 2020 dataset, provided by the International Skin Imaging Collaboration (ISIC), is widely used for skin lesion classification. It contains images of seven different types of skin lesions, including melanoma, nevus, lentigo, lichenoid keratosis, seborrheic keratosis, solar lentigo, and café-au-lait macule. This dataset is known for presenting significant challenges, such as class imbalance and underrepresentation of dark skin images.

4.1.1 Dataset

For this study, we used the ISIC 2020 dataset, which consists of images of seven different types of skin lesions, including:

Melanoma

Description: An aggressive type of skin cancer that develops from melanocytes, the cells that produce pigment in the skin.

Nevus

Description: Also known as a "mole"or "birthmark,"it is a benign pigmented lesion on the skin.

Lentigo NOS

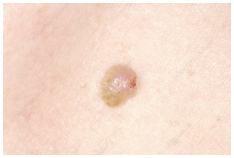
Description: A type of dark spot on the skin, often caused by prolonged sun exposure.

Lichenoid Keratosis

Description: A chronic inflammatory skin lesion that can become scaly and rough over time.

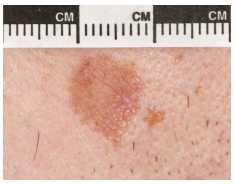
Seborrheic Keratosis

Description: Also known as "seborrheic wart,"it is a common benign lesion that may appear as a dark brown or black patch on the skin.



Solar Lentigo

Description: A benign pigmented lesion caused by chronic sun exposure, common in sun-exposed areas.



Café-au-lait Macule

Description: A skin spot that is lighter than the surrounding skin and may vary in size. It is generally benign.

4.2 Skin Color Classification

We used the Fitzpatrick Scale [14] to classify the skin color in the images. This scale categorizes skin types into six groups based on sun response and pigmentation. To determine the skin tone in the images, we applied computer vision techniques, including conversion to the CIELAB color space and calculation of the Individual Typology Angle (ITA).

$$\text{ITA} = \arctan\left(\frac{L-50}{B}\right) \times \left(\frac{180}{\pi}\right)$$

The calculated ITA° values were shown to reflect an individual's sensitivity to UV radiation and, consequently, the skin's susceptibility to sun exposure and skin damage, such as pigmentation changes, cancer, and aging. From a practical standpoint, this translates into determining the most effective type of protection and/or treatments for specific skin types.

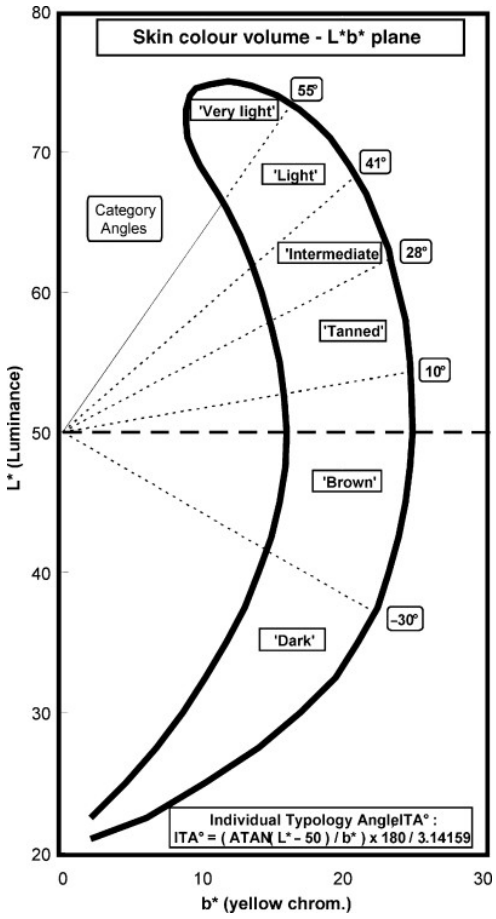


Figura 4.1: Skin color volume projected on the L*b* plane of the Lab color space (Image source: [4]).

Figura 4.2: Fitzpatrick Scale

4.3 Data Preprocessing

The images were resized to 224x224 pixels and normalized. Data augmentation techniques were applied to mitigate class imbalance, including rotation, shifting, shearing, zoom, and horizontal flipping.

4.4 Data Augmentation

Data augmentation was performed using the `ImageDataGenerator` class from the Keras library. The following transformations were applied:
Rotation up to 20 degrees
Horizontal and vertical shift up to 20%
Shearing up to 20%
Zoom up to 20%
Horizontal flip

4.5 Data Splitting

The dataset was split into training and validation sets using the `train_test_split` function from the Scikit-learn library, with 80% of the data for training and 20% for validation. This split ensures an objective evaluation of the model's performance.

4.6 Model Training

We used a pre-trained VGG19 model, with the final layers adjusted for the classification of seven categories of skin lesions. The model was trained using the cross-entropy loss function and the Adam optimizer. Training was performed over 20 epochs with a learning rate adjustable via a callback.

4.7 Model Evaluation

The model's performance was evaluated using metrics such as accuracy, precision, recall, and F1-score. The confusion matrix was used to visualize the model's performance in correctly or incorrectly classifying skin lesion images.

5 Performance Metrics

In the context of machine learning and classification, performance metrics are essential for evaluating a model's effectiveness. The most common metrics include **accuracy**, **precision**, **recall**, and **F1-score**. Each provides a different perspective on the model's performance and is useful in different scenarios.

5.1 Accuracy

Accuracy is one of the simplest and most widely used metrics. It measures the proportion of correct predictions relative to the total number of evaluated samples. The formula for calculating accuracy is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where:

- TP (True Positives) are the correctly predicted positive cases,
- TN (True Negatives) are the correctly predicted negative cases,
- FP (False Positives) are the incorrectly predicted positive cases,
- FN (False Negatives) are the incorrectly predicted negative cases.

Accuracy is an appropriate metric when the classes are balanced, meaning there is a similar number of samples in each class.

5.2 Precision

Precision measures the proportion of correctly predicted positive cases out of all positive predictions made by the model. It is particularly useful when the cost of false positives

is high. The formula for calculating precision is:

$$Precision = \frac{TP}{TP + FP}$$

A high precision indicates that, among the samples classified as positive, many are indeed positive.

5.3 Recall

Recall (or *sensitivity*) measures the proportion of true positives relative to the total number of actual positive samples. It is important in scenarios where correctly identifying positive cases is crucial, such as in disease diagnosis. The formula for calculating recall is:

$$Recall = \frac{TP}{TP + FN}$$

A high recall indicates that the model is capable of identifying most positive samples, even if it makes some errors by classifying negative samples as positive.

5.4 F1-Score

The **F1-score** is the harmonic mean between precision and recall. It is useful when a balance between these two metrics is necessary, especially in classification problems with imbalanced classes. The formula for calculating the F1-score is:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The F1-score is a balanced metric that combines the model's ability to avoid false positives (precision) with its ability to detect most positives (recall).

5.5 Considerations on the Use of Metrics

Depending on the application context, different metrics may be more relevant. Accuracy can be misleading when the classes are imbalanced, whereas the F1-score offers a more

balanced view by considering both precision and recall. The choice of metric should reflect the nature of the problem and the costs associated with false positives and false negatives.

5.5.1 Class Imbalance

One of the significant challenges when using the ISIC 2020 dataset is the imbalance between the different skin lesion categories. Some classes, as shown in Table 5.2, may have fewer available examples compared to other benign lesions. This imbalance can affect the model's ability to effectively learn to distinguish between minority classes, resulting in lower overall accuracy for these underrepresented categories.

Class	**Number of Images**
0	32542
1	584
Total Result	**33126**

Tabela 5.1: Image Distribution by Class

5.5.2 Classifying Skin Color According to the Fitzpatrick Scale

Algoritmo 1 Calculation of the Individual Typology Angle (ITA) of an Image

Input: An image in the LAB color space. **Output:** The ITA value for the image.

Step 1: Extraction of Luminance (L) Obtain the luminance (L) values of the image, ranging from 0 (black) to 100 (white). Ignore pixels with a luminance value equal to 0, as they are completely black.

Step 2: Calculation of Luminance Mean and Standard Deviation

- Calculate the mean of the remaining luminance values.
- Calculate the standard deviation, which indicates how much the luminance values vary around the mean.

Step 3: Filtering Luminance Values

- Keep only the luminance values that are within ±1 standard deviation from the mean.

Step 4: Extraction and Filtering of the b Component (Yellow–Blue)

- Obtain the values of the b component of the image, which represent the yellow–blue hue.
- Calculate the mean and standard deviation of these values.
- Keep only the b values that are within ±1 standard deviation from the mean.

Step 5: ITA Calculation

- Calculate the ITA using the relationship between the adjusted luminance mean and the mean of the b component:

$$ITA = \arctan\left(\frac{\text{Mean Luminance} - 50}{\text{Mean b Component}}\right) \times \frac{180}{\pi}$$

Step 6: Return ITA

- Return the ITA value, which will be used to classify the skin color of the image.

