OncoVision

Rocco Iuliano
r.iuliano13@studenti.unisa.it

Simone Della Porta
s.dellaporta6@studenti.unisa.it

ACM Reference Format:
Rocco Iuliano and Simone Della Porta. 2023. OncoVision. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 PROJECT CONTEXT

Addressing health-related issues represents the next frontier of artificial intelligence. Regarding the healthcare context, several attempts have already been conducted to develop machine learning solutions, genetic algorithms and others. For instance, in the Aravind Eye Care System in India, ophthalmologists and computer scientists are working together to test and deploy an automated image classification system to screen millions of retinal photographs of diabetic patients. Since the mid-twentieth century, researchers have proposed and developed many clinical decision-support systems to help physicians. Rule-based approaches were proposed in the 1970s that allow us to:

(1) diagnose diseases;
(2) choose appropriate treatments;
(3) provide interpretations of clinical reasoning;
(4) assist physicians in generating diagnostic hypotheses in complex patient cases.

It seems a good approach but it presents several problems, such as:

(1) it is costly to build;
(2) it requires explicit expressions of decision rules and require human-authored updates;
(3) it is difficult to encode higher-order interactions among different pieces of knowledge authored by different experts;
(4) the performance of the systems is comprehensible only by a medical knowledge;
(5) it was difficult to implement a system that integrates deterministic and probabilistic reasoning to narrow down relevant clinical context, prioritize diagnostic hypotheses, and recommend therapy.

Instead, recent AI research has leveraged machine learning methods, which can account for complex interactions, to identify patterns from the data. Moreover, machine-learning methods enable the development of AI applications that facilitate the discovery of previously unrecognized patterns in the data without the necessity to specify decision rules for each specific task [6]. Therefore, the main goal of an AI system in Medicine is to help the physicians in the clinical process, namely detect something, cluster the patient on some criteria, etc. On the other hand, AI systems pose new risks for medicine. Failures in medical AI could erode public trust in Healthcare. For example, bias in AI can deliver erroneous medical evaluations. Moreover, AI models magnify existing cyber-security risks, potentially threatening patient privacy and confidentiality. In general, in a software project, we need three types of expert groups:

- the group developing the algorithm;
- a group of validators;
- the operational staff.

These groups are also needed in the healthcare sector to overcome the following 3 key challenges in AI:

- Conceptual challenges in formulating a problem that AI can solve;
- Technical challenges in implementing an AI solution;
- Humanistic challenges regarding AI's social and ethical implications.

Furthermore, the incapability to address these challenges could erode public trust in medical AI, which could in turn undermine trust in Healthcare institutions themselves. In the end, unlike physicians, AI cannot draw upon "common sense" or "clinical intuition" but using good data and the correct AI model configuration, we can obtain a good performance [4]. In this project, we focused on skin cancer, in particular, the melanoma. Melanoma is a malignancy of melanocytes, which are pigment-producing cells of neuroectodermal origin that can be found throughout the body (including in the skin, iris and rectum). The cutaneous form of the disease is common in the Western world causing the majority (75%) of deaths related to skin cancer; indeed, its global incidence is 15–25 per 100,000 individuals. Survival rates in patients with melanoma (cumulative for all forms) have shown huge differences between countries in Europe, ranging between <50% in Eastern Europe to >90% in northern and central Europe for 5-year survival after primary diagnosis [5]. Figure 1 shows the incidence and mortality of cutaneous melanoma in the world. In according with Istituto Superiore di Sanità, the general melanoma diagnosis process is the following: the patient will conduct a dermoscopic examination and if the doctor has some suspects, he takes a skin sample in order to conduct histological analysis and then formulate the diagnosis. The biggest problem with this cancer type is a lack of early detection that does not allow therapy to treat the disease on time. That is why we decided to develop a melanoma detection AI system to help physicians with this disease. We have compared the performance obtained by our model with the proposed one by Di Biasi et al [3].

1.1 Related project

Di Biasi et al. proposed a system that combines Genetic Algorithms (GAs) with Convolutional Neural Networks (CNNs) to detect melanoma. They used GA for improving the architecture of CNN. Indeed, they defined a population of neural networks (NN) that are codified in vectors where each vector element represents a

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

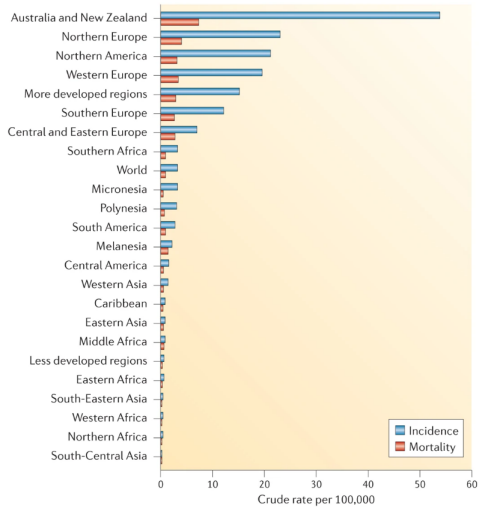


Figure 1: Incidence and mortality of cutaneous melanoma[5]

type of layer or pre-processing routine. Whereas for the hyper-parameters of each individual, they fixed them to these values: *'sgdm', 'MaxEpochs', 16, 'MiniBatchSize', 12, 'Shuffle', 'every-epoch', 'InitialLe-arnRate', 0.0001, 'ExecutionEnvironment', 'auto'.*
The authors defined the following constraints to the Genetic Algorithm in order to obtain a correct population of NN at each evolution step:

- The first gene of each entity must be an image input (II) or one of the pre-processing routines;
- If the gene q is a pre-processing routine, then the gene q + 1 must be a II layer or another pre-processing layer;
- The latest gene of an entity must be a classification layer.

