Challenge Kaggle 2024

Image

IMA-205 - Learning for image and object recognition

Zakaria AKIL

7 août 2025
Palaiseau, France

Table des matières

1 Challenge Description :

1.1 Introduction

A skin lesion is defined as a superficial growth or patch of the skin that is visually different and/or has a different texture than its surrounding area. Skin lesions, such as moles or birthmarks, can degenerate and become cancer, with melanoma being the deadliest skin cancer. Its incidence has increased during the last decades, especially in the areas mostly populated by white people.

The most effective treatment is an early detection followed by surgical excision. This is why several approaches for skin cancer detection have been proposed in the last years (non-invasive computer-aided diagnosis (CAD)).

1.2 The goal of the challenge

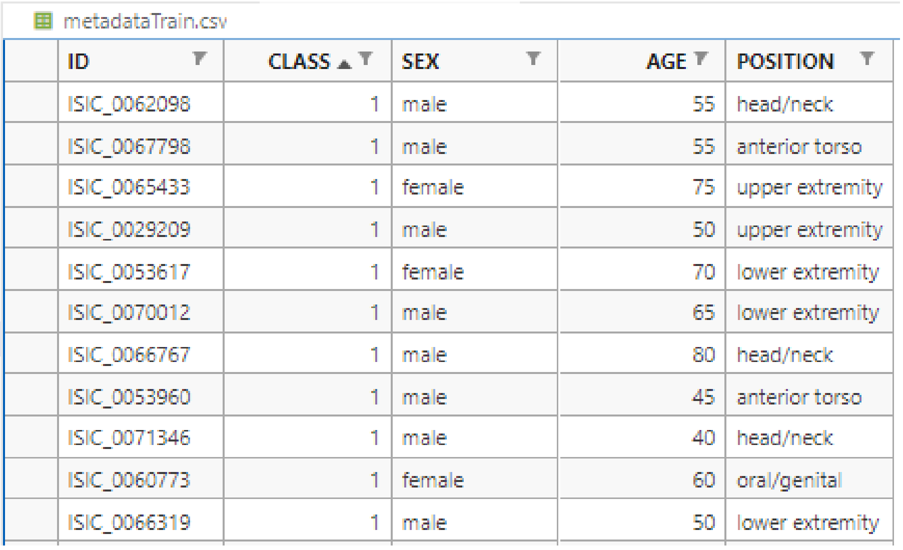
The goal of this challenge is to classify dermoscopic images of skin lesions among eight different diagnostic classes :

1. Melanoma
2. Melanocytic nevus
3. Basal cell carcinoma
4. Actinic keratosis
5. Benign keratosis
6. Dermatofibroma
7. Vascular lesion
8. Squamous cell carcinoma

In order to do that, we will extract features such as the Asymmetry, the Border irregularity, the Colour and the Dimension of the lesion (usually called the ABCD rule). After that, we will use machine learning algorithms to classify the images.

1.3 Data

We possess a data-set of **25331** dermoscopic images of skin lesions with, when available, their relative segmentation and metadata (age, sex and anatomical position).



metadataTrain.csv

ID	CLASS	SEX	AGE	POSITION
ISIC_0062098	1	male	55	head/neck
ISIC_0067798	1	male	55	anterior torso
ISIC_0065433	1	female	75	upper extremity
ISIC_0029209	1	male	50	upper extremity
ISIC_0053617	1	female	70	lower extremity
ISIC_0070012	1	male	65	lower extremity
ISIC_0066767	1	male	80	head/neck
ISIC_0053960	1	male	45	anterior torso
ISIC_0071346	1	male	40	head/neck
ISIC_0060773	1	female	60	oral/genital
ISIC_0066319	1	male	50	lower extremity