In addition, there is a notable underrepresentation of individuals with darker skin

tones in the dataset. Only about 10% of the images are of individuals with skin phototypes V and VI. This is problematic because models trained with similar imbalances may not generalize well to populations with different skin tones. As a result, automated diagnoses may be less accurate for these groups, exacerbating inequalities already present in traditional medical practice.

Recent research highlights the importance of addressing these challenges to improve the fairness and accuracy of automated skin cancer diagnostic systems. Strategies to mitigate class imbalance include data augmentation techniques specifically for minority classes, adjusting class weights during model training, and advanced regularization methods to prevent the model from focusing solely on majority classes.

Diagnosis	**Number of Images**
atypical melanocytic proliferation	1
cafe-au-lait macule	1
lentigo NOS	44
lichenoid keratosis	37
melanoma	584
nevus	5193
seborrheic keratosis	135
solar lentigo	7
unknown	27124*
Total Result	**33126**

Tabela 5.2: Image Distribution by Diagnosis

5.5.3 Image Selection

The images were selected according to skin tone. This information did not exist in the original dataset; therefore, it was necessary to use an algorithm to identify the skin color and classify the images according to the Fitzpatrick Scale. Initially, only scales V and VI were selected, but for comparison purposes, we also used classes I and II (light skin), III and IV (intermediate skin), and V and VI (dark skin). Another filter was applied so that the pathologies were different from "unknown,"even if the indicator was marked with 0 or 1 for the presence of melanoma. In this way, only **6,072** images were evaluated during the execution of the algorithm tests and validations.

In our tests, we evaluated the algorithm by filtering images by target (0/1), where 1

represents melanoma and 0 represents other types of diagnoses, as shown in Table 5.1, and diagnoses different from "unknown"(unknown diagnosis).

Tabela 5.3: Number of Selected Images by Class and Skin Type

Category	**Class 1 and 2**	**Class 3 and 4**	**Class 5 and 6**	**Classes 1 to 6**
Number of Images	3684	1094	1292	6072
20% for training	737	219	258	1214

5.5.4 Preprocessing

The images were resized to 224x224 pixels, normalized, and divided into training and test sets. Different data augmentation techniques were used to mitigate class imbalance.

5.5.5 Data Augmentation

Data augmentation is an essential technique in training machine learning models, especially in problems where the amount of available data is limited. By applying random transformations to training images, it is possible to generate new samples from the existing ones, thus increasing data variability and helping to prevent overfitting.

In this work, we used the `ImageDataGenerator` class from the Keras library to perform data augmentation. The augmentation algorithm is configured with a series of parameters that define the transformations applied to the images:

- **rotation_range=20**: Randomly rotates images by up to 20 degrees.
- **width_shift_range=0.2**: Horizontally shifts images by up to 20% of the total image width.
- **height_shift_range=0.2**: Vertically shifts images by up to 20% of the total image height.
- **shear_range=0.2**: Applies a shear transformation to images by up to 20%.
- **zoom_range=0.2**: Randomly zooms into images by up to 20%.
- **horizontal_flip=True**: Horizontally flips images.
- **fill_mode='nearest'**: Fills empty pixels generated by transformations with the values of the nearest pixels.

These transformations increase the robustness of the model, making it more capable of generalizing to new, unseen data during training. Data augmentation is particularly effective in medical image classification tasks, where variations in the input images can be crucial for correctly identifying different classes of skin lesions.

By applying these data augmentation techniques, we aim to create a more diverse and representative training dataset, contributing to improved performance of the VGG19 model in skin lesion classification.

5.5.6 Splitting the Data into Training and Validation Sets

Splitting the data into training and validation sets is a fundamental step in the machine learning model development process. This step ensures that the model is trained on a subset of the data and evaluated on a separate subset, allowing for an objective assessment of its performance on unseen data.

In this work, we used the `train_test_split` function from the Scikit-learn library to perform the data split. This function divides the input data `X` and the corresponding labels `y` into training and validation sets, according to the proportion specified by the `test_size` parameter. The code used for this split is presented below: Splitting the data into training and validation sets is crucial to ensure that the trained model can generalize well to new data. By reserving part of the data for validation, we can monitor the model's performance during training and adjust the hyperparameters as needed to avoid overfitting. This approach is especially important in medical image classification tasks, where model accuracy has a direct impact on diagnostic quality.

5.5.7 Model Training

Our solution is based on a pre-trained VGG19 model [36], with the final layers adjusted for the classification of seven categories. The categories were grouped in a binary format: positive (when melanoma is present) or negative (any of the other six categories). The cross-entropy loss function was used, and the Adam optimizer was employed. Figure 5.1 illustrates the training process.

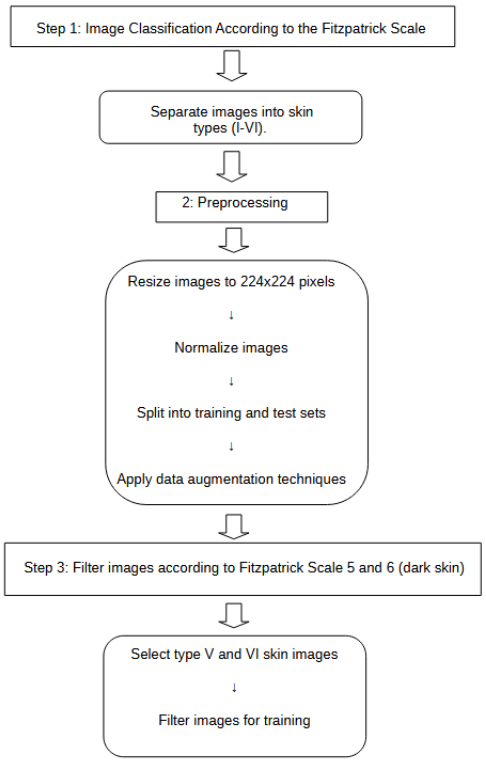


Figura 5.1: Training process diagram

5.6 Algorithm

Import necessary libraries:
Import libraries such as pandas, os, numpy, cv2, tensorflow, and other relevant packages.

Mount Google Drive:
Mount Google Drive for access to stored data.

Load tabular ISIC 2020 data:

Load the CSV file containing ISIC 2020 training data.

Filter data by removing entries with unknown anatomical location or diagnosis.

Load skin color data:
Load CSV file containing skin color information.

Remove rows with empty values in the ITA column.

Filter only images with skin colors within levels 1 to 6 of the Fitzpatrick scale.

Merge data:
Merge diagnosis data with skin color–filtered data.

Filter data to include only entries with target labels 0 or 1.

Data filtering and sampling:
Define the maximum number of images per diagnosis type.
Filter and sample data to ensure a balanced number of images per diagnosis type.
Filter and sample images with target label equal to 1.
Combine filtered data to obtain a final balanced dataset.
Remove unnecessary columns and ensure that all records have non-null values in relevant columns.

Function to load and preprocess images:
Define a function to load and preprocess images using VGG19.
Resize images, convert to RGB format, and apply VGG19-specific preprocessing.

Load images and labels:
Define a function to load images and their respective labels.
Load images from the specified directory and apply preprocessing.

Split data into training and validation sets:
Split data into training and validation sets using an 80% training and 20% validation ratio.

Load pre-trained VGG19 model:
Load pre-trained VGG19 with ImageNet weights, excluding the top layer.

Freeze VGG19 convolutional layers:
Freeze all convolutional layers of the VGG19 model to prevent them from being trained.

Add additional layers:
Add extra layers to the VGG19 model for the specific binary classification task:
Flatten layer to flatten VGG19 output.
Dense layer with 512 neurons and ReLU activation.
Dropout layer with a rate of 0.5.
Dense layer with 256 neurons and ReLU activation.
Dropout layer with a rate of 0.5.
Dense layer with 1 neuron and sigmoid activation.

Compile the model:
Compile the model using the Adam optimizer, binary cross-entropy loss function, and accuracy, AUC, precision, and recall metrics.

Callback to reduce learning rate:
Define a callback to reduce the learning rate when validation loss stops improving.

Train the model:
Train the model on the training data, validating it with the validation data for 20 epochs.

Save model and history:
Save the trained model and training history to Google Drive.

Make predictions on the validation set:
Make predictions on the validation set using the trained model.

Convert predictions to binary labels (0 or 1).

Save labels and predictions:
Save true labels and predictions to CSV files in Google Drive.

Confusion matrix and classification report:
Compute and print the confusion matrix and classification report to evaluate model performance.

6 Results

6.1 Performance Metrics

Performance metrics such as accuracy, precision, recall, and F1-score are crucial in evaluating models used in medical diagnostics. Understanding these metrics accurately is essential due to the direct impact that false positives and false negatives can have on diagnostic outcomes [18].

False positives occur when the model incorrectly identifies a case as positive, indicating the presence of a condition when it is actually absent. This can lead to unnecessary medical procedures, patient anxiety, and waste of medical resources. For example, in cancer diagnosis, a false positive may result in unjustified invasive biopsies or aggressive treatments.

On the other hand, false negatives occur when the model fails to correctly identify a positive case, incorrectly indicating that the condition is absent. This can result in delayed treatment, allowing the disease to progress without proper intervention. In cancer diagnosis, a false negative can lead to a late diagnosis, significantly reducing the chances of effective and potentially life-saving treatment.