Furthermore, they forced the possible values that the genes of an individual can assume in this range: *Convolution, ReLu, Cross Channel Normalization, Max Pooling Grouped Convolution, Fully Connected Layer, Dropout and Softmax.*
The authors stopped the GA after 100 evolutional steps or if no accuracy improvement was detected for ten consecutive evolution steps. In conclusion, their dataset contains skin images divided in two classes, melanoma (positive class) and moles (negative class). The number of instances for the positive class is 70 and 100 for the negative one.

2 PROJECT GOALS

In this work we will reply to the following research questions:

RQ$_1$: *What are the limitations of the selected approach?*

We decided to investigate this aspect to understand problems that afflict the selected approach to develop a better model. The results of this **RQ** influence the input of **RQ$_3$** because we are going to try to define a model that is less affected by these problems.

RQ$_2$: *What are the most useful data types and what are the data problems in the melanoma domain?*

We decided to investigate this aspect to understand which are the relevant data types to train the model for melanoma detection and which are the problems with the data. The results of this **RQ** influence **RQ$_4$** because, based on data problems, we can identify the techniques to improve their quality. Moreover, this **RQ** influences **RQ$_1$** because we can understand if the detected problems are also in the existing approach.

RQ$_3$: *How to improve melanoma detection model performance?*

We decided to investigate this aspect because, based on the other **RQs**' results, we can understand the changes to apply to the model to improve performance.

RQ$_4$: *How to raise the data quality?*

We decided to investigate this aspect to understand which are the techniques for improving data quality to highlight relevant features that the model will learn for making more precise predictions.

Based on the previous research questions, we have defined the following project goals:

(1) Conduct a detailed investigation of the baseline approach selected from the literature to understand the performance of the approach and its limitation;
(2) Understand the problems related to the datasets, namely lack of relevant features, few samples, low data quality, etc;
(3) Definition of an AI pipeline that might be used for cancer detection and is not affected by the problems which the baseline approach selected from the literature suffers.

Our project is available on GitHub at this link: https://github.com/Rocco000/OncoVision

3 AI MODEL SELECTION

In this study, we decided to use Convolutional Neural Network (CNN) because it represents various benefits. Before all, we must first explain its features and how it works. CNN is a classic artificial neural network (ANN) with input, hidden and output layers. The input layer propagates the input in the network, instead, the output layer provides the model prediction. Each layer can contain one or more neurons linked with some neurons of the previous layer. Moreover, each link has an associated weight. Each neuron calculates its activation as the weighted sum of its inputs and then adds a bias value. After which, it applies an activation function to the obtained value. CNN has an important peculiarity which is the convolutional layer. How affirmed by Kun-Hsing Yu et al [6], this layer type is useful for extracting spatial or temporal relations in an image and allows us to summarize and transform clusters of image pixels to extract high-level features. Moreover, according with Albawi et al [1], CNN has these benefits:

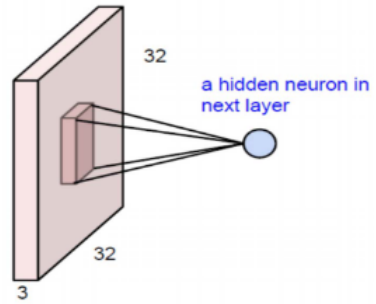


Figure 2: Example of how CNN hidden layer's neuron is responsible for a local region in the picture

(1) The assumption about problems that are solved by CNN is that the problems have no spatially dependent features, therefore, it provides an opportunity to detect and recognize features regardless of their positions in the image. For example, in face detection, we do not need to pay attention to where the faces are located in the images;
(2) Each layer can detect a particular feature when input propagates toward the network. For example, the first layer detects the edges, the second detect shapes, etc;
(3) It can reduce the number of parameters to improve the computational complexity of the training process.

To understand the third point we can use an image detection problem as an example. Consider an image 32x32x3 as input. If we use classic ANN, for connecting the input layer with a hidden layer's neuron we need 32x32x3 weight connections and this number must be multiplied by the number of neurons in the hidden layer. This amount of parameters is required only to connect the input layer with the first hidden layer's neurons! Instead, in the convolutional neural network, each neuron in the hidden layer is responsible for an image local region as shown in Figure 2. It means that a neuron of the next layer can only get input from the corresponding part of the previous layer (i.e. the neuron is connected only with specific neurons of the previous layer). For example, it can only be connected to 5×5 neurons. Therefore, if we want 32×32 neurons in the next layer, we will have 5×5×3 by 32x32 connections, which is 76,800 connections compared with 3,145,728 in ANN. Another assumption for simplification; is to keep the local connection weights fixed for the entire neurons of the next layer. This will connect the neighbour neurons in the next layer with exactly the same weight to the local region of the previous layer. With this assumption, the layers are similar to a window of 5×5×3 sliding on input neurons and mapping the generated output to the corresponding place. For this reason, this type of layer is named convolutional and it looks like a filter in image processing. Therefore, each layer can be associated with different filters. The network architecture and the configuration of the training algorithm hyperparameters of our project have been defined by the genetic algorithm output.

3.1 PEAS Specification

The problem that we have treated has the PEAS specification that is shown in Table 1.

