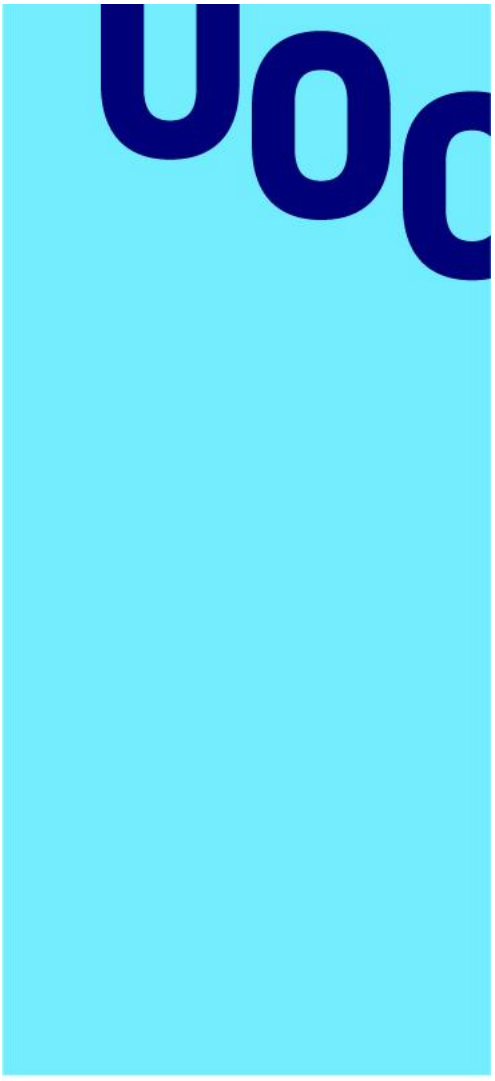
# CAA1: Definition and work plan

Student: Maria Lucas Gascón
Tutor: Romina Astrid Rebrij
Date: 16/10/2023



Universitat Oberta de Catalunya

## Index

## FINAL PROJECT'S SHEET

| | |
|---|---|
| **Project's title:** | *Development of a Web App for Skin Alteration Classification Using Machine Learning* |
| **Author's name:** | *Maria Lucas Gascón* |
| **Consultant's name:** | *Romina Astrid Rebrij* |
| **PRA's name:** | *Carles Ventura Royo* |
| **Delivery date:** | *02/2024* |
| **Program:** | *Master's degree in Bioinformatics and biostatistics UOC-UB* |
| **Final project's area:** | *Statistical Bioinformatics and Machine Learning* |
| **Language:** | *English* |
| **Keywords** | *Machine Learning, Dermatology, Diagnosis* |

**Abstract**

The purpose of this master's degree final project is to create an accessible and user-friendly web application empowered by a robust machine learning algorithm, designed to accurately classify a wide range of skin alterations based on user-submitted images. This project addresses the growing need for accessible and efficient tools in the field of dermatology, where timely identification of skin conditions can significantly impact patient outcomes.

The context of this web app is open and free access to anyone with an internet connection. By eliminating entry barriers, we aim to democratize the process of diagnosing skin conditions, making it available to individuals worldwide.

Our methodology centers on the utilization of Python for developing a supervised machine learning algorithm. We explore and compare various machine learning methods, selecting the most effective approach for accurately classifying skin lesions from submitted images. Additionally, we employ the Django framework to build the web application, ensuring a seamless user experience.

The primary objective of this project is to demonstrate the successful implementation of an operational and user-friendly method for classifying diverse skin lesions through image analysis. The web application's intuitive interface simplifies the process of uploading images and provides real-time classification results.

In conclusion, this master's degree final project makes a valuable contribution to healthcare technology by fostering early detection and empowering individuals to seek appropriate medical help and guidance when indicated by the app. The open-access nature of the web app ensures its widespread utility, promoting global skin health awareness and enhancing the accessibility of dermatological expertise.

# Abbreviations list

| | |
|---|---|
| AI | Artificial Intelligence |
| BCC | Basal Cell Carcinoma |
| CNN | Convolutional Neural Network |
| ISIC | International Skin Imaging Collaboration |
| kNN | k-Nearest Neighbors |
| ML | Machine Learning |
| NIH | National Cancer Institute |
| ReLU | Rectified Linear Unit |
| SCC | Squamous Cell Carcinoma |
| SGC | Sebaceous Gland Carcinoma |
| SGD | Sustainable Development Goals |
| SVM | Support Vector Machines |
| UI | User Interface |

# 1.Background and justification of the MTP

## 1.1 General description

This project encompasses the development of a user-friendly, freely accessible web application equipped with a powerful machine learning algorithm designed to classify a wide range of skin lesions based on submitted images. This innovative solution aims to revolutionize the field of dermatology by providing an efficient and accessible means for individuals to obtain preliminary assessments of their skin conditions, regardless of their geographical location or socioeconomic status.

### Key components

1. **Machine Learning Algorithm:** At the project's core will be a robust machine learning algorithm developed in Python. The algorithm will be meticulously designed to analyze and classify skin lesions, drawing from a diverse dataset and employing various machine learning techniques to ensure accuracy and reliability.
2. **Web Application:** The project will include the creation of an intuitive web interface, built using the Django framework, to facilitate seamless user interaction. Users can easily upload images of their skin conditions through the web app and receive real-time classification results.
3. **Accessibility**: One of the project's primary objectives is to eliminate barriers to entry. The web app is accessible to anyone with an internet connection, making it available to a global audience, including regions with limited access to dermatological expertise.
4. **Education and Awareness:** Beyond classification, the web app also serves as an educational platform. Users can access information about various skin conditions, fostering awareness and encouraging proactive skin health practices.

## 1.2 Context and justification

### Dermatology

Dermatology, the branch of medicine specializing in the study, diagnosis, and treatment of skin disorders and conditions, plays a pivotal role in healthcare worldwide. The skin, as the body's largest organ, serves as a protective barrier between the body's internal systems and the external environment. It encompasses a vast spectrum of issues, ranging from common and benign ailments to severe and

life-threatening conditions, necessitating thorough examination and diagnosis for effective treatment (Braun-falco, et al., 2013).

Skin cancer has emerged as one of the most prevalent and concerning forms of cancer in recent times (Craythorne and Al-Niami, 2017). Given that the skin is the body's largest organ, it comes as no surprise that skin cancer ranks as one of the most commonly diagnosed types of cancer among humans (Leiter and Garbe, 2008). Broadly categorized into melanoma and nonmelanoma skin cancer, these conditions demand our utmost attention (Netscher et al., 2011).

In the realm of skin cancer management, the crux lies in the early detection of these conditions. Traditional diagnostic methods often involve the painful, time-consuming, and invasive procedure of skin lesion biopsy (Abhishek and Khunger, 2015).

The different skin conditions this project studies include actinic keratosis, basal cell carcinoma, dermatofibroma, melanoma, nevus, squamous cell carcinoma, seborrheic keratosis, solar lentigo, and vascular lesions. To shed light on the importance of this project, let's delve into some specific statistics and insights regarding these skin conditions.

Melanoma, in particular, represents a perilous, relatively rare, and potentially lethal variant of skin cancer. Despite accounting for only 1% of all reported cases, melanoma's fatality rate is notably higher (Holme, et al., 2000). This malignant cancer originates in melanocytes, the cells responsible for pigment production, and occurs when these melanocytes lose control over their growth, leading to the formation of cancerous tumors. Melanoma can manifest in various parts of the body, with a predilection for areas exposed to sunlight, such as the hands, face, neck, and lips. Early diagnosis is paramount for effective treatment, as untreated melanomas can metastasize to other parts of the body, resulting in devastating consequences for the individual (Khan, et al., 2019). Notably, melanoma encompasses diverse subtypes, including nodular melanoma, superficial spreading melanoma, acral lentiginous melanoma, and lentigo maligna (Katalinic, et al., 2003).