Therefore, understanding and minimizing false positives and false negatives is crucial to ensuring the accuracy of automated medical diagnoses, providing reliable results that support appropriate clinical decisions and improve patient care.

Metric	Description
True Positive (TP)	Positive cases correctly identified.
False Positive (FP)	Negative cases incorrectly identified as positive.
True Negative (TN)	Negative cases correctly identified.
False Negative (FN)	Positive cases incorrectly identified as negative.

Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6.1}$$

Accuracy is a fundamental performance measure in classification models, used to evaluate the proportion of correct predictions relative to the total predictions made by the model. Simply put, accuracy measures the overall correctness of the model in classifying samples.

Precision:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6.2}$$

Precision is a crucial metric for evaluating the performance of a classification model, especially in contexts where the cost of false positives is high. Precision measures the proportion of true positives relative to the total positive predictions made by the model.

Recall:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6.3}$$

Recall, also known as sensitivity or true positive rate, is an essential metric for evaluating classification model performance, particularly in situations where identifying all positive cases is critical. Recall measures the proportion of true positives relative to all actual positive samples.

F1-Score:

$$\text{F1-Score} = 2 \times \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right) \tag{6.4}$$

The F1-score is a metric that combines precision and recall into a single measure, providing a balanced view of a classification model's performance. It is particularly useful when there is an imbalance between data classes, which is common in many real-world

problems.

These metrics are crucial for understanding and evaluating the effectiveness of classification models across various applications.

6.1.1 Confusion Matrix

The confusion matrix chart is a powerful tool for quickly visualizing the number of instances correctly or incorrectly classified by the model.

Below is an example of a confusion matrix resulting from a classification model: A total of 540 images were evaluated, with 20% used for testing, totaling 108 images.

For an example classification of 108 images:

	Class 0	**Class 1**
Class 0	37 (TN)	9 (FP)
Class 1	12 (FN)	50 (TP)

Tabela 6.1: Confusion Matrix, with a total of 108 images

- **37 (True Negative - TN)**: The model correctly classified 37 instances of class 0 (benign lesions).
- **9 (False Positive - FP)**: The model incorrectly classified 9 instances of class 0 as class 1 (malignant lesions).
- **12 (False Negative - FN)**: The model incorrectly classified 12 instances of class 1 as class 0.
- **50 (True Positive - TP)**: The model correctly classified 50 instances of class 1 (malignant lesions).

These values help to understand the model's performance, allowing us to identify not only the number of correct predictions but also the types of errors made.

6.2 Algorithm Evaluation Results

This section presents the evaluation results of using a CNN with VGG19 for feature extraction from dermoscopic image patches, analyzing the data according to Fitzpatrick scales 1 and 2, 3 and 4, and 5 and 6.

6.2.1 All Classes (1 to 6)

The model was trained with 6,072 images, covering all classes, so that the results could be evaluated without separation by skin tone. Validation was performed with 1,215 images.

- Training Accuracy: 0.9555
- Validation Accuracy: 0.9350
- Training Loss: 0.0852
- Validation Loss: 0.2143
- Training AUC: 0.9897
- Validation AUC: 0.9616
- Training Precision: 0.9772
- Validation Precision: 0.9162
- Training Recall: 0.7618
- Validation Recall: 0.7353

Class	**Precision**	**Recall**	**F1-Score**	**Support**
0	0.94	0.98	0.96	977
1	0.92	0.74	0.82	238
Average Accuracy	0.93			
Macro Average	0.93	0.86	0.89	1215
Weighted Average	0.93	0.93	0.93	1215

Tabela 6.2: Classification Report for the Model

Class	**Number of Images**
0	961
1	16
2	63
3	175

Tabela 6.3: Confusion Matrix

Table 6.2 shows the classification report for the model, while Table 6.3 presents the corresponding confusion matrix.

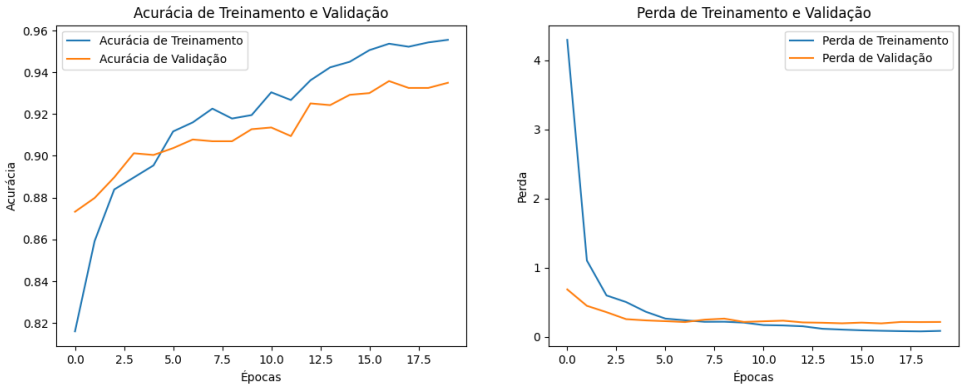


Figura 6.1: Results of CNN + VGG19 feature extraction for classes 1 to 6

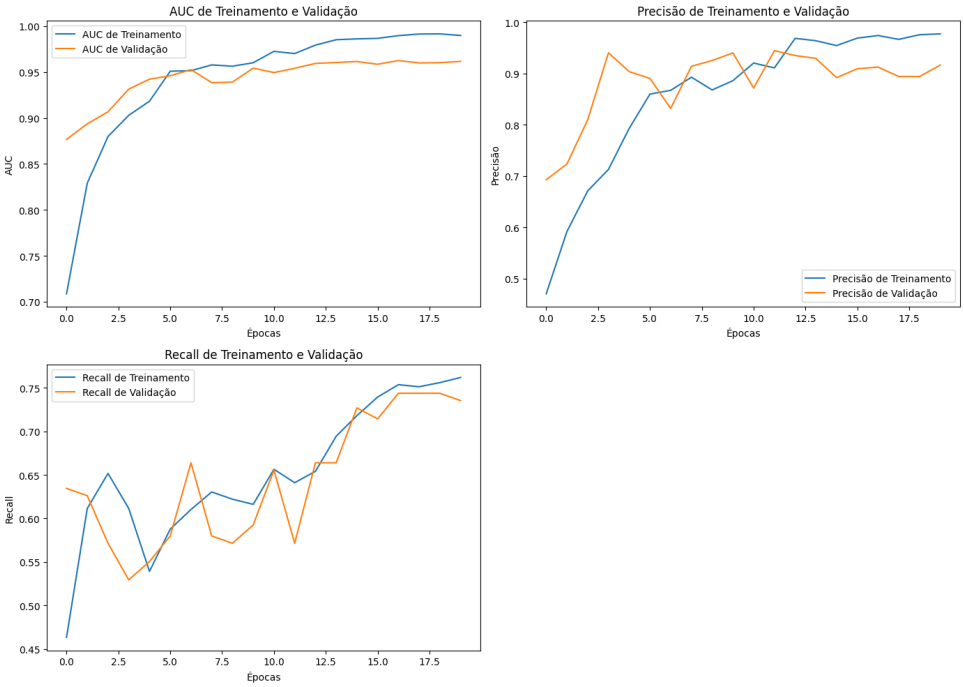


Figura 6.2: AUC for classes 1 to 6

6.2.2 Classes 1 and 2

Classes 1 and 2 represent lighter skin tones and have more samples in the dataset. A total of 3,684 images were filtered, with 20% representing 737 images.

- Training Accuracy: 0.95
- Validation Accuracy: 0.93

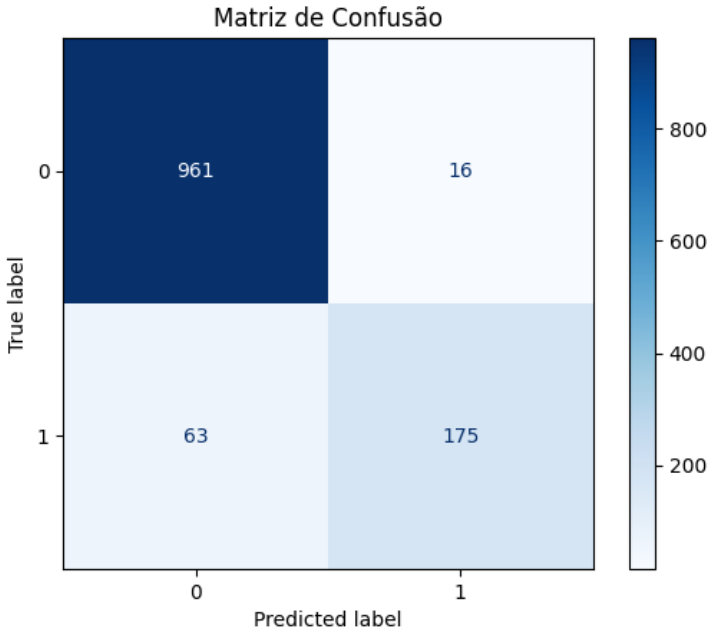


Figura 6.3: Confusion Matrix for classes 1 to 6

- Training Loss: 0.08
- Validation Loss: 0.28
- Training AUC: 0.98
- Validation AUC: 0.93
- Training Precision: 0.97
- Validation Precision: 0.86
- Training Recall: 0.74
- Validation Recall: 0.74

6.2.3 Classes 3 and 4

Classes 3 and 4 represent intermediate skin tones in the dataset. A total of 1,095 images were selected, with 219 representing the 20%.