Table 1: PEAS Specification

Performance	For evaluating our model, we used these metrics: accuracy, recall, precision and F1-score. In particular, we will focus on precision and recall because we need to minimize the number of false positives and false negatives.
Environment	Our model operates in an environment that has these features: (1) **Fully observable**: because the sensors provide our model full environment state; (2) **Static**: because the environment doesn't change after an action that our model did or during the time; (3) **Discrete**: because the environment limits the perceptions and actions of our model; (4) **Episodic**: because an action that our model did, is independent of the previous actions; (5) **Single-agent**
Actuators	The model interacts with the environment through the standard output of the machine
Sensors	The model can receive the environment input through the standard input of the machine

4 METHODICAL STEPS

Based on the goals set defined in Section 2, the methodical steps that we have conducted to address it are:

- Define a survey for physicians to understand data problems, data types that are useful for defining a good model and how to measure data quality. We conducted this step by using Google Forms;
- Investigate some image processing techniques that are useful to improve data quality;
- Search one or more new datasets in order to train and test our models. We conducted these researches by using Kaggle and ISIC Archive;
- Develop a Convolutional Neural Network for melanoma detection tasks;
- Develop two genetic algorithms (GAs) that produce a population of Convolutional Neural Networks (CNNs) that are optimized from two points of view:
  - The goal of the first one is to improve the training algorithm hyperparameters of the base model. In this paper,we will call it ***GA1***;
  - The goal of the second one is to improve both training algorithm hyperparameters and network architecture. In this paper,we will call it ***GA2***.

  At the end of each genetic algorithm, we selected the best individual in the last obtained population. For developing these alternatives we used PyGAD;

- Test the obtained models on the related project dataset to verify their performances;
- Test the obtained models on similar diseases to verify their behavior;
- Understand why the models have done a certain prediction. To conduct this step we used an explainability library called grad-cam;
- Verify the obtained models behavior on real-time images.

To coordinate the project development, we used Trello to apply the Kanban model. Moreover, to track all the experiments, we used MLflow through DagsHub. In the end, we compared the results obtained by the models and selected the best one. Figure 3 summarizes the methodical steps that we conducted.

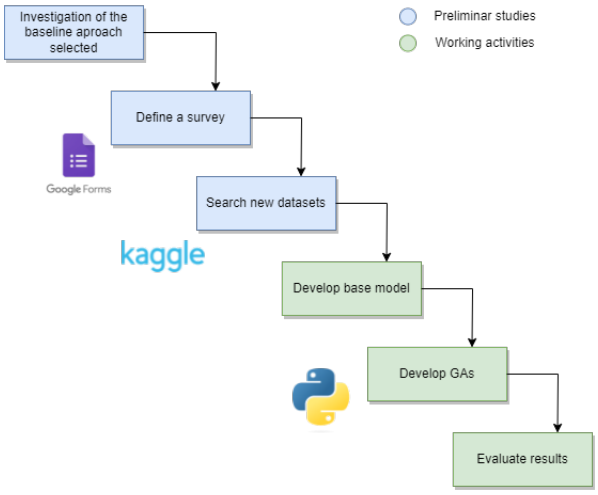


Figure 3: Methodical steps

5 SURVEY

The survey has two goals: the first is that it can guide us to select useful data for training the model and the techniques for improving image quality. The second one is to understand the problems that the melanoma diagnosis process suffers and, as a consequence, the problems that the selected approach suffers. The physicians are involved by sharing the survey's Google link on the Prolific platform to involve as many as possible doctors. The survey's questions are:

(1) $\mathbf{Q}_1$: What is your role?
(2) $\mathbf{Q}_2$: How many job experience years do you have?
(3) $\mathbf{Q}_3$: How much experience do you have in dermatology?
(4) $\mathbf{Q}_4$: How much does dermoscopic image affect melanoma diagnosis? The possible response is a value between one and five where one represents "little impact" and five represents "impactful";
(5) $\mathbf{Q}_5$: How much does histological image affect melanoma diagnosis? The possible response is a value between one and five where one represents "little impact" and five represents "impactful";
(6) $\mathbf{Q}_6$: If only the dermoscopic image is available, how reliable is the diagnosis? The possible response is a value between one and five where one represents "unreliable" and five represents "very reliable";
(7) $\mathbf{Q}_7$: If only the histological image is available, how reliable is the diagnosis? The possible response is a value between one and five where one represents "unreliable" and five represents "very reliable";
(8) $\mathbf{Q}_8$: If I have both the histological and dermoscopic images available, how reliable is the diagnosis? The possible response is a value between one and five where one represents "unreliable" and five represents "very reliable";
(9) $\mathbf{Q}_9$: What are the problems that can be encountered when acquiring the dermoscopic image? The possible choices are: little lighting, presence of hair, movement of the patient, incorrect position of dermatoscope and other;
(10) $\mathbf{Q}_{10}$: What are the problems that can be encountered when acquiring the histological image? The possible choices are: poor preparation of skin tissue, noise in the skin tissue, non-uniform sectioning and other;
(11) $\mathbf{Q}_{11}$: What are the useful metrics to evaluate a dermoscopic image quality? The possible choices are: Signal-to-noise ratio, Contrast-to-noise ratio, image sharpness and other;
(12) $\mathbf{Q}_{12}$: What are the relevant dermoscopic image features on which you focus more? It can be replied through a text box;
(13) $\mathbf{Q}_{13}$: What are the relevant histological image features on which you focus more? It can be replied through a text box;
(14) $\mathbf{Q}_{14}$: Do you think there are other relevant issues in melanoma diagnosis that the survey did not treat? It can be replied through a text box.

The questions gave us the following informations:

- $\mathbf{Q}_1$ and $\mathbf{Q}_2$ allowed us to understand the role and experience years of the participants.
- $\mathbf{Q}_3$ and $\mathbf{Q}_4$ allowed us to understand the most relevant data type.
- $\mathbf{Q}_5$, $\mathbf{Q}_6$, and $\mathbf{Q}_7$ allowed us to understand how much is reliable the prediction having a certain type of data.
- $\mathbf{Q}_8$ and $\mathbf{Q}_9$ allowed us to understand the problems that occur during data acquisition.
- $\mathbf{Q}_{10}$ allowed us to understand the useful metrics for evaluating data quality.
- $\mathbf{Q}_{11}$ and $\mathbf{Q}_{12}$ allowed us to understand the most relevant data features.
- $\mathbf{Q}_{13}$ allowed us to understand if we had not considered some relevant aspects.