Conversely, the majority of skin cancer cases fall under the nonmelanoma category, which includes basal cell carcinoma (BCC), squamous cell carcinoma (SCC), and sebaceous gland carcinoma (SGC). BCC, SCC, and SGC originate in the middle and upper layers of the epidermis. Fortunately, these nonmelanoma cancers typically exhibit a lower propensity to spread to other parts of the body, making them more amenable to treatment (Madan, et al., 2010).

According to the National Cancer Institute (NIH), melanoma of the skin is the 5th most common type of cancer, representing approximately 5% of all cancer cases. Furthermore, melanoma cases have been rising on average 1.2% each year over 2010-2019 (NIH, 2022). Recognizing and categorizing melanoma based on visual cues is a critical step in saving lives. The “ABCDE” is the standard to watch for (NIH, 2019):

- Asymmetry - When one side doesn't mirror the other.
- Border - The edges appear jagged, blurred, or uneven.
- Color - Uneven coloring, possibly with mixtures of black, brown, and tan.
- Diameter - There is a noticeable change in size, often an increase.
- Evolving - The mole has changed in recent weeks or months.

Actinic keratosis, for instance, is a precursor to skin cancer, particularly affecting those with a history of excessive sun exposure. It is a prevalent condition, with millions of cases diagnosed globally. Dermatofibroma, while benign, affects numerous individuals, and it's essential to distinguish it from potentially malignant growths.

Nevus, commonly referred to as moles, can vary in size, shape, and color. Some moles are atypical and may warrant closer examination due to their potential link to skin cancer. Seborrheic keratosis, while benign, can often be mistaken for more sinister growths, making accurate identification essential. Solar lentigo, or age spots, frequently appear with age and sun exposure and can lead to cosmetic concerns.

Vascular lesions, a broad category of skin conditions, involve various blood vessel-related issues. Conditions like port-wine stains, hemangiomas, and telangiectasias, often referred to as spider veins, exhibit diverse appearances and require precise diagnosis for appropriate management (Braun-falco, et al., 2013).

## Machine Learning

Machine learning (ML) is a pivotal subfield within the realm of artificial intelligence (AI) that revolves around the development of algorithms and models with the capacity to learn from data and make predictions or decisions without explicit programming. It's a systematic approach to solving intricate problems by leveraging statistical and computational techniques. At its core, machine learning revolves around data, which comprises observations or examples.

This data is usually represented as a collection of input features (attributes or characteristics) and output labels (the target or desired outcome) (Lantz, 2021).

Central to the concept of machine learning is the machine learning model or algorithm. A machine learning model is essentially a mathematical construct designed to capture the underlying patterns and relationships inherent in the data. These models can vary in complexity, from straightforward linear equations to highly intricate neural networks. Training is the process by which a machine learning model is taught. During training, the model is provided with a dataset that contains both input and output data, and it endeavors to learn and adapt its internal parameters to minimize the disparity between its predictions and the actual outcomes in the training data. Upon successful training, a machine learning model can then make predictions or inferences on new, previously unseen data. It does this by applying the learned patterns and relationships it has extracted during the training phase.

Algorithms serve as the set of rules and instructions that guide a machine learning model in making predictions based on the patterns it has learned from the training data. Numerous algorithms exist in the realm of machine learning, including decision trees, support vector machines, and deep neural networks (Lantz, 2019).

Machine learning encompasses two primary categories: supervised and unsupervised learning. Supervised learning involves training a model using labeled data, where the correct answers are provided, enabling it to learn to map inputs to corresponding outputs. In unsupervised learning, models explore patterns and structures in unlabeled data, often through clustering or dimensionality reduction (Donalek, 2011).

After training a machine learning model, it is essential to evaluate its performance using separate validation or test data to ensure its ability to generalize effectively to new, unseen data. Various metrics, such as accuracy, precision, recall, and F1 score, are used to assess model performance (Handelman, et al., 2019).

One of the critical challenges in machine learning is avoiding both overfitting and underfitting (Dietterich, 1995). Overfitting occurs when a model learns noise in the training data and fails to generalize, while underfitting transpires when the model is too simplistic to capture the underlying patterns in the data (Roelofs, et al., 2019).

Machine learning algorithms often have hyperparameters that need to be tuned to optimize a model's performance. These hyperparameters control aspects such as the learning rate, model complexity, or regularization strength (Probst, et al., 2019).

Deep learning is a subfield of machine learning that focuses on the use of artificial neural networks to model and solve complex problems. These neural networks are composed of multiple layers, allowing them to automatically learn and represent hierarchical patterns and features from data (LeCun, et al., 2015). Deep learning has gained prominence due to its remarkable ability to excel in tasks such as image recognition, natural language processing, and speech recognition (Erickson, et al., 2017). It's particularly well-suited for tasks involving large, high-dimensional datasets and has been a driving force behind recent advancements in AI, including technologies like self-driving cars and advanced recommendation systems (Shinde and Shah, 2018).

Once a machine learning model is successfully trained and evaluated, it can be deployed to make real-world predictions or automate decision-making in various applications. Machine learning is an integral part of modern technology and data-driven decision-making processes.

## Machine learning in dermatopathology

Dermatopathologists have traditionally played a central role in diagnosing skin conditions through the examination of histopathological images and clinical data. However, this process can be time-consuming, resource-intensive, and subject to variations in diagnostic accuracy among experts (Norman, et al., 1989).

The integration of AI and ML techniques in dermatology has brought about transformative changes in the field (Jartarkar, et al., 2023). AI and ML algorithms have the capacity to process vast amounts of medical data and recognize intricate patterns, significantly enhancing the accuracy and efficiency of dermatological practices (Efimenko, et al., 2020).

AI plays a pivotal role in early skin cancer detection. By analyzing images of skin lesions, AI systems can recognize potential indicators of skin cancer in its initial stages, facilitating timely intervention and enhancing patient outcomes (Wells, et al., 2021). This capability holds particular significance within the realm of skin cancer, where timely detection is a decisive factor in successful treatment (Khan, et al., 2019).

Dermoscopy, a technique that uses specialized magnifying tools for skin lesion examination, has also benefited from AI advancements (Young, et al., 2020). AI systems can automate the analysis of dermoscopy images, assisting dermatologists and reducing the subjectivity inherent in manual analysis (Saravanan, et al., 2020).

Teledermatology, enabled by AI, allows patients to submit images of skin issues for remote consultations. AI-driven platforms assist dermatologists in triaging cases and providing initial assessments, extending healthcare access to remote or underserved areas (Du-Harpur, et al., 2020).

In addition to diagnosis and treatment, AI systems also contribute to population health management by analyzing large patient datasets to identify disease trends and evaluate the effectiveness of treatments (Reisinho, et al., 2020). Such insights aid in public health planning and interventions. Furthermore, AI-driven drug discovery in dermatology is a promising area. These algorithms can predict the efficacy of compounds and analyze potential side effects, expediting the development of dermatological drugs (Yuan, et al., 2023).

However, as AI continues to advance, addressing ethical and privacy considerations is essential. Protecting sensitive patient data and mitigating bias in AI algorithms are critical aspects of ensuring the responsible use of AI in dermatology (Kroll, 2018).

In summary, the integration of AI and ML in dermatology is revolutionizing the field by enhancing diagnostic accuracy, treatment planning, and patient care. It holds promise for continued advancements and improved patient outcomes as technology evolves and matures.

### *Convolutional neural networks*

Convolutional Neural Networks (CNNs), are a type of machine learning model particularly well-suited for image analysis, and they play a significant role in the medical field, especially in the classification of skin conditions. These networks are inspired by the organization of neurons in the human brain and are designed to automatically extract features and patterns from images (O'Shea and Nash, 2015). In the context of medical imaging, CNNs are employed to analyze skin images and make distinctions between different skin conditions, such as rashes, moles, or other abnormalities.

One of the key operations in CNNs is the convolution operation, which involves sliding small filters, or kernels, over the input image. These filters extract local features, like edges or corners, from different parts of the image. As the network progresses through multiple convolutional layers, it learns to recognize increasingly complex patterns in the images, enabling it to identify specific features related to different skin conditions (Alwabi, et al., 2017).