Figure 1 – Head of the train meta data

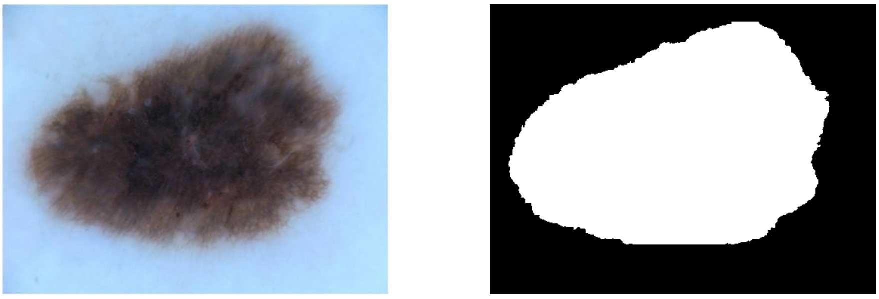


Figure 2 – Example of an dermoscopic image and its segmentation

1.4 Remark :

2 Data Preparation

We are given a huge amount of data divided into two types : images and their segmentations, and metadata. However, we are obliged to go through some preprocessing approaches due to the imperfections present in this data. Specifically, the metadata contains a significant number of missing values, which is why we will use data cleaning and data completion techniques.

Additionally, we have a large number of images without corresponding segmentations, which are crucial for the classification task using convolutional neural networks (CNNs). Using the original dermoscopic images directly as inputs can work, but it is performance-limited because the machine learning network will focus on learning the differences contained in the borders instead of the main differences between the eight diagnostic classes.

Another important reason for performing image segmentation is that it is crucial for extracting ABCD features, which are required for the challenge.

Also, we have a class imbalance problem, and as a solution, we will perform data augmentation.

2.1 Data Cleaning

- The figure below shows the missing values presented in the meta data :

feature	Train NaN values	Test NaN values
CLASS	0	0
SEX	284	100
AGE	324	113
POSITION	1970	661

Table 1 – Data Table

- The solution I adopted consists in replacing the NaN values by the median value as the features are categorical except the age feature that is numerical but integer.

After performing a data cleaning and completion, this is a visualisation of my meta data distribution.

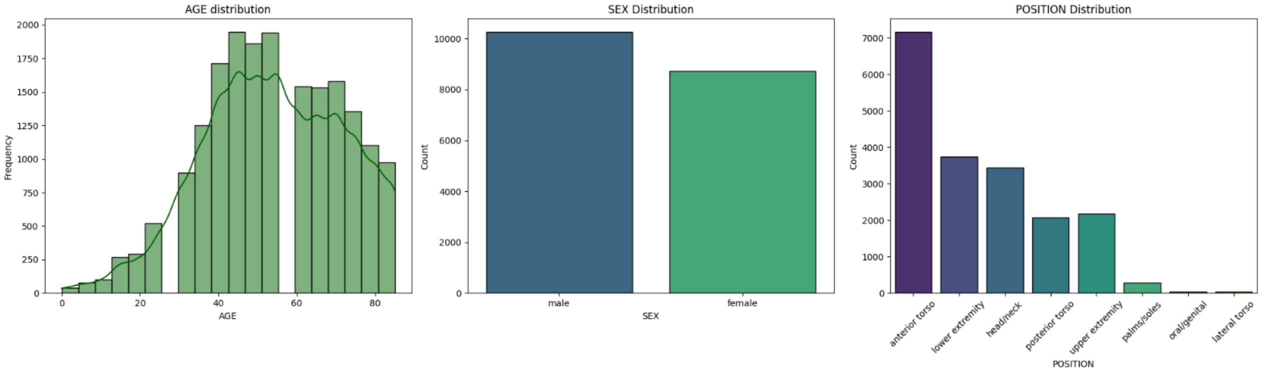


Figure 3 – Metadata distribution

2.2 Segmentation

- During the challenge I tried two different methods of segmentations which I will present here.

2.2.1 Deformable model - Active contour

- The first thing I thinked about is to use one of the segmentation methods we have seen during the module IMA-204 for medical imaging. Thus I tried the Active contour method or what we call "The snake".

However, Active contour models rely on several parameters that need to be tuned appropriately **for each image** or **class of images**. Finding the right combination of parameters can be challenging and may require manual intervention, which can be impractical when dealing with a large dataset as it is the case in our problem.