5.1 Survey results

Sharing the survey, we collected 66 answers on Google Forms. Then, we conducted a pre-screening to remove the non-valid responses, namely, the people who have no experience in dermatology or are not doctors. Therefore, we obtained 49 valid answers and we evaluated them using these techniques:

- The survey results with an integer value have been analyzed with statistical techniques, namely mean, median, standard deviation and the box plot of the distribution;
- The survey results with a text value have been analyzed with thematic analysis to find common patterns.

The following results were obtained:

- $\mathbf{Q}_1$: As shown in Figure 4a, the most frequent employments of the participants are "doctor" and "dermatologist";
- $\mathbf{Q}_2$: As shown in Figure 4b, the participants have an average of 7 job experience years;
- $\mathbf{Q}_3$: As shown in Figure 4c, the participants have a good experience in dermatology;
- $\mathbf{Q}_4$: As shown in Figure 5a, the dermoscopic images are moderately considerable for diagnosis;
- $\mathbf{Q}_5$: we have found a high relevance of histological images on diagnosis, indeed; we obtained a lot of answers with a value equal to 5 on the Likert scale;
- $\mathbf{Q}_6$: As shown in Figure 5b, the diagnosis is moderately reliable; if only the dermoscopic image is available;
- $\mathbf{Q}_7$: we have found that the diagnosis has high reliability, although only the histological image is available. Indeed we obtained a lot of answers with a value equal to 4 or 5 on the Likert scale;
- $\mathbf{Q}_8$: As shown in Figure 5c, the diagnosis is more reliable; if both image types are available;
- $\mathbf{Q}_9$: As shown in Figure 6a, the frequent issues encountered in dermoscopic images are the incorrect angle of the dermatoscope, the presence of body hair, little lighting, and patient movement;
- $\mathbf{Q}_{10}$: we have found that the frequent issues encountered during the acquisition of histological images are: poor tissue preparation, non-uniform cutting, noise in the tissue;
- $\mathbf{Q}_{11}$: As shown in Figure 6b, the generally used metrics to evaluate the dermoscopic images are SNR, CNR, and image brightness;
- $\mathbf{Q}_{12}$: we have found that the relevant dermoscopic image features are: colour, asymmetry, pigment networks, boundary irregularities, and ABCDE principles;
- $\mathbf{Q}_{13}$: we have found that the relevant histological image features are: nests of melanocytes, cytological atypia, and lymphovascular invasion.

Based on these results, we can confirm that the dermoscopic images are less considerable than the histological images for diagnosis. Moreover, the diagnosis is moderately reliable; if only the dermoscopic image is available. To the issues highlighted by question Q9, we must measure the dermoscopic images' quality.

6 DATA PIPELINE

6.1 Data Collection

On **Kaggle**, we have found several datasets related to the skin cancer problem. We chose two datasets:

- The first one is available at this link. There are 10015 images classified into more category groups. Moreover, there are two folders that contain all the images and *csv* file named *HAM10000_metadata* that report the matching between the image ID and the related category of problem. The categories are:
  - **akiec**, Actinic keratoses and intraepithelial carcinoma / Bowen's disease. There are 327 samples of this category in the entire dataset;
  - **bcc**, basal cell carcinoma. There are 514 samples of this category in the entire dataset;
  - **bkl**, benign keratosis-like lesions such as solar lentigines / seborrheic keratoses and lichen-planus-like keratoses. There are 1099 samples of this category in the entire dataset;
  - **df**, dermatofibroma. There are 115 samples of this category in the entire dataset;
  - **mel**, melanoma. There are 1113 samples of this category in the entire dataset;
  - **nv**, melanocytic nevi. There are 6705 samples of this category in the entire dataset;
  - **vasc**, vascular lesions such as angiomas, angiokeratomas, pyogenic granulomas and hemorrhage. There are 142 samples of this category in the entire dataset.
- The second one available at this link has 2750 images divided into three folders: ***train***, ***test***, ***valid***. In each folder, there are three sub-folders called ***melanoma***, ***nevus*** and ***seborrheic_keratosis.***

The first operation was to reorganize the images into folders:

- For the first dataset we had different classes, therefore, in this case, we had to collapse these different categories into fewer classes. We did this after conducting some research to verify if each class is benign or malignant. In particular, we collapsed the classes in this way:
  - **bkl**, **df**, ***nv*** and ***vasc*** classes into ***benign*** folder
  - **bcc** class in ***carcinoma*** folder
  - **akiec** class in ***bowenDisease*** folder
  - **mel** class in ***melanoma*** folder

  We used *benign* and *melanoma* folders to build our dataset. Instead, *carcinoma* and *bowenDisease* folders were used for in vivo testing of the models.
- For the second dataset, we extracted the contents of the *train*, *test* and *valid* folders. We then moved the images contained in the seborrheic_keratosis folder into the benign folder.