Pooling layers are another essential component of CNNs, which follow the convolutional layers. Pooling layers downsample the feature maps created by the convolutional layers, helping to reduce the spatial dimensions of the data while retaining the most critical information (Nigat, et al., 2023). This process helps decrease computational complexity and allows the network to focus on the most relevant details.

Furthermore, CNNs include fully connected layers towards the end of the network, which connect every neuron to every neuron in the previous layer. These layers facilitate learning high-level features and enable the network to make predictions based on the patterns it has extracted from the input images (Zhang, et al., 2020). Activation functions like ReLU (Rectified Linear Unit) introduce non-linearity into the model, which is crucial for the network to learn complex relationships in the data.

In a CNN, layers are stacked one after the other, creating a hierarchical architecture (**Figure 1**). As you move deeper into the network, the convolution layers learn to recognize increasingly complex and abstract features. The training process of a CNN involves learning from a labeled dataset. During training, the network attempts to minimize the difference between its predictions and the actual labels. This optimization process, known as backpropagation, helps the network learn and adjust its internal parameters to make accurate predictions (Cullel-Dalmau, et al., 2021).

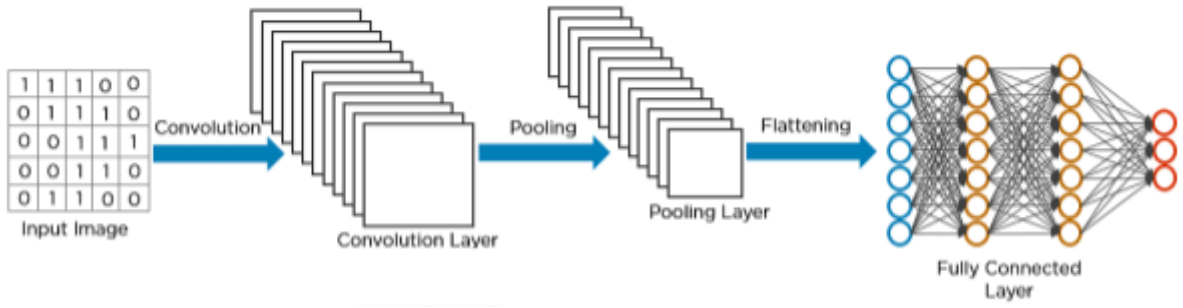



**Figure 1: Architecture of a CNN.** The CNN architecture comprises distinct layers. Initially, an image, presented as a pixel array, undergoes convolution layers, which extract key features. Subsequently, pooling layers reduce data dimensions while preserving crucial information for interpretation by fully connected layers. Image extracted from Biswal, 2023.

The application of CNN-based deep learning models in dermatopathology has shown promising results in significantly enhancing the diagnostic accuracy of skin conditions. Several studies have demonstrated that AI algorithms, when trained on large and diverse datasets, can outperform human experts in specific diagnostic tasks (Brinker, et al., 2019). This can reduce the likelihood of misdiagnoses and ensure more accurate treatment decisions.

By reducing diagnostic variability, CNNs provide a consistent and standardized approach to skin condition diagnosis. Furthermore, their efficiency and speed in processing whole-slide images can significantly expedite the diagnostic process, benefiting both patients and healthcare providers. Resource optimization is another advantage, as AI algorithms can triage cases, allowing experts to focus on complex scenarios. Additionally, AI-driven diagnostics can extend expert knowledge to underserved regions with limited access to specialist pathologists (Esteva, et al., 2017).

In light of these factors, there is a compelling need to explore and develop CNN-based models for skin condition classification. This scientific paper aims to contribute to the growing body of knowledge in dermatopathology by presenting a CNN-based approach to enhance the accuracy, efficiency, and consistency of skin condition diagnosis.

In summary, the development of this web-based platform with a machine learning algorithm for skin alteration classification addresses a pressing need in healthcare. It combines the power of technology and accessibility to bridge gaps in dermatological care, promotes early detection and appropriate medical intervention, and empowers individuals to take charge of their skin health. This project has the potential to make a significant positive impact on global healthcare and public awareness of skin conditions.

## 1.3 State-of-the-art

Skin condition classification using deep learning techniques has undergone significant advancements, with Convolutional Neural Networks (CNNs) playing a pivotal role in achieving remarkable accuracy and reliability in diagnosing dermatological conditions (Gore, 2020). This section provides an overview of the current state of the art in dermatology image classification with a focus on CNN-based approaches.

In recent years, researchers have emphasized the importance of data augmentation and preprocessing techniques to improve model performance. Augmentation methods, such as rotation, scaling, and flipping, have been employed to expand training datasets, mitigating data scarcity issues and enhancing the network's ability to generalize to unseen cases (Yanagisawa, et al., 2023).

State-of-the-art CNN architectures have evolved to leverage deep learning capabilities. Models such as DenseNet, Inception, and ResNet have been adapted for dermatological image classification, demonstrating exceptional feature extraction

capabilities (Wei, et al., 2023). Customized CNN architectures, often incorporating novel techniques such as attention mechanisms and skip connections, have been designed to capture fine-grained details specific to skin condition images (Xie, et al., 2020).

Transfer learning has emerged as a highly effective approach, with researchers leveraging pretrained models on large-scale datasets like ImageNet (Morid, et al., 2021). Fine-tuning these models for skin condition classification tasks has accelerated convergence and improved performance. In the pursuit of more accurate classification, researchers have also explored the integration of diverse data sources. Combining visual image data with clinical text, patient history, and dermoscopic images has shown promise in enhancing model accuracy (Höhn, et al., 2021). Multimodal CNNs have emerged as a key research focus in this regard (Li, et al., 2022).

As the reliability and safety of CNN-based diagnostic systems are paramount in clinical applications, research has increasingly focused on the interpretability and explainability of these models. Techniques such as Grad-CAM, LIME, and SHAP values have been applied to visualize and interpret regions of interest within images that influence the model's decision-making process (Nunnari, et al., 2021).

Community-driven initiatives have led to the development of benchmark datasets and challenges that serve as testing grounds for skin condition classification models. Notable examples include the ISIC (International Skin Imaging Collaboration) challenge and the HAM10000 dataset as well as the BCN20000 used in this project, which have spurred healthy competition and fostered innovation in the field.

Transitioning from research to practical clinical applications, validation and integration of CNN-based skin condition classifiers are essential. Studies demonstrating real-world applicability, including large-scale clinical trials, are increasingly prevalent in the literature (Brinker, et al., 2018).

In conclusion, the field of skin condition classification with CNNs is evolving rapidly, driven by data augmentation, novel architectures, transfer learning, multimodal approaches, and an increasing emphasis on interpretability and clinical validation. Researchers are working towards more accurate, reliable, and clinically relevant diagnostic tools, which have the potential to revolutionize dermatological practice.

# 2. Objectives

## 2.1 Main Objective

Develop a web-based platform equipped with a machine learning algorithm capable of accurately classifying various skin alterations based on user-submitted images.

## 2.2 Specific Objectives

1. Develop and evaluate a CNN machine learning model in Python to classify a diverse range of skin alterations.
2. Optimize the model to achieve a minimum precision rate of 85%.
3. Develop an interactive and user-friendly web application, integrating the model for skin alteration classification.

# 3. Sustainable Development Goals

In the context of Sustainable Development Goals (SDGs) established by the United Nations, our project aligns with several critical SDGs, contributing to global sustainability and development. The primary objectives of our project focus on developing a web-based platform with a machine learning algorithm capable of classifying various skin alterations based on user-submitted images. This innovative solution holds a significant potential to support the following SDGs (**Figure 2**):


**Figure 2:** An image depicting all the Sustainable Development Goals (SDGs) incorporated within this project. Extracted from https://sdgs.un.org/goals.