- Training Accuracy: 0.99
- Validation Accuracy: 0.96

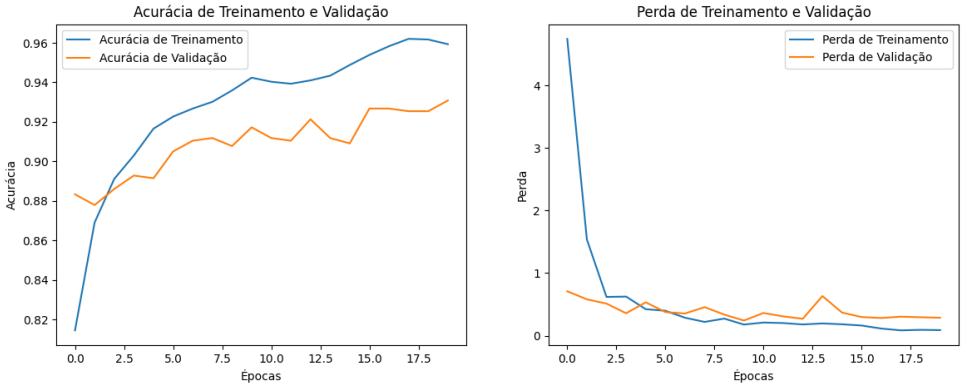


Figura 6.4: Results of CNN + VGG19 feature extraction for classes 1 and 2

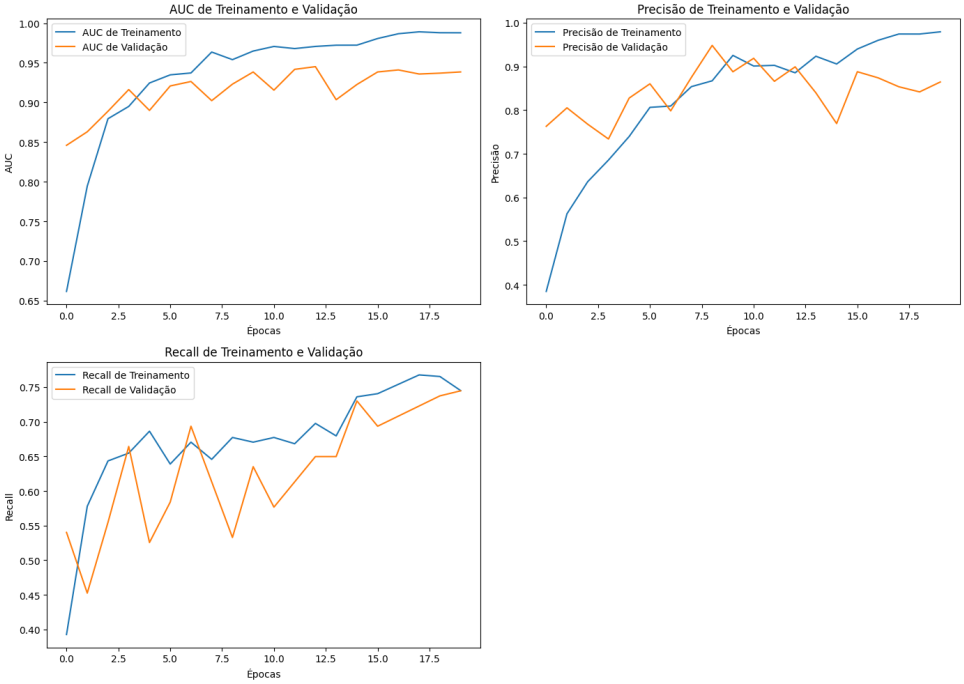


Figura 6.5: AUC for classes 1 and 2

- Training Loss: 0.03
- Validation Loss: 0.40
- Training AUC: 0.99
- Validation AUC: 0.97
- Training Precision: 0.97
- Validation Precision: 0.94
- Training Recall: 0.98

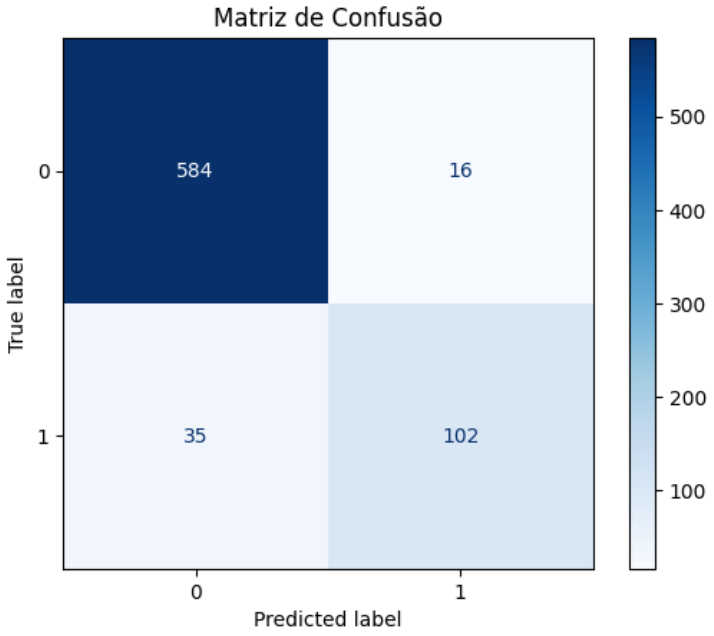


Figura 6.6: Confusion Matrix for classes 1 and 2

- Validation Recall: 0.92

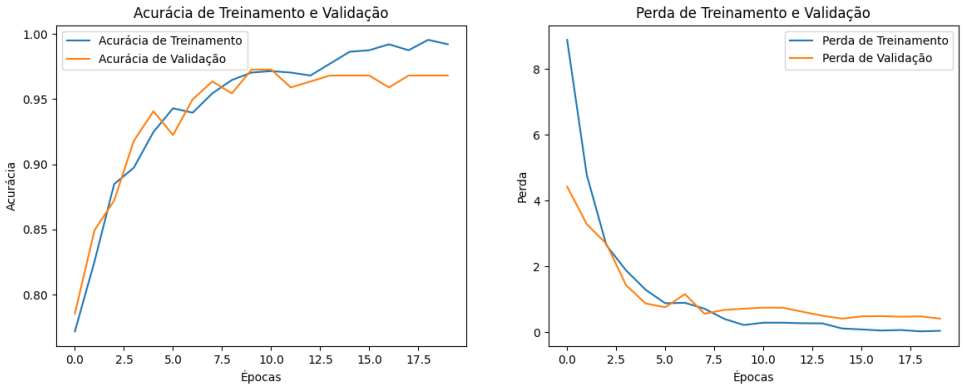


Figura 6.7: Results of CNN + VGG19 feature extraction for classes 3 and 4

6.2.4 Classes 5 and 6

Classes 5 and 6 represent darker skin tones and have fewer samples in the dataset. A total of 1,292 images from classes 5 and 6 were analyzed, with 20% (259 images) used for validation.

- Training Accuracy: 0.99
- Validation Accuracy: 0.98

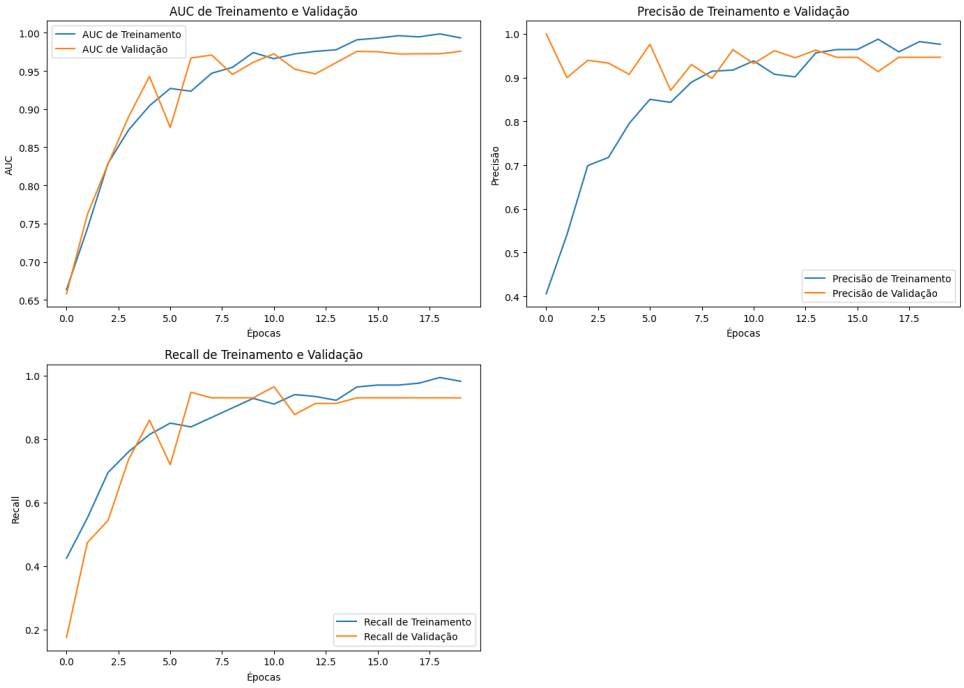


Figura 6.8: AUC for classes 3 and 4

- Training Loss: 0.05
- Validation Loss: 0.43
- Training AUC: 0.99
- Validation AUC: 0.99
- Training Precision: 0.99
- Validation Precision: 0.92
- Training Recall: 0.99
- Validation Recall: 1.0

6.2.5 Class 6

The class corresponding to Fitzpatrick phototype VI was trained in isolation with the aim of assessing the potential tendency to create bias in the opposite direction to that observed in the analysis of phototypes I and II. The results reinforced the need for greater variability in the data and training images to achieve a less biased and more accurate model. This highlights that diversity in the samples used is essential to avoid biases in the classification process.