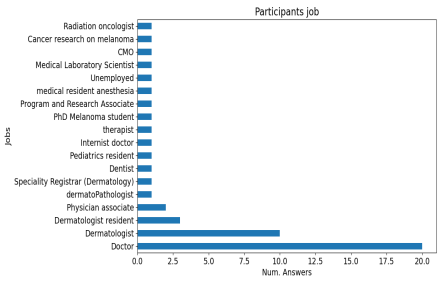
After these steps, we unified the two datasets into one called Final-Dataset. It contains two folders named benign and melanoma. At this point, we noticed that we had 10289 negative images and only 1634 positive images. So we needed to have more positive images. To increase the number of positives we extracted melanoma images from ISIC challenges datasets. In particular, we used ISIC 2018, ISIC 2019 and ISIC 2020 challenge datasets available at this link. For ISIC 2018 challenge, we found 6 datasets: train, test and validation set for tasks 1-2 and train, test and validation set for task 3. We extracted melanoma images only from 5 of these because the training dataset for task 3 corresponds with our first dataset. At the end of this operation, we obtained an overall of 5515 positive samples.

6.2 Data Preparation

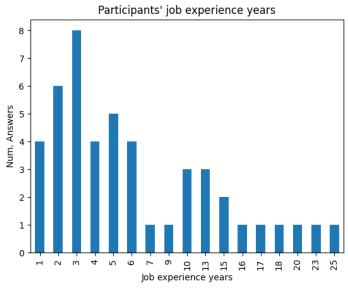
After the previous step, we had a look at the data collected and we noticed that there are a lot of images with black borders and some other types of noise like pieces of colored cloth. So we decided to hand crop the images presenting these kinds of noise to prevent the model from learning irrelevant features. Once cropped images, we resized these to 450x600.

6.3 Data Quality

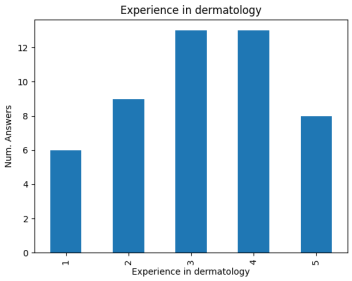
Based on survey results, we computed these metrics on images:



(a) Employment of the participants

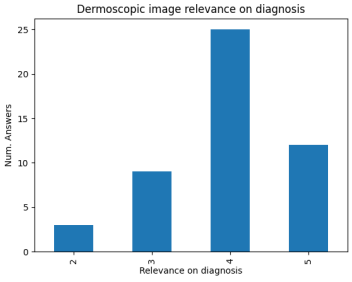


(b) Job experience of participants

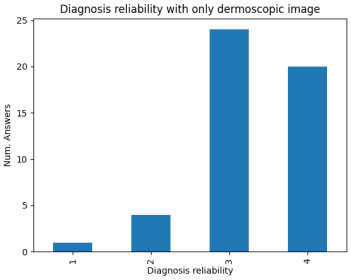


(c) Experience in dermatology

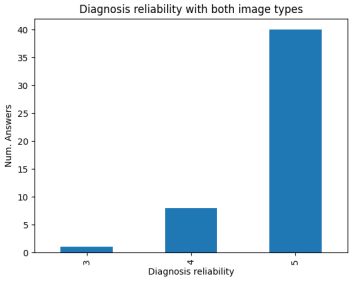
Figure 4: Participants background



(a) Dermoscopic image relevance on the diagnosis

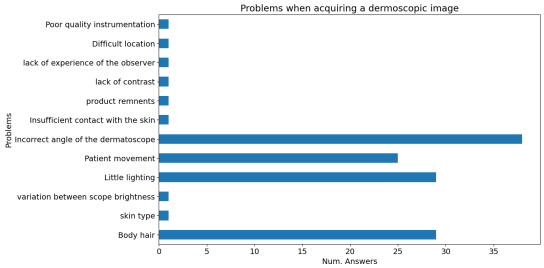


(b) Diagnosis reliability (only dermoscopic image)

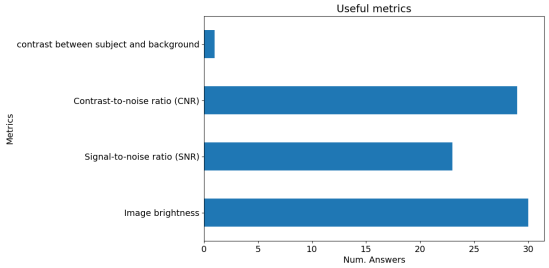


(c) Diagnosis reliability (both image types)

Figure 5: Dermoscopic images features



(a) Issues in dermoscopic images



(b) Helpful metrics to evaluate the images

Figure 6: Dermoscopic images quality

- **Sharpness**, measures the clarity level of detail of an image. We computed it with the following formula:

$$SH = \sum_{x,y} \frac{LP(x,y)}{\mu_{xy}} \quad , \quad LP(x,y) = \frac{\delta^2 L}{\delta x^2} + \frac{\delta^2 L}{\delta y^2}$$

  where μ_{xy} denotes the average luminance around pixel (x, y) The results achieved were shown in Figure 7a;
- **Brightness**, refers to the overall lightness or darkness of an image. It is a perceptual attribute that describes the amount of light or luminance present in an image. We calculated the brightness of the image as the *average of the V channel of the image converted to HSV format.* After calculating the average we normalized it in the range [0,1]. The results achieved were shown in Figure 7b;
- **Signal to Noise Ratio** (**SNR**) is a measure of the image signal in a given region to the background. We computed it with the following formula:

$$SNR = \frac{\mu_{sig}}{\sigma_{sig}}$$

where μ_{sig} is the average signal of the region of interest and σ_{sig} is the noise of the region of interest. The results achieved were shown in Figure 7c;

- **Contrast to Noise Ratio** (**CNR**) in a medical image is a measure of the contrast between the tissue of interest and the background. We computed it with the following formula:

$$CNR = \frac{|S_{roi} - S_{background}|}{\sigma_o}$$

where S_{roi} is the average signal of the region of interest, $S_{background}$ is the average signal of background and σ_o is the noise of the region of interest. The results achieved were shown in Figure 7d;