2.2.2 DeepLearning UNet Finetuning

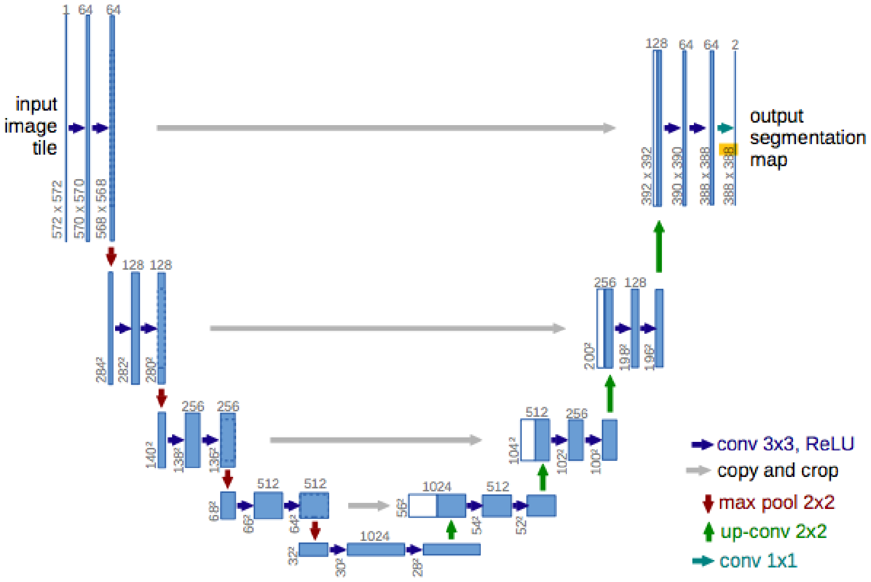


Figure 4 – U-Net architecture, source : Blent.AI

The **U-Net** model has gained widespread recognition as one of the best choices for image segmentation tasks due to several key strengths. First and foremost, its architecture features a U-shaped design with an encoder-decoder structure that allows for effective feature extraction and precise localization. This means that it can capture both global context and local details, which is crucial for accurate segmentation. Additionally, it is an architecture that employs skip connections that facilitate the seamless integration of low-level and high-level features, enhancing the model's ability to capture intricate patterns and boundaries in the data. It is also a versatile architecture and can then be adapted to our lesion segmentation problem.

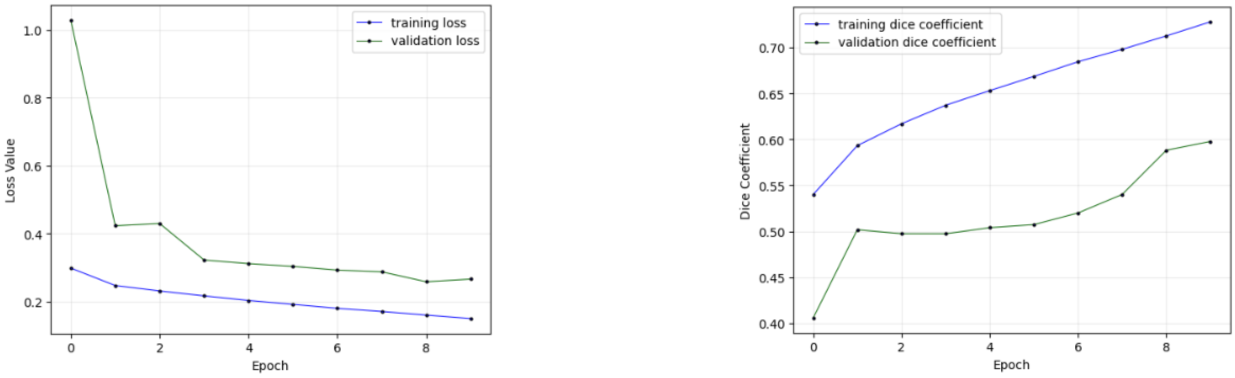


Figure 5 – Training/Validation : loss and Dice score

Here an example of the the performed segmentation using UNet :