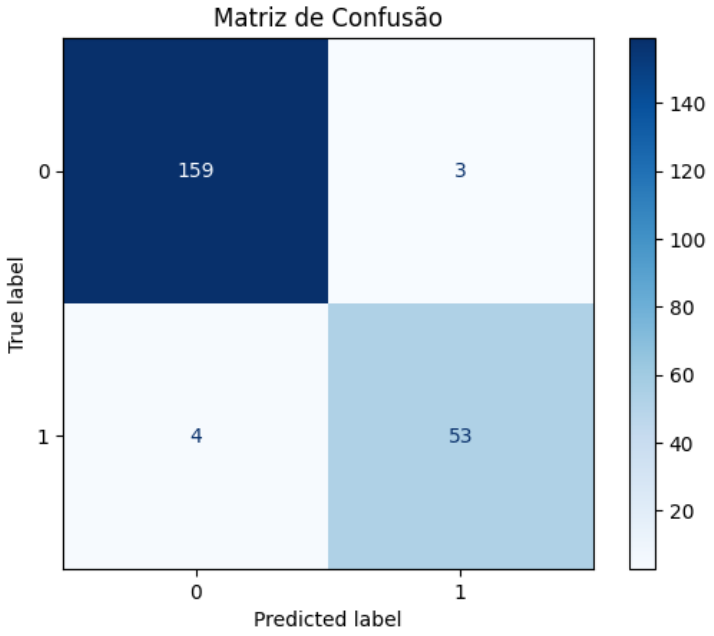


Figura 6.9: Confusion Matrix for classes 3 and 4

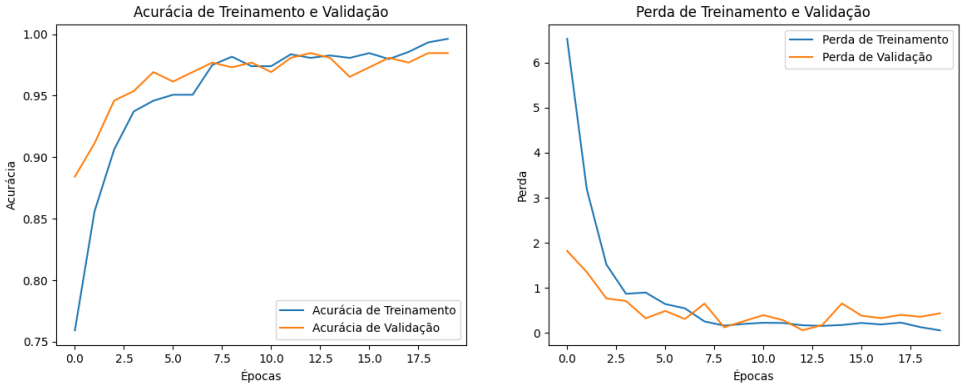


Figura 6.10: Results of CNN + VGG19 feature extraction for classes 5 and 6

6.2.5.1 Analysis of Class 6 (Darker Skin)

The evaluation of the model for Class 6, representing darker skin tones, reveals a performance distinct from other classes.

A total of 960 images filtered for Class 6 were analyzed, with 20% (192 images) used for validation.

During training, the model achieved an accuracy of 99.35% and an AUC of 0.997, indicating excellent fit to the training data. However, in the validation set, accuracy dropped to 92.71% and AUC to 0.895, suggesting potential overfitting. The validation precision was 86.96%, but recall was relatively low at 83.33%, indicating that the model struggles to correctly identify all positive instances (melanoma) within this class. The confusion

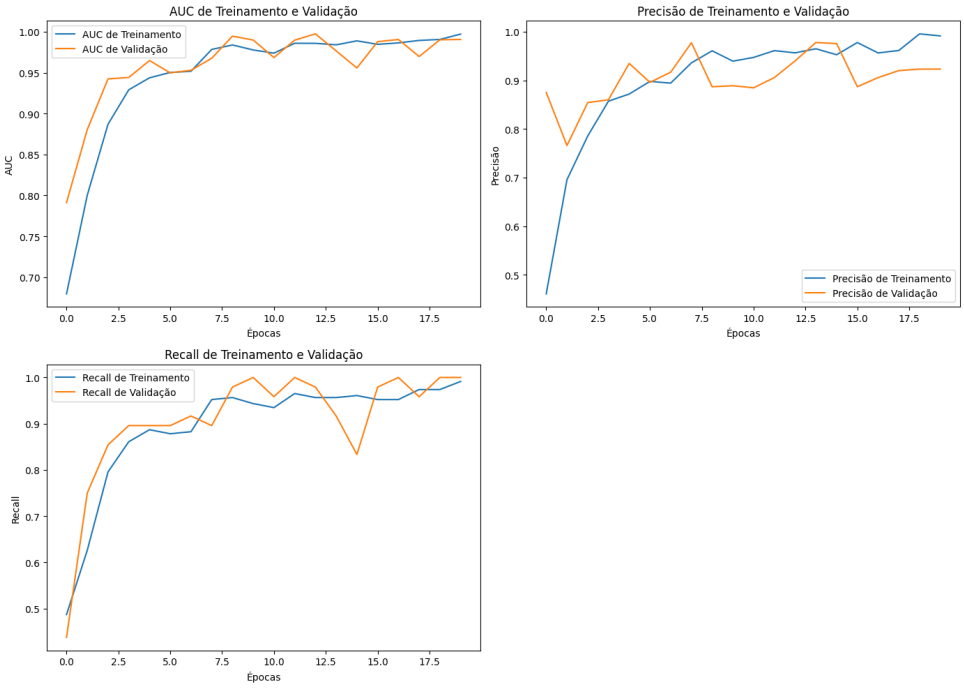


Figura 6.11: AUC for classes 5 and 6

matrix and classification report corroborate these findings, showing that although overall precision is reasonable, the ability to detect melanomas in darker skin still requires improvement. The F1-score for both classes is approximately 0.85, reflecting a moderate balance between precision and recall. These results emphasize the need to improve model generalization for darker skin tones, possibly through data augmentation techniques and hyperparameter tuning to better address the specific characteristics of this class.

- Training Accuracy: 0.9935
- Validation Accuracy: 0.9271
- Training Loss: 0.0781
- Validation Loss: 1.4131
- Training AUC: 0.9974
- Validation AUC: 0.8954
- Training Precision: 0.9735
- Validation Precision: 0.8696
- Training Recall: 1.0

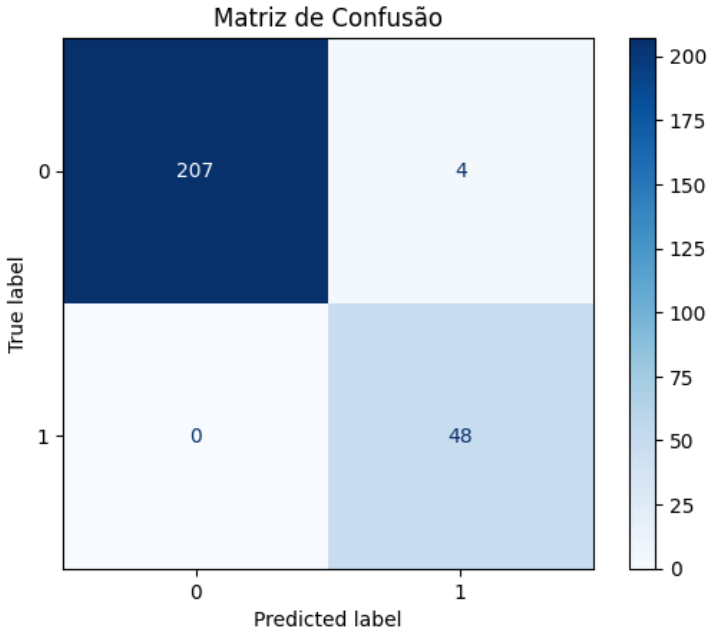


Figura 6.12: Confusion Matrix for classes 5 and 6

- Validation Recall: 0.8333

Class	Precision	Recall	F1-Score	Support
0	0.95	0.96	0.95	144
1	0.87	0.83	0.85	48
Accuracy	0.93			
Macro Average	0.91	0.90	0.90	192
Weighted Average	0.93	0.93	0.93	192