- **Image contrast** refers to the difference in brightness, color, or grayscale intensity between different parts of an image. It determines the level of distinction and separation between objects or areas within the image. We computed it with the following formula:

$$Contast = |S_{roi} - S_{background}|$$

where S_{roi} is the average signal of the region of interest, $S_{background}$ is the average signal of background. The results achieved were shown in Figure 7e;

- **RMS contrast**, also known as **Root Mean Square Contrast**, is a measure of contrast used to quantify the average difference in intensity between adjacent pixels in an image. We computed it with the following formula:

$$\sqrt{\frac{1}{MN}\sum_{i=0}^{N-1}\sum_{j=0}^{M-1}(I_{ij} - \bar{I})^2}$$

where I_{ij} is the intensity of i-th j-th pixel of an image MxN, $\bar{I}$ is the average intensity of all pixel values. The image I has its pixel values normalized in the range $[0, 1]$ therefore, the values for this metric will be in the range $[0, 1]$. The results achieved were shown in Figure 7f.

During training time, we randomly improved the sharpness and contrast of the images with a probability of p, using *RandomAdjustSharpness()* and *RandomAutocontrast()* of *torchvision.transforms*.

6.4 Data splitting

After data pre-processing, we divided the dataset into three sets, namely train, test, and valid. We allocated 70%, 20% and 10% respectively. Then, we plotted the instances number of each class to each set in order to visualize the sum of samples collected. The plots are shown in Figure 8.

6.5 Data balancing

As shown in Figure 8a there is a wide gap between the two classes in the train set; therefore, we applied the undersampling technique to balance the majority class with the minority class by randomly removing the majority instances. To preserve the train set representativeness of reality, we decided to keep a small difference between the two classes, indeed we have a difference of 100 instances as shown in Figure 8b.

7 BASE MODEL

As mentioned in Section 3, we developed a convolutional neural network to melanoma detection. Figure 9 shows our model architecture and the configuration of its hyper-parameters is as follows:

lr=0.001, batch size=64, epoch=64, optimizer=Adam

Since the last network layer has two neurons, we used the CrossEntropyLoss function to obtain the model prediction and loss value during training. Moreover, we weighed the loss function (during the training step only), weighting the two classes (0.4 to the negative class and 0.6 to the positive class) to tell the model to "pay more attention" to samples from the under-represented class (the melanoma class).

8 GA FOR MODEL IMPROVEMENT

Genetic Algorithms are based on biological evolution (i.e. Darwins's theory of evolution) and they are research algorithms for solving optimization problems. GA is capable of evolving a population of individuals iteratively. Each step produces a new generation of individuals improved over a fitness function. It repeats this improvement until a stop criterion is not verified. There are three operations that allow us to enhance the individuals, namely selection, crossover and mutation. Each individual is composed by m features called **genes**.

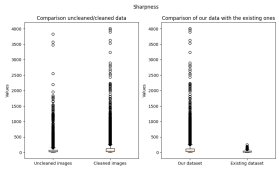
As shown in Figure 10 the lifecycle of GA is based on six phases:

(1) **Initialization**: in this phase, an initial population is defined, in most cases, randomly;
(2) **Selection**: in this phase, a set of parents has chosen from the previous generation based on fitness values. Those individuals are going to compose the new generation and they are going to be used to generate new individuals.
(3) **Crossover**: this phase represents the genetic operator used to combine the genetic information of two parents to generate new individuals;
(4) **Mutation**: this phase represents the genetic operator used to maintain genetic diversity, just like in biological mutation;
(5) **Evaluation**: in this phase, the modified population is assessed against the fitness function;
(6) **Replacement**: in this phase, the obtained population replaces the old one and a new iteration starts; Steps 2-6 are iterated until a stop criterion is not verified.

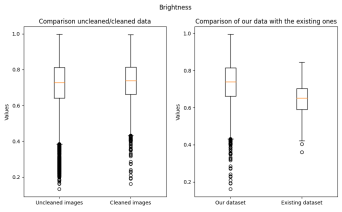
According to *Bochinski et al.*[2], the optimal choice of hyperparameters is one of the major challenges when applying CNN-based methods. Furthermore, since a broad range of CNN network structures yields good results, it stays unclear if a current solution obtains optimal or near-optimal configurations or if different structures can improve previous results. Genetic Algorithms can deal with these challenging conditions. That is why to improve the architecture and hyperparameter configuration of the model we used Genetic Algorithms (GA).

8.1 Genetic Algorithms Configuration

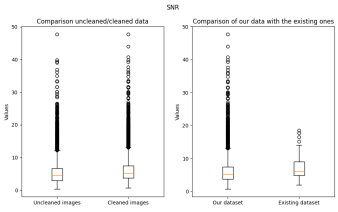
As mentioned in Section 4, we developed two genetic algorithms focused on different aspects, namely hyperparameters of the training algorithm and network architecture. We defined the same objective function, fitness function, stop criterion and initial population size