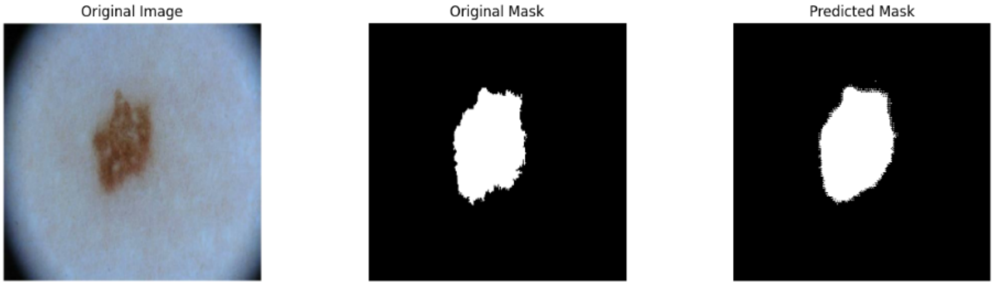


Figure 6 – Example of predicting masks in test data

2.3 Data Augmentation

We give below the distribution of our classes in train data.

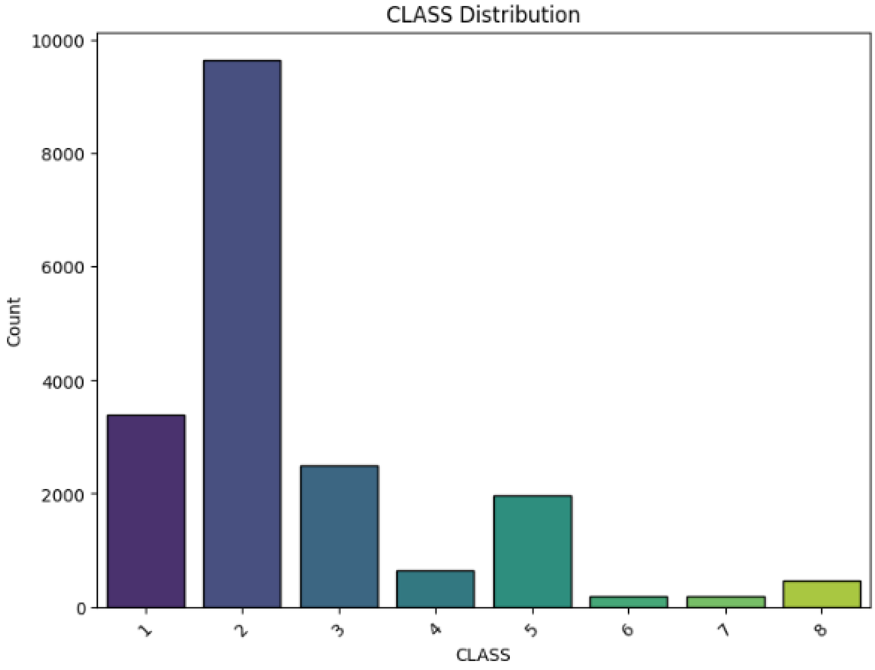


Figure 7 – Imbalance classification problem

We can see that we have a classification imbalance problem. In fact, a simple model may tend to choose the second class consistently and may still achieve a pretty good accuracy. That is why we will perform a data augmentation. And we note that the metric score of kaggle is using classes weights.

One may perform data augmentation by using basic geometric transformations or or some high performing methods like **GAN** on images, but it is not efficient as a method. A cleverer solution would be to perform data augmentation after features extraction and there will be no need to handle images.For this purpose we can use the random over-sampler of the library **Imblearn**.

2.4 Feature extraction

The given meta data is obviously not enough to decide between the eight features. After performing a segmentation model and a data augmentation we possess a large balanced complete image data. One may think to perform a CNN model applied to the ROIs of images (by multiplying the mask with the image), however a better performance can be obtained by using the metadata as well.