**SDG 3: Good Health and Well-Being**. Our project directly advances SDG 3 by addressing the goal of good health and well-being. The web-based platform offers an accessible tool for early detection and assessment of a wide range of skin conditions. By enabling individuals to submit images of their skin alterations and receive rapid and accurate assessments, our project empowers users to seek timely medical intervention when needed. This proactive approach to skin health enhances overall well-being and supports the broader goal of good health for all.

**SDG 4: Quality Education**. The project integrates an educational component that aligns with SDG 4, emphasizing quality education. In addition to skin condition classification, the web application provides information about various skin conditions. This educational feature fosters awareness, improves understanding of skin health, and encourages proactive practices. By offering knowledge and resources, our project contributes to the promotion of quality education, aligning with the United Nations' sustainable development agenda.

**SDG 9: Industry, Innovation, and Infrastructure.** Our project exemplifies the principles of innovation and technological advancement outlined in SDG 9. The development of a web-based platform equipped with a machine learning algorithm represents a significant leap in the application of technology to address healthcare challenges. By leveraging innovative approaches and technology-driven solutions, our project advances industry, innovation, and infrastructure, particularly within the healthcare sector.

**SDG 10: Reduced Inequalities**. The project is rooted in the principle of reducing inequalities, as outlined in SDG 10. By providing open and free access to skin condition assessments, our project eliminates geographic and financial barriers to healthcare. This accessible platform ensures that individuals worldwide, regardless of their geographical location or socioeconomic status, can benefit from early detection and assessment services. Through its inclusive design, the project contributes to the reduction of inequalities in healthcare access.

In conclusion, our project aligns with these SDGs by utilizing technology, education, and innovative approaches to improve health outcomes, reduce inequalities in access to healthcare, and promote awareness. The development of a web application and the application of machine learning technology exemplify the potential for technological solutions to address global challenges, ultimately contributing to sustainable development and the broader United Nations agenda.

While our project primarily focuses on improving healthcare accessibility and awareness, it's essential to acknowledge potential negative impacts and ethical considerations. First, the operation of machine learning algorithms and web servers consumes energy, contributing to environmental impact, which we recognize and aim to mitigate through responsible server hosting and energy-efficient practices. Second, data privacy and security are paramount, as user-submitted images may contain sensitive information, and we are committed to safeguarding this data. Third, we acknowledge the potential for algorithm bias and emphasize our efforts to mitigate bias and ensure fairness in the classification results. Furthermore, we stress that our web application should complement, not replace, professional medical advice, promoting informed decision-making. By addressing these concerns transparently and ethically, our project aims to ensure responsible, impactful, and equitable contributions to healthcare accessibility and awareness.

# 3. Approach and Methodology

## Dataset

The dataset employed for the purpose of this study is the BCN20000 dataset, which encompasses a total of 12,413 dermoscopic images depicting various skin abnormalities. These images were captured between the years 2010 and 2016 at the Clinic's Hospital of Barcelona. The dataset comprises a spectrum of potential diagnoses, including actinic keratosis, basal cell carcinoma, dermatofibroma, melanoma, nevus, squamous cell carcinoma, seborrheic keratosis, solar lentigo, and vascular lesions. Additionally, the metadata associated with each image provides valuable information such as age, gender, and the anatomical location where the image was shot.

This dataset is publicly accessible through the ISIC Archive, a resource established with the primary objective of enhancing the diagnosis of skin cancer. The ISIC Archive serves as a repository for standardizing skin imaging practices, collecting and disseminating dermatological images, and fostering collaboration among medical practitioners and experts in the field of computer vision.

### *Initial data exploration*

An essential aspect of understanding the dataset is to examine the distribution of images across different skin condition classes. This helps in identifying potential class imbalances and informs data preprocessing decisions. The class distribution can be visualized in the following table (**Table 1**):

| Diagnosis | Sample Size | Percentage |
|---|---|---|
| Nevus | 4206 | 33,88 |
| Melanoma | 2857 | 23,01 |
| Basal cell carcinoma | 2809 | 22,62 |
| Seborrheic keratosis | 929 | 7,48 |
| Actinic keratosis | 737 | 5,93 |
| Squamous cell carcinoma | 431 | 3,47 |
| Solar lentigo | 209 | 1,68 |
| Dermatofibroma | 124 | 0,99 |
| Vascular lesion | 111 | 0,89 |

**Table 1: Class distribution.** Table presenting the diagnosis distribution of the BCN20000 dataset. Self-made table obtained by the execution of the "**Code 1**" in the annex.

A severe class imbalance can be observed, with 4 classes getting less than 5% of the total cases. Class imbalance refers to a situation in machine learning where the number of instances in different classes of a dataset significantly varies, leading to an unequal distribution of classes. Unbalanced classes can substantially impact the study by causing the model to be biased towards the majority class, resulting in poor recognition of minority classes (Buda, et al., 2018). This imbalance may lead to a high accuracy rate, but it fails to address the primary objective of identifying rare or critical instances, such as dermatofibroma in this scenario.

To mitigate this issue, techniques like data augmentation, resampling, weighted loss functions, transfer learning or Ada-Boosting can be employed (Taherkhani, et al., 2020). As a last resource, reducing the scope of the project to include only the most common types of diagnosis may be possible. These approaches seek to create a more equitable representation of all classes, thus enhancing the model's ability to accurately classify rare classes and improving the overall robustness and reliability of the image classification system in a technical and formal manner.

As part of the initial data exploration, we have examined relevant metadata variables to gain a deeper understanding of our dataset (**Code 1**). The metadata includes information on "sex," "anatomical site," and "melanocytic" characteristics. Their inclusion is driven by the recognition that factors beyond image content might have relevance for the classification process.

- **Sex Distribution:** The sex distribution in the dataset reveals that it is reasonably balanced, with 6,499 male and 5,840 female instances, ensuring that any potential gender-based biases are minimized in our analysis.
- **Anatomical Site Distribution:** The data encompasses a variety of anatomical sites. The majority of samples are from the "anterior torso" (5,086), followed by "head/neck" (3,223), "lower extremity" (2,173), "upper extremity" (1,319), "palms/soles" (380), and "oral/genital" (59). Understanding this distribution may help us account for anatomical variations that could impact the classification of skin conditions.
- **Melanocytic Classification:** The "melanocytic" variable, denoting whether a condition is melanocytic or not, showcases that 7,063 instances are melanocytic, while 5,350 are not. This distinction is critical as melanocytic skin conditions often exhibit distinct visual characteristics.

*Image Characteristics*

Understanding the inherent characteristics of the images within our dataset is of paramount importance in preparing for the development of our Convolutional Neural Network (CNN) model. These characteristics significantly influence the design and optimization of our model architecture and data preprocessing. The dataset exhibits the following key attributes (**Code 2**):

- **Height**: The height of all the images in the dataset is 1024 pixels
- **Width**: Similarly, the width of all images in the dataset is 1024 pixels.
- **Channels**: All images in the dataset are configured with three color channels (Red, Green, and Blue), consistent with the typical RGB format.

These results are expected, as the dataset was carefully prepared by ISICS with the objective of machine learning development in mind. Comprehending these image characteristics plays a crucial role in the development of our CNN model, guiding decisions related to input shape determination, resource allocation, data augmentation, and preprocessing procedures. This knowledge enables us to tailor our CNN model and data processing to effectively accommodate these specific attributes.

However, meticulously designing the model to accept images of specific attributes (1024x1024 pixels and RGB color channels), poses the question of whether user-contributed images must adhere to these specifications. This issue may be necessary to address in the future with the implementation of a resizing and preprocessing of user-submitted images within the web application.

## Machine learning algorithms

As stated previously, CNNs are a fitting choice for image classification tasks, including the categorization of skin conditions. Its suitability stems from its prowess in recognizing intricate patterns, scalability to handle high-dimensional image data, automated feature extraction, ability to generalize from training data to new examples, adaptability to improve with more data and fine-tuning, and effective handling of image complexities (Zafar, et al., 2020).