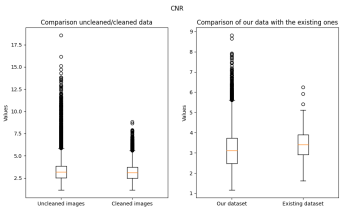
(a) Sharpness boxplot



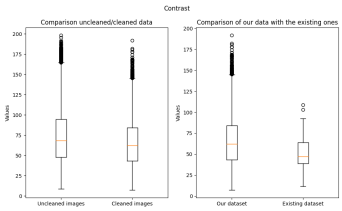
(b) Brightness boxplot



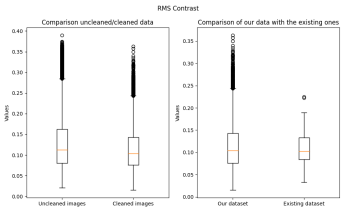
(c) SNR boxplot



(d) CNR boxplot

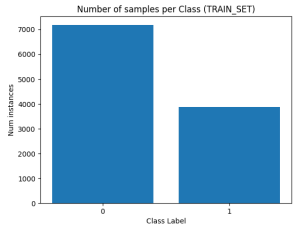


(e) Contrast boxplot

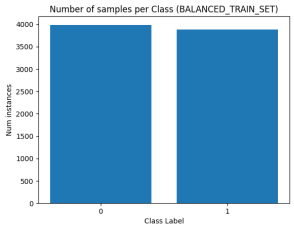


(f) RMS Contrast boxplot

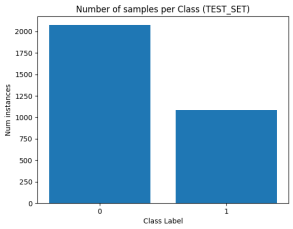
Figure 7: Image metrics



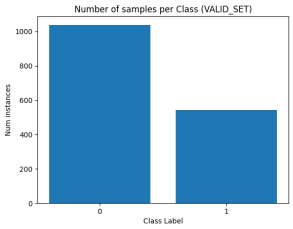
(a) Train plot



(b) Balanced train plot



(c) Test plot



(d) Valid plot

Figure 8: Plot of sets

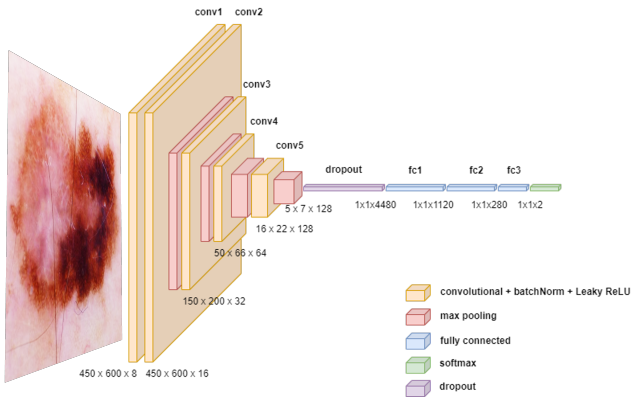


Figure 9: Base model architecture

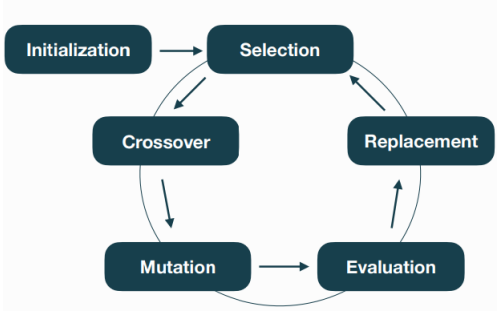


Figure 10: GA lifecycle

for both GA versions. The **objective function** is defined as follows:

$$f(s) = max\ w_1 * accuracy_s + w_2 * precision_s + w_3 * recall_s$$

w_i is a weighting factor that determines the trade-off between accuracy, precision and recall; it is a value between 0 and 1. We defined these values as follows $w_1 = 0.20$, $w_2 = 0.35$, $w_3 = 0.45$. We gave more weight to recall and precision because we need to reduce the number of false negatives (indeed, high recall indicates a low number of false negatives) and false positives to create a capable model to predict each class correctly. The **evaluation function** is equal to the objective function, instead the **fitness function** is

defined as follows:

$$fitness(s) = \frac{f(s)}{\sum_{s_i \in P-\{s\}} f(s_i)}$$

where P is the individuals set and s is an individual. In this way, we could evaluate an individual based on the evaluation of the entire population. The **initial population** consists of 8 random solutions. The **selection technique** that we used for all GA versions is K-way Tournament Selection, where $K = 4$ and $M = 4$. Therefore, we will execute the tournament four times to select $M = 4$ solutions. Moreover, at each tournament, $K = 4$ solutions are involved. The **mutation technique** for all GA versions is Random Resetting and the **stop criterion** is to stop the genetic algorithm after 10 generations or if there is not a fitness function improvement after 6 consecutive steps. To guarantee the AI models' uniqueness in the same generation, we defined a function to replace the clones in the population. The replacement technique is a simple random generator that iteratively generates a new solution until it is not in the actual population. Moreover, all solutions of both GA versions have been trained and tested on the same sets.

8.1.1 ***GA1***.

The goal of this version is to find an optimal training algorithm configuration of the base model. Therefore, we defined a set of individuals that represent a possible training algorithm configuration. We defined each solution as a vector $S = \{C_1, ..., C_m\}$ of m genes ($m = 4$) and the gene position in the genome represents a certain hyperparameter. The hyperparameters that we chose to optimize are:

(1) **Learning rate**: it refers to the step size used during optimization to update the model weights. Therefore, it represents how much model weights must be modified to improve model performance. The range that we chose for this parameter is between 0.001 and 0.010. This parameter is represented by the first gene in the genome;
(2) **Number of epochs**: it represents the number of epochs to train the model. An epoch is an iteration over the entire dataset. The values that it can assume are: {64,96,128}. This parameter has represented by the second gene in the genome;
(3) **Batch size**: it represents the number of samples in a mini-batch and the values that it can assume are: {32, 64}. It has represented by the third gene in the genome;
(4) **Optimizer type**: it represents the optimizer algorithm used to improve the model parameters to enhance the model performance. We chose this optimizer type:
    - **Adam**: it adapts the learning rate based on the historical gradients and the second moment of the gradients. In this way, it helps in achieving faster convergence;
    - **Adadelta**: it also adapts the learning rate but keeps a running average of recent squared gradients and uses this average to compute the learning rate by taking into account only a local neighborhood of the parameters;
    - **NAdam**: it combine Adam and Nesterov momentum. The gradient used to update the parameters is not the current gradient but an estimation based on the lookahead parameters. NAdam adjusts the lookahead parameters by considering both the momentum term and the gradient calculated using Adam.