Tabela 6.4: Classification Report for Class 6

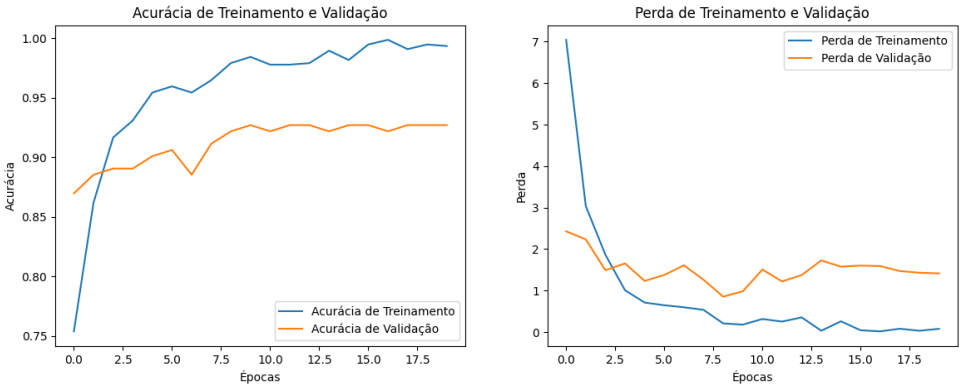


Figura 6.13: Accuracy for Class 6

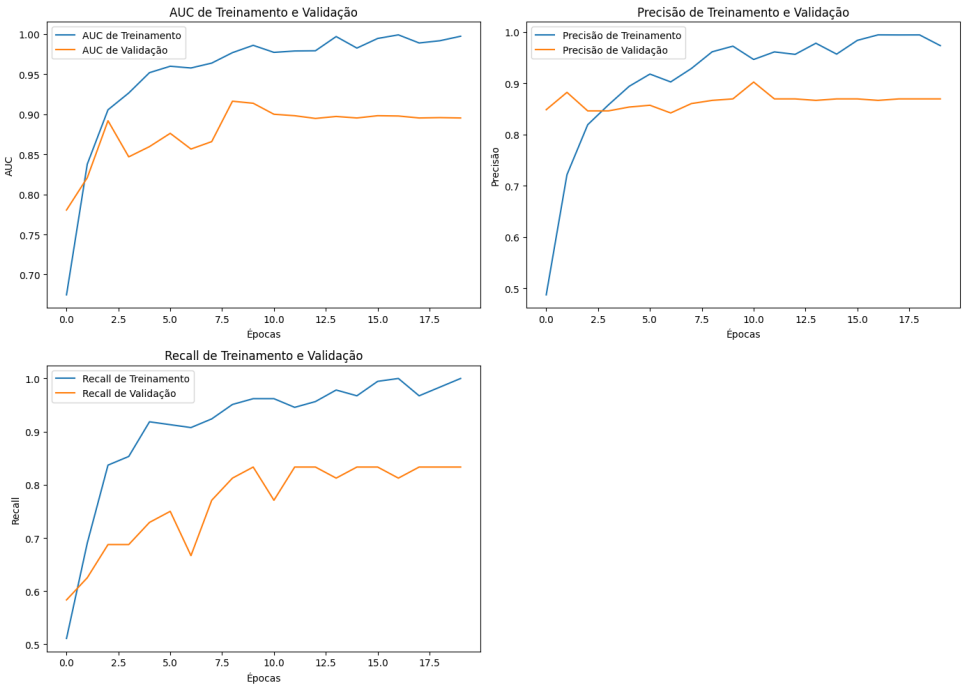


Figura 6.14: AUC for Class 6

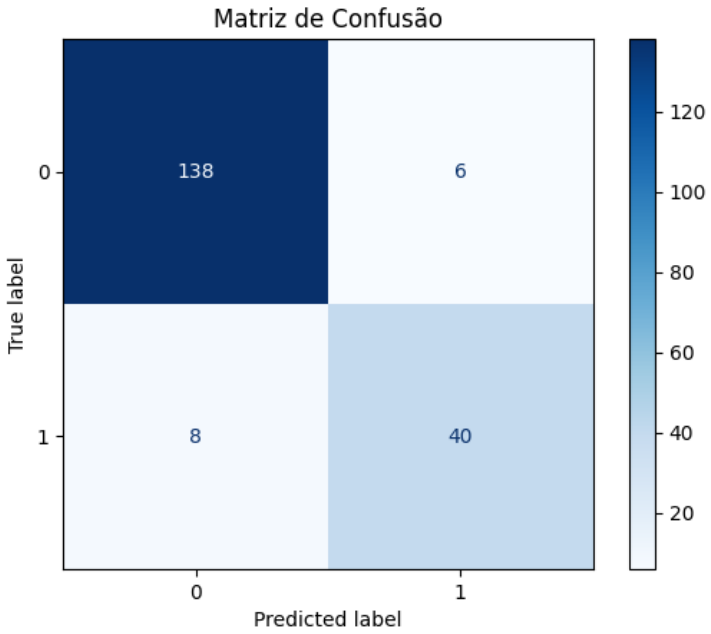


Figura 6.15: Confusion Matrix for Class 6

7 Comparison with ResNet50

The same tests were conducted with the same filtering criteria using the ResNet50 architecture. The results are presented in the table below.

	Tipos 1-2		Tipos 3-4		Tipos 5-6		Todos (1 a 6)	
Número de Imagens	3684		1094		1292		6072	
Métrica	ResNet50	VGG19	ResNet50	VGG19	ResNet50	VGG19	ResNet50	VGG19
Acurácia de Treinamento	0,9528	**0,9593**	0,9783	**0,992**	0,9884	**0,9961**	0,9133	**0,9555**
Acurácia de Validação	0,9254	**0,9308**	**0,9772**	0,968	0,9768	**0,9846**	0,8988	**0,935**
Perda de Treinamento	0,0963	**0,0879**	0,2725	**0,0378**	0,0699	**0,0509**	0,159	**0,0852**
Perda de Validação	0,3648	**0,286**	**0,3043**	0,409	**0,3071**	0,43	0,3757	0,2143
AUC de Treinamento	0,9868	**0,9879**	0,9691	0,9932	0,998	0,9971	0,959	0,9897
AUC de Validação	**0,9439**	0,9385	**0,9872**	0,9756	**0,9922**	0,9905	0,9293	**0,9616**
Precisão de Treinamento	0,975	**0,9792**	0,9453	**0,9762**	0,9886	**0,9913**	**0,9774**	0,9772
Precisão de Validação	**0,8796**	0,8644	0,8194	**0,9462**	0,8889	**0,9231**	0,8859	**0,9162**
Recall de Treinamento	0,7043	**0,7449**	0,9401	**0,982**	0,9609	**0,9913**	0,513	**0,7618**
Recall de Validação	0,6934	**0,7445**	**1**	0,9298	**1**	**1**	0,5546	**0,7353**

Figura 7.1: Comparison ResNet50 vs VGG19

When analyzing the results presented in the comparison table between the VGG19 and ResNet50 architectures, it is evident that VGG19 demonstrates several significant advantages over ResNet50 in various aspects. VGG19 shows notable superiority in crucial metrics such as validation accuracy, achieving an average rate of 0.9350 for all types, compared to 0.8988 for ResNet50. This superior validation performance indicates better generalization capability for VGG19, also reflected in metrics such as validation precision, where VGG19 reaches an average of 0.9162, while ResNet50 presents 0.8859. Furthermore, VGG19 proves more effective in reducing validation loss, with an average value of 0.2143 compared to 0.3757 for ResNet50, suggesting greater performance stability during validation. These results suggest that VGG19, with its deeper architecture and simplified approach to building convolutional networks, can deliver better performance in classification tasks, especially in scenarios where accuracy and generalization are critical. The robustness of VGG19 compared to ResNet50 may be attributed to its ability to extract more discriminative and robust features, resulting in more consistent and reliable overall performance.

8 Discussion

The results obtained from applying the CNN model with VGG19 feature extraction for skin lesion classification were analyzed with a special focus on evaluating Fitzpatrick skin types 1 and 2, 3 and 4, and 5 and 6.

8.1 Model Performance

The model's performance during training was promising, achieving high accuracy, recall, and AUC on the training set for all classes. However, when evaluating the validation set, we observed a significant drop in performance, with accuracy around 75–76%, AUC between 0.78–0.85, precision varying between 70–80%, and recall between 66–76%. This disparity between training and validation sets suggests the occurrence of *overfitting*, where the model becomes overly fitted to the training data and fails to generalize well to new data.

8.2 Impact of Class Imbalance

The class imbalance, particularly notable in darker skin types (V and VI), posed a significant challenge. Only 10% of the samples in the dataset belonged to these skin types, which may lead to training bias and affect the model's ability to accurately detect skin lesions in patients with darker skin tones. Strategies such as data augmentation and class weight adjustments may be necessary to effectively address this issue.