Convolutional Neural Networks (CNNs) have emerged as the de facto standard for image classification, owing to their remarkable efficacy and unique design principles. CNNs excel in this realm due to their innate ability to automatically unearth complex hierarchical features from raw image data (Cheng-Hong, et al., 2021). A detailed list of advantages and disadvantages can be found in the following table (**Table 2**).

While traditional machine learning algorithms like Support Vector Machines (SVM), k-Nearest Neighbors (kNN), or Naive Bayes are versatile tools across various domains, their application to image classification poses inherent limitations. Unlike Convolutional Neural Networks (CNNs), these traditional algorithms require handcrafted feature engineering from high-dimensional image data, a cumbersome and potentially inadequate process for capturing intricate patterns in images (Lantz, 2019). That said, you can use traditional machine learning algorithms for image classification, but you'll typically need to hand-craft feature vectors from the images, which can be a complex and time-consuming process. Additionally, the performance of these algorithms may not match that of CNNs on image data, especially when dealing with large and diverse datasets.

So, while you can technically apply kNN, Naive Bayes, Decision Trees, or SVM to image data by flattening the images into feature vectors, it's generally more practical and effective to use CNNs for image classification tasks due to their ability to automatically learn relevant features and hierarchical representations from raw image data. For this reason, in this project we will focus on the development of a CNN algorithm that can precisely classify images of skin conditions.

| Pros and cons of CNN algorithms | |
|---|---|
| **Pros** | **Cons** |
| Automatic feature learning | Computational resources |
| Hierarchical Representation | Need for large datasets |
| Translation Invariance | Overfitting |
| Scalability | Interpretability |
| Generalization | Hyperparameter tuning |
| Adaptability | |
| Flexibility | |
| State of art performance | |

**Table 2: Pros and cons of CNN algorithms.** Table showing pros and cons of using CNN algorithms (Thompson, 2022).

## Programming language and libraries

In the construction of such an algorithm, the choice of the programming language is pivotal, and both R and Python emerge as viable options. Although there are several robust deep learning frameworks available, including TensorFlow, Keras and

MXNet, the deliberate choice for the programming language in this project is Python. Python's selection is underpinned by its robust attributes, notably its extensive library ecosystem, and its status as one of the most pervasive and widely-supported programming languages. The magnitude of its community ensures ready access to valuable resources, which is particularly advantageous in the context of this project. Furthermore, Python's versatility renders it apt for diverse tasks encompassing data preprocessing, data visualization, statistical analysis, and the availability of well-established deployment frameworks such as Django and Flask, which are conducive to the realization of the intended web application.

Concurrently, TensorFlow, a preeminent deep learning framework, asserts its prominence in the landscape of machine learning tools. Within TensorFlow, the high-level Keras API stands as an accessible and user-intuitive vehicle for conceiving, training, and assessing convolutional neural network (CNN) models. The endorsement of TensorFlow and Keras is substantiated by their comprehensive documentation and an expansive network of support, rendering them a superlative choice for developers. Familiarity with this library, owing to prior usage in the development of other machine learning algorithms, serves as a compelling rationale for its adoption in this context. While a multitude of alternative libraries such as PyTorch, Caffe, MXNet, and Fastai could be considered, the aforementioned factors collectively contribute to the preference for TensorFlow and Keras in this project.

# 4. Planning with milestones and calendar

## 4.1 Main tasks and prioritization

1. **Definition of the work plan**: Define the project's scope, objectives, methodology and expected outcomes. Create a project charter outlining the project's purpose and goals as well as a calendar with milestones and dates.
2. **Data Collection and Preparation**: Curate a comprehensive and diverse dataset of skin alteration images, ensuring that it represents various skin conditions and ages to train and validate the machine learning model. Split the data into training, validation, and test sets. Preprocess the images, normalizing pixel values to the range, and augmenting the data if needed.
3. **Algorithm Development**: Develop a machine learning algorithm in Python capable of accurately classifying a diverse range of skin alterations.
4. **Web Application Development**: Design and develop a user-friendly web application using the Django framework to enable users to upload skin images for classification.

## 4.2 Extra tasks

1. **Algorithm Evaluation**: Evaluate the performance of the machine learning algorithm by conducting rigorous testing, including metrics such as accuracy, sensitivity, specificity, and precision, to ensure its effectiveness in skin condition classification.
2. **Algorithm Optimization:** Investigate the potential of utilizing a pre-trained model such as VGG16, ResNet50, or MobileNet. Identify the most effective pre-processing technique to enhance the model's performance.
3. **Accessibility and Scalability**: Ensure that the web application is accessible to anyone with an internet connection, and design it to be scalable for potential future enhancements and improvements.
4. **Educational Content Integration**: Incorporate educational content within the web application to provide users with information about various skin conditions, fostering awareness and encouraging proactive skin health practices.
5. **User Experience Testing**: Conduct user experience testing to refine the web application's interface and functionality, ensuring ease of use and accessibility for a wide range of users.
6. **Data Security and Privacy**: Implement robust data security and privacy measures to protect user-submitted images and information, complying with relevant regulations and standards.

7. **Deployment and Maintenance**: Deploy the web application to a reliable hosting environment and establish a plan for ongoing maintenance and updates to ensure continued functionality and accuracy.

## Calendar

This Gantt chart (**Figure 3**) provides a visual representation of the project's timeline, helping to track progress and ensure that the project's objectives are met within the specified timeframes.

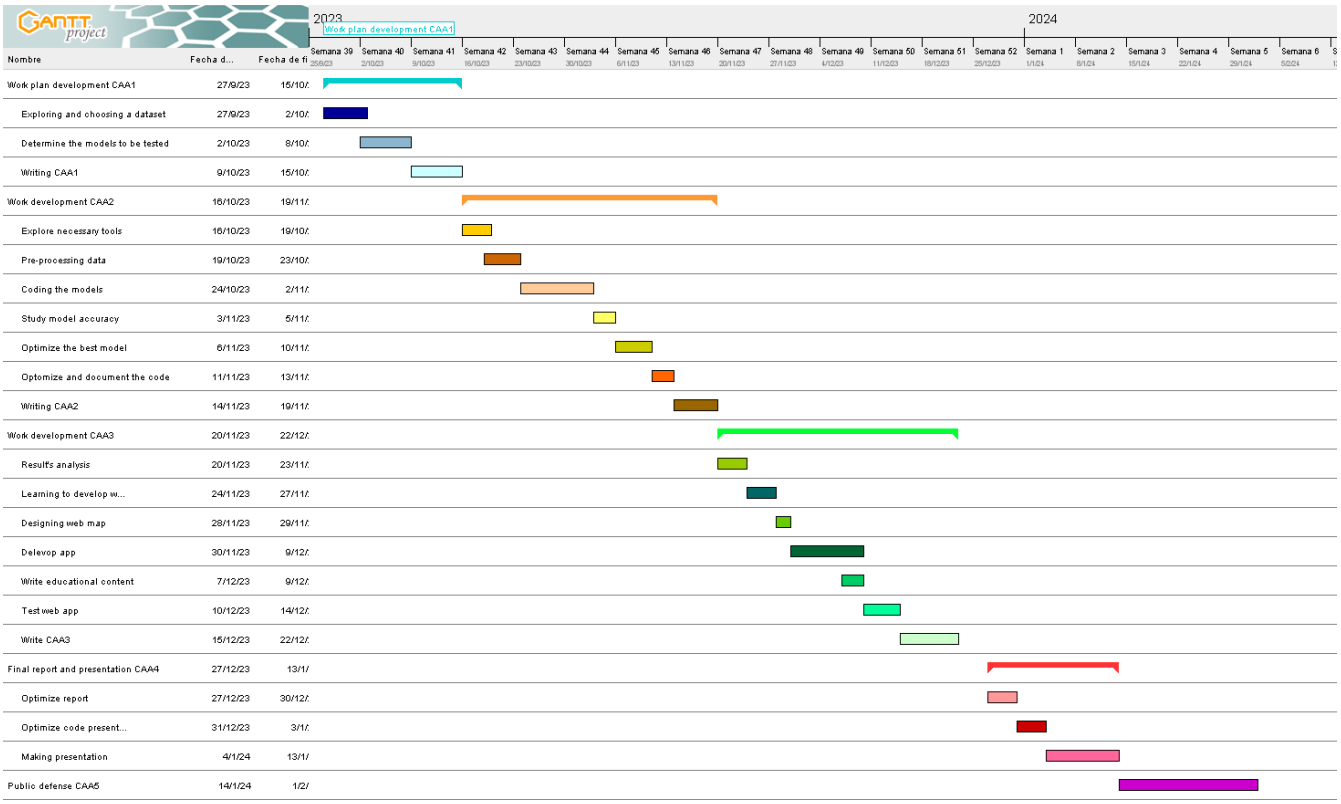



**Figure 3: Project timeline.** Gantt chart providing an overview of the timeline of the project. Self-made image using the “GanttProject” program.