ABCD feature extraction is one of the process to extract the important feature. The results of this process are used to distinguish between the different classes. There are four important global features i.e. **Asymmetry**, **Border** Irregularity, **Color** Variation, and **Diameter**. [1]

Another features can be used in order to better characterize the ROIs of our skin lesions are **GLCM texture features**. These features are bases on the co-occurrence matrix. In practice, the co-occurrence matrix is often used to extract textural features from an image. From this matrix, different statistical measures such as, homogeneity, correlation, entropy, etc., can be calculated to characterize the texture of the image and assist in object segmentation or pattern recognition. This is exactly what interests us because we are wishing to find the difference between textured images.

GLCM features can be obtained from the feature extractor of the **Pyradiomics** library, which is a direct implementation of radiomics technology implemented by : Joost JM van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper et Hugo JWL Aerts.

After computing all these features for every image we have, I stored them in a data frame in order to use it for training. The table below presents all theses features.

A	B	C	D	GLCM
ID, AREA, PERIMETER, CIRCULARITY, BULKINESS, SOLIDITY, ECCENTRICITY	GRAD_R_MEAN, GRAD_R_STD, GRAD_G_MEAN, GRAD_G_STD, GRAD_B_MEAN, GRAD_B_STD	AVERAGE_R, AVERAGE_G, AVERAGE_B, STDEV_R, STDEV_G, STDEV_B	DIAMETER	Autocorrelation, ClusterProminence, ClusterShade, ClusterTendency, Contrast, Correlation, DifferenceAverage, DifferenceEntropy, DifferenceVariance, Id, Idm, Idmn, Idn, Imc1, Imc2, InverseVariance, JointAverage, JointEnergy, JointEntropy, MCC, MaximumProbability, SumEntropy, SumSquares

3 Classification problem

3.1 Fine Tuning : EfficientNet

The first thing I tried to do was to learn from images without their segmentations and without metadata using CNNs. I tried multiple architectures from scratch, but the results were not satisfactory. This is normal since it's a really complex problem, and using small CNNs will not help to learn from the ROIs of skin lesions (as images are not segmented). So, always with the goal of achieving good classification without too much preprocessing and/or manual feature extraction, I chose to use fine-tuning techniques. Fine-tuning involves using the layers of a large model that has been trained on images of the same field. That's why I used EfficientNet, which is one of the largest CNN models that has been trained on a large amount of medical images from ImageNet.

This method does not use the maximum of given information as it does not use the meta data neither the the segmentations. However, it is still a consistent model as it allowed me to have more than **67%** of test accuracy on kaggle. (0.69 in public score and 0.67 in private score).

3.2 RandomForest Classifier

After completing all preprocessing steps, I ended up with a substantial dataset containing a sufficient number of features. This turned the problem into a balanced classification challenge. To tackle it, I initially opted for either a Decision Tree or a RandomForest model. However, RandomForest models can be black-box in nature, computationally demanding, memory-intensive, and may have limited extrapolation capabilities beyond such a large training dataset, factors that I didn't carefully consider at the time. Consequently, the RandomForest model failed to yield the desired results, and the outcomes did not show any improvement.

In fact, In the context of a skin lesion classification problem, opting for Random Forest (RF) over Multilayer Perceptron (MLP) or Support Vector Machines (SVM) can be a strategic decision. RF's ensemble learning approach is advantageous, blending multiple models to enhance accuracy and mitigate overfitting, crucial in medical image analysis. Moreover, RF's ability to assess feature importance aids in understanding which aspects (like texture, color, or shape) are most influential in classification, adding interpretability to the model's decisions—an essential factor in medical applications.

The deviation I got from expected performance was abnormal because the expectation was to achieve good results ; unfortunately, due to time constraints before the deadline, I couldn't thoroughly investigate what went wrong.

4 Références

1. Article:Computational Radiomics System to Decode the Radiographic Phenotype
2. IEEE Xplore Digital Library
3. https://www.sciencedirect.com/science/article/pii/S0933365712001108#bib0180
4. https://www.sciencedirect.com/science/article/pii/S0933365713001589
5. https://www.nature.com/articles/nature21056
6. https://hal-univ-bourgogne.archives-ouvertes.fr/hal-01250955/document