8.3 Performance Metrics

The analysis of performance metrics such as precision, recall, and F1-score revealed important insights into the model's ability to distinguish between positive (melanoma) and

negative (non-melanoma) cases. Precision ranged from 70–80%, indicating the proportion of positive cases correctly identified out of all positive predictions made by the model, while recall ranged from 66–76%, highlighting the model's ability to correctly identify all positive cases present in the dataset.

8.4 Limitations and Future Considerations

It is crucial to acknowledge the limitations of this study. The lack of adequate representation of darker skin types in the dataset may have negatively impacted the model's ability to generalize to these population groups. For future research, it is recommended to explore advanced data augmentation techniques specific to variations in skin tone, as well as to consider using more diverse and representative datasets.

Challenges and public datasets play a crucial role in the development of reliable and effective algorithms for clinical use [35]. Such resources can improve algorithm accuracy and generalization, in addition to promoting transparency and collaboration among researchers. The use of high-quality public datasets ensures that algorithms are trained in realistic and diverse scenarios, increasing their reliability in clinical practice.

To improve model performance and reduce *overfitting*, several strategies can be considered:

- **Increasing the Number of Epochs:** Extending the training epochs can enable the model to better learn data features, although careful monitoring is needed to avoid *overfitting*. Techniques like *EarlyStopping* can help determine the optimal stopping point.
- **Using More Images:** Collecting and using a larger, more diverse dataset can enhance the model's ability to generalize to new data. Adding images from darker Fitzpatrick skin types can better balance the dataset and improve accuracy for these groups.
- **Adding More Layers:** Incorporating additional dense and convolutional layers may help capture more complex patterns in the data. However, this should be done carefully to avoid excessive model complexity, which could lead to *overfitting*.
- **Regularization Techniques:** Applying more robust regularization methods such as *dropout* and L2 regularization can help mitigate the *overfitting* observed in the

model results. These techniques introduce noise during training, forcing the model to generalize better.

- **More Complex Network Architectures:** Exploring more advanced network architectures such as *EfficientNet*, *ResNet*, or *Inception*, which have shown excellent performance in various computer vision tasks. These architectures can provide better feature extraction and generalization capabilities.

- **Data Augmentation:** Implementing data augmentation techniques can help increase the variability of the training dataset, improving model robustness. Techniques such as rotation, translation, zoom, horizontal flip, and brightness adjustment can be applied.

- **Hyperparameter Tuning:** Fine-tuning model hyperparameters, such as learning rate, number of neurons in dense layers, and regularization parameters, can help optimize model performance.

- **Class Balancing:** Using class balancing techniques, such as class weight adjustments or targeted data augmentation for minority classes, can improve the model's ability to learn equitably across all classes.

These approaches provide a clear path for future model improvements, increasing its generalization and performance on unseen data. Exploring these directions may result in more robust and effective algorithms for clinical practice.

8.5 Ethics and Bias in AI Systems

Ethics and bias in artificial intelligence (AI) systems are critical topics in the application of CNNs for medical diagnosis. Studies by [32] show that AI algorithms can perpetuate or even amplify existing biases in training data, leading to disparities in healthcare. Such biases can result in less accurate diagnoses for underrepresented populations, exacerbating existing health inequalities.

8.5.1 Bias Analysis

Bias analysis is essential to ensure that AI systems are fair and equitable. [7] highlight that the lack of diversity in training data can lead to algorithms that work well for some groups

but not for others. In our study, the inclusion of the Fitzpatrick Scale and the detailed analysis of different skin phototypes are important steps in identifying and mitigating these biases.

8.5.2 Bias Mitigation

To mitigate bias, it is important to adopt a holistic approach. This includes collecting diverse data, using data augmentation techniques to simulate greater case variety, and continuously validating models across different population subgroups. Studies by [33] suggest that transparency in development processes and collaboration with affected communities are crucial to creating fairer AI systems.

8.5.3 Future Trends in Computer Vision for Medical Diagnostics

Research in computer vision for medical diagnostics is constantly evolving, with several emerging trends promising to transform clinical practice. Among these trends, the use of more advanced neural network architectures, such as Transformers in computer vision, stands out, showing great potential in image segmentation and classification tasks. Studies by Dosovitskiy et al. (2020) demonstrate that these architectures can outperform traditional CNNs in various performance metrics [11].

8.5.3.1 Personalized Medicine

Another important trend is the integration of computer vision models with genomic and clinical data to promote personalized medicine. Studies by Miotto et al. (2018) suggest that combining image data with other sources of information can lead to more accurate diagnoses and personalized treatments, significantly improving patient outcomes [29].

8.5.4 Augmented and Virtual Reality

The application of augmented reality (AR) and virtual reality (VR) in medical diagnostics is another promising area. Studies by Maresky et al. (2019) show that these technologies can be used for advanced visualization of medical data, allowing physicians to view diagnostic images in 3D and in real-time, potentially improving diagnostic accuracy and efficiency [28].

8.5.5 Synthesis and Potential Impact on Clinical Practice and Future Research

This study demonstrated the effectiveness of using VGG19 in skin lesion classification, highlighting the importance of including ethnic variability in assessing model performance. The integration of the Fitzpatrick Scale allowed for a more detailed analysis of how different skin tones can influence diagnostic accuracy, providing a significant contribution to research in computer vision and medical diagnostics.

8.5.5.1 Clinical Relevance

The clinical relevance of this study is evident in the potential application of the results to improve early melanoma diagnosis, especially in underrepresented populations. Implementing more inclusive models can lead to increased diagnostic accuracy and equity in healthcare, benefiting patients from diverse ethnic backgrounds.

8.5.6 Future Directions

For future research, it is crucial to continue exploring and developing techniques that consider ethnic diversity and data variability. Combining CNNs with other data sources, such as genomic and clinical data, promises significant advances in personalized medicine. Furthermore, the application of emerging technologies, such as augmented and virtual reality, may open new frontiers in medical data visualization and interpretation, further enhancing clinical practice.

9 Conclusion

This study proposes the use of VGG19 to classify skin lesions, including melanoma, highlighting that it is a relatively underexplored approach, as shown in Table 2.1. The ability of VGG19 to extract detailed features from medical images can significantly enhance diagnostic accuracy, benefiting both patients and healthcare professionals.

Another study reveals a strong correlation between the fatal incidence of melanoma in less sunny climates and skin pigmentation, especially among populations with minimal migration history in Europe. Conversely, in the United States, there is an association between melanoma incidence and lower latitudes with higher UV indices, affecting both non-Hispanic white individuals and Hispanic and Black populations.

Although skin cancer is one of the most commonly diagnosed types, it does not rank among the leading causes of mortality according to SEER [1]. This underscores the importance of understanding not only the incidence of the disease but also the factors influencing its prevalence and the challenges faced in prevention and treatment.

These risk factors suggest common mechanisms in the development of melanoma across different ethnic groups [15]. However, it is crucial to note that, according to the Fitzpatrick Scale, the ISIC 2020 dataset was evaluated and found to contain only 10% of samples representing skin tones V and VI (brown and dark). This underrepresentation may result in deficiencies in classifying skin lesions in individuals with darker skin tones. To address this challenge and improve accuracy, machine learning techniques incorporating skin color categorization could significantly enhance results. Future studies could assess outcomes separately by skin color category, exploring how different models and machine learning techniques can be adapted to account for ethnic diversity and improve dermatological lesion classification accuracy.

The integration of models such as VGG19 into clinical decision support systems represents a promising advancement for dermatological practice, enabling early and accurate detection of skin lesions, potentially leading to improved clinical outcomes and progno-

ses for patients. Furthermore, a study conducted by Haenssle et al. [17] compared the diagnostic performance of a convolutional neural network (CNN), concluding that the technology can provide a robust and effective approach to dermatoscopic image analysis, highlighting its potential to significantly enhance diagnostic accuracy and efficiency.

Nevertheless, it is important to note that, according to the Fitzpatrick Scale, darker skin is not adequately balanced within the dataset, potentially leading to underfitting in classification. This imbalance may adversely affect the model's ability to correctly identify skin lesions in patients with darker skin. To address this challenge, future studies should explore robust data augmentation techniques and model adaptation strategies to better capture the ethnic diversity of patients.

In summary, integrating models such as VGG19 into clinical decision support systems represents a promising step forward for dermatological practice, enabling early and precise detection of skin lesions, which may lead to improved clinical outcomes and patient prognoses.

REFERENCIAS

[1] National Cancer Institute (NCI). *SEER StatFacts: Common Cancer Facts*. Accessed on July 11, 2024. 2024. URL: https://seer.cancer.gov/statfacts/.

[2] Association of American Medical Colleges. *Why Are So Many Black Patients Dying from Skin Cancer?* https://www.aamc.org/news/why-are-so-many-black-patients-dying-skin-cancer. Accessed: 2024-07-11.