This chart can be found with better resolution in the annex, as **supplementary figure 1**. There a detailed list of the tasks that appear on the calendar can also be found with the starting date and finish date (**supplementary figure 2**).

# 4.3 Milestones

- **CAA1: October 16, 2023** - Achievement of the development milestone encompassing the formulation of the comprehensive work plan.

- **CAA2: November 19, 2023** - Successful culmination of Phase 1 of the development process, characterized by the establishment of a model with the ability to accurately classify images, achieving an 85% accuracy rate.
- **CAA3: December 22, 2023** - Successful completion of Phase 2 of the development project, marked by the realization of a fully operational web application incorporating the implemented model.
- **CAA4: January 13, 2024** - Accomplishment of the project's conclusion, including the delivery of a comprehensive project report, a presentation, and an explanatory video.
- **CAA5: February 1, 2024** - Public presentation and defense of the project outcomes.

## 4.4 Risk analysis

| Risk | Severity | Likelihood | Mitigation |
|---|---|---|---|
| Data availability and quality | Moderate | Low | Perform an extensive data exploration and explore alternative sources and datasets. |
| Algorithm performance and usage | High | Moderate | Conduct rigorous algorithm evaluation and testing using cross-validation techniques and investigate optimization methods. |
| User adoption and awareness | High | Moderate | Include proper warnings about proper use of the app as well as health information. |
| Technical challenges | Moderate | High | Perform proper exploration of the libraries and seek guidance and mentorship from professors or experts in relevant fields. |
| Legal and ethical compliance | High | Low | Adhere to academic integrity guidelines and ethical standards when conducting research.<br>Properly cite and reference sources, code, and data used in the project. |
| Resource constraints | Moderate | Moderate | Create a well-defined project schedule with milestones and allocate sufficient time for each phase. Plan for contingencies. |

**Table 3: Risk analysis.** This table presents various risks associated with the project, along with their severity, likelihood, and potential mitigation measures.

# 5. Expected results

**CAA1:** This PDF document contains the comprehensive work plan and calendar, meticulously outlining all requisite tasks for project completion. It also defines the project's objectives and includes a robust risk management strategy.

**Report:** PDF document providing an exhaustive account of the project's processes, encompassing investigation, development, results, conclusions, and discussions.

**Web-application:** link to the web application developed in Django, which hosts the machine learning algorithm capable of classifying various skin abnormalities.

**GitHub Repository:** GitHub Repository housing all the requisite code essential for the project's replication, encompassing both the machine learning model and the web application.

**Virtual presentation:** Presentation in PPT format offering support and a platform for the comprehensive presentation of the undertaken work. A video recording, accompanied by explanatory narration, will also be available to enhance the understanding of the project.

# 6. Structure of the MPT

The concluding report will encompass the subsequent sections:

1. **Cover with Basic Information**: Includes the project's title, my name, the date, and the name of the tutor and responsible teacher.
2. **Summary Sheet:** A brief, condensed overview of the project, highlighting its key points and findings.
3. **Abbreviations List:** A compilation of commonly used abbreviations throughout the project, accompanied by their full explanations to aid readers in comprehension.
4. **Index:** An organized list of all sections and subsections in the project, offering a quick reference for readers to locate specific information.
5. **Introduction**:
    - *Context and Justification:* Background information and the rationale behind the project's significance.
    - *State of Art*: A review of the current state of knowledge in the relevant field.
    - *Objectives*: Clearly defined project goals and the intended achievements.
    - *SGD*: Details regarding the alignment of the project with Sustainable Development Goals, explaining how it contributes to these global objectives.
    - *Methodology*: An overview of the methods and approaches utilized.
    - *Planning and Calendar*: Information about the project's timeline and scheduling.
    - *Final Product*: A description of the project outcome.
6. **Materials and Methods:** Detailed information about the materials, equipment, and methodologies employed in the project, enabling others to replicate the work.
7. **Results:** Presentation of the project's data, observations, and outcomes.
8. **Conclusion:** A summary of the key findings and their implications, providing a conclusion to the project.
9. **Discussion**: A critical analysis and interpretation of the results, along with their relevance to the project's objectives and their impact on the field.
10. **Annex:**
    - *Bibliography*: A list of all sources, references, and literature consulted during the project.
    - *Code*: All code essential for comprehending the report and the project.
    - *Supplementary figures*: Visual aids, such as charts, graphs, and images, supporting the project.

# 7. Annex

## Bibliography

Abhishek, K., & Khunger, N. (2015). Complications of skin biopsy. *Journal of cutaneous and aesthetic surgery*, 8(4), 239.

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017, August). Understanding of a convolutional neural network. *In 2017 international conference on engineering and technology (ICET)* (pp. 1-6). Ieee.

Biswal, A. (2023, April 24). Convolutional Neural Network Tutorial. *Simplilearn*. Retrieved October 15, 2023, from https://www.simplilearn.com/tutorials/deep-learning-tutorial/convolutional-neural-network

Braun-Falco, O., Plewig, G., Wolff, H. H., & Winkelmann, R. K. (2013). Dermatology. *Springer Science & Business Media.*

Brinker, T. J., Hekler, A., Enk, A. H., Klode, J., Hauschild, A., Berking, C., ... & Schrüfer, P. (2019). A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. *European Journal of Cancer, 111,* 148-154.

Brinker, T. J., Hekler, A., Utikal, J. S., Grabe, N., Schadendorf, D., Klode, J., ... & Von Kalle, C. (2018). Skin cancer classification using convolutional neural networks: systematic review. *Journal of medical Internet research*, 20(10), e11936.

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, *106*, 249-259.

Cheng-Hong, Y., Jai-Hong, R., Huang, H. C., Li-Yeh, C., & Po-Yin, C. (2021). Deep Hybrid Convolutional Neural Network for Segmentation of Melanoma Skin Lesion. *Computational Intelligence and Neuroscience: CIN, 2021*.

Craythorne, E., & Al-Niami, F. (2017). Skin cancer. Medicine, 45(7), 431-434.

Cullell-Dalmau, M., Noé, S., Otero-Viñas, M., Meić, I., & Manzo, C. (2021). Convolutional neural network for skin lesion classification: understanding the fundamentals through hands-on learning. *Frontiers in Medicine, 8,* 644327.

Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM computing surveys (CSUR), 27(3),* 326-327.

Donalek, C. (2011, April). Supervised and unsupervised learning. In *Astronomy Colloquia*. USA (Vol. 27, p. 8).

Du-Harpur, X., Watt, F. M., Luscombe, N. M., & Lynch, M. D. (2020). What is AI? Applications of artificial intelligence to dermatology. *British Journal of Dermatology, 183(3),* 423-430.

Efimenko, M., Ignatev, A., & Koshechkin, K. (2020). Review of medical image recognition technologies to detect melanomas using neural networks. *BMC bioinformatics, 21,* 1-7.

Erickson, B. J., Korfiatis, P., Akkus, Z., Kline, T., & Philbrick, K. (2017). Toolkits and libraries for deep learning. Journal of digital imaging, 30, 400-405.
Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542(7639),* 115-118.

Gore, J. C. (2020). Artificial intelligence in medical imaging. *Magnetic resonance imaging, 68,* A1-A4.

Handelman, G. S., Kok, H. K., Chandra, R. V., Razavi, A. H., Huang, S., Brooks, M., ... & Asadi, H. (2019). Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. *American Journal of Roentgenology, 212(1),* 38-43.

Höhn, J., Hekler, A., Krieghoff-Henning, E., Kather, J. N., Utikal, J. S., Meier, F., ... & Brinker, T. J. (2021). Integrating patient data into skin cancer classification using convolutional neural networks: systematic review. *Journal of medical Internet research, 23(7),* e20708.

Holme, S. A., Malinovszky, K., & Roberts, D. L. (2000). Changing trends in non‐melanoma skin cancer in South Wales, 1988–98. *British Journal of Dermatology, 143(6),* 1224-1229.

Jartarkar, S. R., Cockerell, C. J., Patil, A., Kassir, M., Babaei, M., Weidenthaler‐Barth, B., ... & Goldust, M. (2023). Artificial intelligence in Dermatopathology. *Journal of Cosmetic Dermatology, 22(4),* 1163-1167.

Katalinic, A., Kunze, U., & Schäfer, T. (2003). Epidemiology of cutaneous melanoma and non-melanoma skin cancer in Schleswig-Holstein, Germany: incidence, clinical subtypes, tumour stages and localization (epidemiology of skin cancer). *British Journal of Dermatology, 149(6),* 1200-1206.

Khan, M. Q., Hussain, A., Rehman, S. U., Khan, U., Maqsood, M., Mehmood, K., & Khan, M. A. (2019). Classification of melanoma and nevus in digital images for diagnosis of skin cancer. *IEEE Access, 7,* 90132-90144.

Kroll, J. A. (2018). Data science data governance [AI ethics]. *IEEE Security & Privacy, 16(6),* 61-70.

Lantz, B. (2019). Machine learning with R: expert techniques for predictive modeling. *Packt publishing ltd*.

Lantz, B. (2021). Overview of Machine Learning Tools. In *The Machine Age of Customer Insight* (pp. 79-90). Emerald Publishing Limited.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature,* 521(7553), 436-444.

Leiter, U., & Garbe, C. (2008). Epidemiology of melanoma and nonmelanoma skin cancer—the role of sunlight. *Sunlight, vitamin D and skin cancer,* 89-103.

Li, Z., Wang, H., Han, Q., Liu, J., Hou, M., Chen, G., ... & Weng, T. (2022). Convolutional Neural Network with Multiscale Fusion and Attention Mechanism for Skin Diseases Assisted Diagnosis. *Computational Intelligence and Neuroscience,* 2022.

Madan, V., Lear, J. T., & Szeimies, R. M. (2010). Non-melanoma skin cancer. *The lancet, 375(9715),* 673-685.

Morid, M. A., Borjali, A., & Del Fiol, G. (2021). A scoping review of transfer learning research on medical image analysis using ImageNet. *Computers in biology and medicine, 128,* 104115.

Netscher, D. T., Leong, M., Orengo, I., Yang, D., Berg, C., & Krishnan, B. (2011). Cutaneous malignancies: melanoma and nonmelanoma types. *Plastic and reconstructive surgery, 127(3),* 37e-56e.

Nigat, T. D., Sitote, T. M., & Gedefaw, B. M. (2023). Fungal Skin Disease Classification Using the Convolutional Neural Network. *Journal of Healthcare Engineering,* 2023.

NIH: National Cancer Institute. (2019, December 16). *Melanoma*. Retrieved October 15, 2023, from https://medlineplus.gov/melanoma.html

NIH: National Cancer Institute. (2022). *Cancer Stat Facts: Melanoma of the Skin.* Retrieved October 15, 2023, from https://seer.cancer.gov/statfacts/html/melan.html

Norman, G. R., Rosenthal, D., Brooks, L. R., Allen, S. W., & Muzzin, L. J. (1989). The development of expertise in dermatology. *Archives of dermatology, 125(8),* 1063-1068.

Nunnari, F., Kadir, M. A., & Sonntag, D. (2021, August). On the overlap between grad-cam saliency maps and explainable visual features in skin cancer images. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction* (pp. 241-253). Cham: Springer International Publishing.

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:*1511.08458.

Probst, P., Boulesteix, A. L., & Bischl, B. (2019). Tunability: Importance of hyperparameters of machine learning algorithms. *The Journal of Machine Learning Research, 20(1)*, 1934-1965.

Reisinho, J., Coimbra, M., & Renna, F. (2020, July). Deep convolutional neural network ensembles for multi-classification of skin lesions from dermoscopic and clinical images. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 1940-1943). IEEE.

Roelofs, R., Shankar, V., Recht, B., Fridovich-Keil, S., Hardt, M., Miller, J., & Schmidt, L. (2019). A meta-analysis of overfitting in machine learning. Advances in *Neural Information Processing Systems, 32*.

Saravanan, S., Heshma, B., Shanofer, A. A., & Vanithamani, R. (2020). Skin cancer detection using dermoscope images. *Materials Today: Proceedings, 33,* 4823-4827.

Shinde, P. P., & Shah, S. (2018, August). A review of machine learning and deep learning applications. In *2018 Fourth international conference on computing communication control and automation* (ICCUBEA) (pp. 1-6). IEEE.

Taherkhani, A., Cosma, G., & McGinnity, T. M. (2020). AdaBoost-CNN: An adaptive boosting algorithm for convolutional neural networks to classify multi-class imbalanced datasets using transfer learning. *Neurocomputing*, *404*, 351-366.

Thompson, D. (2022, July 03). The Pros And Cons Of Convolution Neural Networks. iTechPost. Retrieved October 16, 2023 from https://www.itechpost.com/articles/109452/20220307/the-pros-and-cons-of-convolution-neural-networks.htm

Wei, M., Wu, Q., Ji, H., Wang, J., Lyu, T., Liu, J., & Zhao, L. (2023). A Skin Disease Classification Model Based on DenseNet and ConvNeXt Fusion. *Electronics, 12(2)*, 438.

Wells, A., Patel, S., Lee, J. B., & Motaparthi, K. (2021). Artificial intelligence in dermatopathology: Diagnosis, education, and research. *Journal of Cutaneous Pathology, 48(8),* 1061-1068.

Xie, F., Yang, J., Liu, J., Jiang, Z., Zheng, Y., & Wang, Y. (2020). Skin lesion segmentation using high-resolution convolutional neural network. *Computer methods and programs in biomedicine, 186,* 105241.

Yanagisawa, Y., Shido, K., Kojima, K., & Yamasaki, K. (2023). Convolutional neural network-based skin image segmentation model to improve classification of skin diseases in conventional and non-standardized picture images. *Journal of dermatological science, 109(1),* 30-36.

Young, A. T., Xiong, M., Pfau, J., Keiser, M. J., & Wei, M. L. (2020). Artificial intelligence in dermatology: a primer. *Journal of Investigative Dermatology, 140(8),* 1504-1512.

Yuan, Y., Han, Y., Yap, C. W., Kochhar, J. S., Li, H., Xiang, X., & Kang, L. (2023). Prediction of drug permeation through microneedled skin by machine learning. *Bioengineering & Translational Medicine,* e10512.

Zafar, K., Gilani, S. O., Waris, A., Ahmed, A., Jamil, M., Khan, M. N., & Sohail Kashif, A. (2020). Skin lesion segmentation from dermoscopic images using convolutional neural network. *Sensors, 20(6),* 1601.

Zhang, N., Cai, Y. X., Wang, Y. Y., Tian, Y. T., Wang, X. L., & Badami, B. (2020). Skin cancer diagnosis based on optimized convolutional neural network. *Artificial intelligence in medicine, 102,* 101756.