[3] Luana Barros, Levy Chaves e Sandra Avila. "Assessing the Generalizability of Deep Neural Networks-Based Models for Black Skin Lesions". Em: *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Ed. por Verónica Vasconcelos, Inês Domingues e Simão Paredes. Cham: Springer Nature Switzerland, 2024, pp. 1–14. ISBN: 978-3-031-49249-5.

[4] Sandra Bino et al. "Relationship between skin response to ultraviolet exposure and skin color type". Em: *Pigment cell research / sponsored by the European Society for Pigment Cell Research and the International Pigment Cell Society* 19 (jan. de 2007), pp. 606–14. DOI: 10.1111/j.1600-0749.2006.00338.x.

[5] Freddie Bray et al. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". Em: *CA: A Cancer Journal for Clinicians* 68.6 (2018), pp. 394–424. DOI: 10.3322/caac.21492.

[6] T. J. Brinker et al. "Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task". Em: *European Journal of Cancer* 113 (2019), pp. 47–54. DOI: 10.1016/j.ejca.2019.05.030.

[7] Joy Buolamwini e Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification". Em: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. PMLR. 2018, pp. 77–91.

[8] R. Cabrera e F. Recule. "Unusual Clinical Presentations of Malignant Melanoma: A Review of Clinical and Histologic Features with Special Emphasis on Dermatoscopic Findings". Em: *American Journal of Clinical Dermatology* 19 (nov. de 2018), pp. 15–23. DOI: 10.1007/s40257-018-0373-6.

[9] MaryBeth B Culp e Natasha Buchanan Lunsford. "Melanoma Among Non-Hispanic Black Americans". Em: *Prev Chronic Dis* 16 (2019). PEER REVIEWED, p. 180640. DOI: 10.5888/pcd16.180640.

[10] Edna Lucia Da Silva e Estera Muszkat Menezes. "Metodologia da pesquisa e elaboração de dissertação". Em: *UFSC, Florianópolis, 4a. edição* 123 (2005).

[11] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". Em: *arXiv preprint arXiv:2010.11929* (2020).

[12] Andre Esteva et al. "Dermatologist-level classification of skin cancer with deep neural networks". Em: *Nature* 542.7639 (2017), pp. 115–118. ISSN: 1476-4687. DOI: 10.1038/nature21056. URL: https://doi.org/10.1038/nature21056.

[13] Andre Esteva et al. "Dermatologist-level classification of skin cancer with deep neural networks". Em: *Nature* 542 (2017), pp. 115–118. DOI: 10.1038/nature21056. URL: https://doi.org/10.1038/nature21056.

[14] T.B. Fitzpatrick. "Ultraviolet-induced pigmentary changes: benefits and hazards". Em: *Curr Probl Dermatol* 15 (1986), pp. 25–38. DOI: 10.1159/000412090.

[15] Philip K. Goon e Susan M. Swetter. "Skin cancer epidemiology in populations of color". Em: *Dermatologic Clinics* 39.1 (2021), pp. 1–10.

[16] H. A. Haenssle et al. "Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists". Em: *Annals of Oncology* 29.8 (2018), pp. 1836–1842. ISSN: 0923-7534. DOI: 10.1093/annonc/mdy166. URL: https://doi.org/10.1093/annonc/mdy166.

[17] HA Haenssle et al. "Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists". Em: *Ann Oncol* 29.8 (ago. de 2018), pp. 1836–1842. DOI: 10.1093/annonc/mdy166.

[18] Trevor Hastie, Robert Tibshirani e Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, 2009.

[19] Shelley Hu, Raynald Soza-Vento e David F. Parker. "Analysis of skin cancer risk factors in black patients: a retrospective study". Em: *Dermatology Research and Practice* 2014 (2014), p. 593020. DOI: 10.1155/2014/593020. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4141864/.

[20] M. Ichihashi et al. "UV-induced skin damage". Em: *Toxicology* 189 (2003), pp. 21–39. DOI: 10.1016/S0300-483X(03)00150-1. URL: https://doi.org/10.1016/S0300-483X(03)00150-1.

[21] Irfan Ali Kandhro et al. "Performance evaluation of E-VGG19 model: Enhancing real-time skin cancer detection and classification". Em: *Heliyon* 10.10 (2024), e31488. ISSN: 2405-8440. DOI: https://doi.org/10.1016/j.heliyon.2024.e31488. URL: https://www.sciencedirect.com/science/article/pii/S2405844024075194.

[22] C. Karimkhani et al. "The global burden of melanoma: Results from the Global Burden of Disease Study 2015". Em: *British Journal of Dermatology* 177 (2017), pp. 134–140. DOI: 10.1111/bjd.15510. URL: https://doi.org/10.1111/bjd.15510.

[23] Staffs Keele et al. *Guidelines for performing systematic literature reviews in software engineering*. Rel. técn. Citeseer, 2007.

[24] Newton M. Kinyanjui et al. "Estimating Skin Tone and Effects on Classification Performance in Dermatology Datasets". Em: *arXiv preprint arXiv:1910.13268* (2019). NeurIPS 2019 Workshop on Fair ML for Health. DOI: 10.48550/arXiv.1910.13268. arXiv: 1910.13268 [cs.CV]. URL: https://arxiv.org/abs/1910.13268.

[25] Barbara Kitchenham et al. "Systematic literature reviews in software engineering–a systematic literature review". Em: *Information and software technology* 51.1 (2009), pp. 7–15.

[26] Renata Garcez da Luz et al. "Analysis of the variation in the number of cases of malignant melanoma and other skin malignancies in all regions of Brazil from 2015 to 2022". Em: *Research, Society and Development* 12.6 (jun. de 2023), e17712641967. DOI: 10.33448/rsd-v12i6.41967. URL: https://rsdjournal.org/index.php/rsd/article/view/41967.

[27] CM Magro, AN Crowson e MC Mihm. "Unusual variants of malignant melanoma". Em: *Modern Pathology* 19 (2006), S41–S70.

[28] Harold S. Maresky et al. "Virtual reality and cardiac anatomy: exploring immersive three-dimensional cardiac imaging, a pilot study in undergraduate medical anatomy education". Em: *Clinical Anatomy* 32.2 (2019), pp. 238–243.

[29] Riccardo Miotto et al. "Deep learning for healthcare: review, opportunities and challenges". Em: *Briefings in Bioinformatics* 19.6 (2018), pp. 1236–1246.

[30] AN Moshell, PD Parikh e WJ Oetgen. "Characteristics of medical professional liability claims against dermatologists: data from 2704 closed claims in a voluntary registry". Em: *Journal of the American Academy of Dermatology* 66.1 (2012), pp. 78–85.

[31] Novology. *A Guide to Finding the Right Sunscreen for Your Skin Tone*. Accessed: 2024-07-25. 2024. URL: https://www.novology.com/blogs/pigmentation/a-guide-to-finding-the-right-sunscreen-for-your-skin-tone.

[32] Ziad Obermeyer et al. "Dissecting racial bias in an algorithm used to manage the health of populations". Em: *Science* 366.6464 (2019), pp. 447–453.

[33] Alvin Rajkomar et al. "Ensuring fairness in machine learning to advance health equity". Em: *Annals of Internal Medicine* 169.12 (2018), pp. 866–872.

[34] HM Rayess et al. "A critical analysis of melanoma malpractice litigation: should we biopsy everything?" Em: *Laryngoscope* 127.1 (2017), pp. 134–139.

[35] Veronica Rotemberg et al. "The Role of Public Challenges and Data Sets Towards Algorithm Development, Trust, and Use in Clinical Practice". Em: *Seminars in Cutaneous Medicine and Surgery* 38.1 (mar. de 2019), E38–E42.

[36] Karen Simonyan e Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". Em: *arXiv preprint arXiv:1409.1556* (2014). Submitted on 4 Sep 2014 (v1), last revised 10 Apr 2015 (this version, v6). URL: https://doi.org/10.48550/arXiv.1409.1556.

[37] Strahil Strashilov e Angel Yordanov. "Aetiology and Pathogenesis of Cutaneous Melanoma: Current Concepts and Advances". Em: *International Journal of Molecular Sciences* 22.6395 (2021). DOI: 10.3390/ijms22126395. URL: https://doi.org/10.3390/ijms22126395.

[38] DB Troxel. "Medicolegal aspects of error in pathology". Em: *Archives of Pathology Laboratory Medicine* 130.5 (2006), pp. 617–619.

[39] P. Tschandl et al. "Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study". Em: *The Lancet Oncology* 21.1 (2020), pp. 138–149. DOI: 10.1016/S1470-2045(19)30761-6.

[40] Alan Turing. "The impact of interdisciplinary collaboration on the development of artificial intelligence". Em: *AI & Society* 33.4 (2018), pp. 465–479.

[41] John O Woolliscroft e Joel D Howell. “Interdisciplinary collaboration in health care: Implications for teaching and learning”. Em: *Journal of Interprofessional Care* 33.5 (2019), pp. 479–482.

[42] Andressa Xavier. *Aprenda a usar as Normas da ABNT: Citação (2 de 4)*. Disponível em https://www.tecmundo.com.br/tutorial/834-aprenda-a-usar-as-normas-da-abnt-citacao-2-de-4-.htm, acessado em 05/10/2021. 2020.