## Code

**Code 1:** Essential code for extracting metadata statistics.

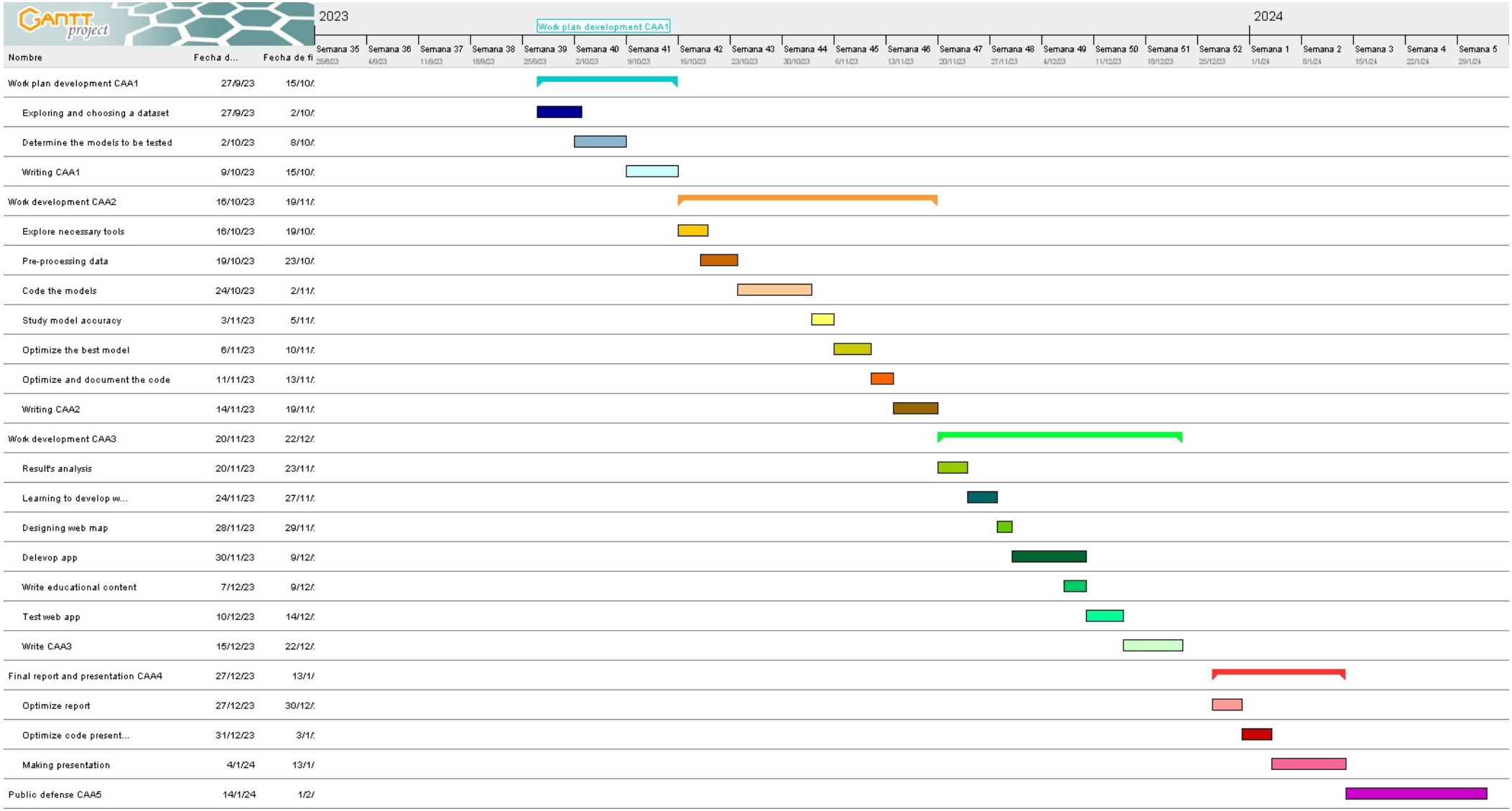
```
import pandas as pd
# Load file
dataset = pd.read_csv('bcn20000_metadata_2023-09-29.csv')

# Examine file
print(f'Variables: {dataset.columns}')
print(f'Number of images: {len(dataset)}')

# Delete non-informative columns
col_drop = ['attribution', 'copyright_license',
'diagnosis_confirm_type', 'image_type', 'lesion_id']
data = dataset.drop(columns = col_drop) # Eliminación de las
columnas
print(f'Variables: {data.columns}')

# Data exploration
diagnosis_counts = data['diagnosis'].value_counts()

# Creation of diagnosis table
diagnosis_table = pd.DataFrame({'Diagnosis': diagnosis_counts.index,
'Sample Size': diagnosis_counts.values})
diagnosis_table['Percentage'] = (diagnosis_table['Sample Size'] /
diagnosis_table['Sample Size'].sum()) * 100
diagnosis_table['Percentage'] =
diagnosis_table['Percentage'].round(2)

# Display the resulting table
print(diagnosis_table)

# Export the DataFrame to a CSV file
diagnosis_table.to_csv('diagnosis_table.csv', index=False)

# Other info
print(data.sex.value_counts())
print(data.anatom_site_general.value_counts())
print(data.melanocytic.value_counts())
```

**Code 2**: Vital code for gathering fundamental image information.

```
import os
import cv2
import pandas as pd

# Image directory
image_directory = 'E:\TFM\BCN20000\Imagenes'

# Function to gather image information recursively
def gather_image_info(directory):
    image_info = []
    for root, _, files in os.walk(directory):
        for filename in files:
            if filename.endswith('.JPG'):  # Images are in JPG format
                file_path = os.path.join(root, filename)
                image = cv2.imread(file_path)
                if image is not None:
                    height, width, channels = image.shape
                    image_info.append({
                        'Height': height,
                        'Width': width,
                        'Channels': channels
                    })
    return image_info

# Call the function to gather image information
image_info = gather_image_info(image_directory)

# Create a DataFrame to store the gathered information
image_df = pd.DataFrame(image_info)

# Print necessary info
print(image_df.Height.value_counts())
print(image_df.Width.value_counts())
print(image_df.Channels.value_counts())
```

# Supplementary figures

**Supplementary figure 1. Project timeline.** Gantt chart providing an overview of the timeline of the project. Self-made image using the "GanttProject" program.



**Supplementary figure 2. Project calendar.** List of tasks with the starting date and ending date.

| Nombre | Fecha de inicio | Fecha de fin |
|---|---|---|
| Work plan development CAA1 | 27/9/23 | 15/10/23 |
| Exploring and choosing a dataset | 27/9/23 | 2/10/23 |
| Determine the models to be tested | 2/10/23 | 8/10/23 |
| Writing CAA1 | 9/10/23 | 15/10/23 |
| | | |
| Work development CAA2 | 16/10/23 | 19/11/23 |
| Explore necessary tools | 16/10/23 | 19/10/23 |
| Pre-processing data | 19/10/23 | 23/10/23 |
| Code the models | 24/10/23 | 2/11/23 |
| Study model accuracy | 3/11/23 | 5/11/23 |
| Optimize the best model | 6/11/23 | 10/11/23 |
| Optimize and document the code | 11/11/23 | 13/11/23 |
| Writing CAA2 | 14/11/23 | 19/11/23 |
| | | |
| Work development CAA3 | 20/11/23 | 22/12/23 |
| Result's analysis | 20/11/23 | 23/11/23 |
| Learning to develop web-app in Django | 24/11/23 | 27/11/23 |
| Designing web map | 28/11/23 | 29/11/23 |
| Delevop app | 30/11/23 | 9/12/23 |
| Write educational content | 7/12/23 | 9/12/23 |
| Test web app | 10/12/23 | 14/12/23 |
| Write CAA3 | 15/12/23 | 22/12/23 |
| | | |
| Final report and presentation CAA4 | 27/12/23 | 13/1/24 |
| Optimize report | 27/12/23 | 30/12/23 |
| Optimize code presentation and documentation | 31/12/23 | 3/1/24 |
| Making presentation | 4/1/24 | 13/1/24 |
| | | |
| Public defense CAA5 | 14/1/24 | 1/2/24 |