    It has represented by the fourth gene in the genome.

Hyperparameters which are not on the list have been set with the default value. The **crossover technique** that we used in this version is the One-Point Crossover. As previously mentioned, all solutions of this genetic algorithm version represent the training algorithm hyperparameters for the base model defined in Section 7.

8.1.2 ***GA2***.

The goal of this version is to find an optimal CNN network structure and an optimal training algorithm configuration. Therefore, we defined a set of individuals that represent both the network architecture and hyperparameters. We defined each individual as a vector $S = \{H_1, H_2, H_3, H_4, L_1, ..., L_m\}$ of $m + 4$ genes ($m = 16$). The first four genes represent the same hyperparameters of the first GA version, instead, the other m genes represent network layers. For each solution, we did not represent the first and the last network layer because we fixed them. The first one is a convolutional layer 2D (8 filters of size 3x3, stride=1, padding=1, bias=True), instead; the second one is the output layer for binary classification (nn.Sequential(Dropout, Linear, Linear)). The genome genes that regard the network architecture can be one of these layer types:

(1) **Convolutoinal layer (Conv2D)**: it can be a convolutional layer with 128, 64, 32, 16 or 8 filters with a size of 3x3. Moreover, we fixed the stride and padding to 1 and the bias parameter to "*True*". Finally, we computed the input channels of these layers with a custom function;
(2) **Pooling**: this layer types down-sampling the input representation to reduce the complexity for further layers. We chose two pooling types: Max Pooling 2D or Average Pooling 2D. Both the pooling layers can have a filter size of 3x3 or 2x2;
(3) **Dropout2d**: it randomly zero out entire channels with a frequency rate (p) at each step during training time. Each channel will be zeroed out independently on every forward call. This helps to prevent overfitting;
(4) **BatchNorm2d**: it applies Batch Normalization on its input to prevent exploding or vanishing gradient issues and potentially reducing overfitting;
(5) **Activation layer**: this layer type represents an activation function and we chose two activation types:
    - **ReLU**: it is defined as follows: $ReLU(x) = max(0, x)$
    - **Leaky ReLU**: it is a variation of the ReLU function and resolves the dying ReLU issue, namely, the neurons can output a zero value. Therefore, they will have no effect on model prediction. The function is defined as follows: $LeakyReLU_\alpha(x) = max(\alpha \cdot x, x)$

The **crossover technique** that we used in this version is the Two-Point Crossover. Moreover, to guarantee the AI models validity, we called a function at each generation to replace the incorrect solutions in the population. The replacement technique is a simple random generator that iteratively generates a new solution until it is not valid (examples of no valid solutions: a BatchNorm layer that does not follow a convolutional layer, an activation layer that does not follow a convolutional or BatchNorm layer, or if there is four consecutive convolutional layers due to physic limits of cloud

provider). In conclusion, we weighed the loss function as in the base model.

9 PRELIMINARY RESULTS AND FINDINGS

9.1 Base model results

Figure 11 shows the evaluation metrics trend obtained by the base model on validation set, and the best values was achived at 29th training step and are the follows:

accuracy = 0.80, precision = 0.66, recall = 0.89, F1 = 0.76

Instead, the evaluation metrics on the test set are shown below and the confusion matrix is shown in Figure 14a:

accuracy = 0.82, precision = 0.70, recall = 0.81, F1 = 0.75

9.2 Genetich algorithms results

9.2.1 *GA1 results*.

The best solution of the GA1 was achived at the 5th generation and its configuration is the follows:

lr = 0.00381, batch size = 32, epoch = 64, optimizer = Adadelta

During the execution of the GA, we tracked the new solutions rate discovered by the genetic algorithm at each generation, as shown in Figure 13a. Moreover, we monitored the fitness function and gene trend during the GA generations. Figure 13b shows the fitness function trend, instead, the gene trend is available at this link.

9.2.2 *GA2 results*.

The best solution of the GA2 was achived at the initial population and its hyperparameters configuration is the follows:

lr = 0.00577, batch size = 32, epoch = 64, optimizer = Adadelta

Instead, the Figure 12 shows its architecture. As we done in GA1, we tracked in GA2 the new solutions rate discovered by the genetic algorithm at each generation, as shown in Figure 13c and monitored the fitness function and genes trend, as shown in Figure 13d. Also in this case, the gene trend is available at this link.

9.3 Results on Test set

The evaluation metrics of base model, GA1 best solution and GA2 best solution are shown in Table 2, instead the confusion matrices are shown in Figure 14.

Table 2: Models evaluation metrics on test set

model	accuracy	precision	recall	F1
base model	0.82	0.70	0.81	0.75
GA1	0.81	0.68	0.86	0.76
GA2	0.80	0.66	0.86	0.75

9.4 In-Vivo results

As mentioned in Section 4, we tested the models on the related project dataset and on diseases similar to the Melanoma, namely Bowen disease and Carcinoma, in order to apply an In-Vivo experiment. This approach helped us to understand the models quality. Therefore, the evaluation metrics obtained by the models are shown in Table 3 and the confusion matrices are shown in Figure 15. Moreover, the predictions made by the base model on similar diseases are shown in Figures 16a and 16b. Instead, the GA1 predictions made on similar diseases are shown in Figures 16c and 16d. Finally, the best GA2 solution produced the results shown in Figures 16e and 16f on Carcinoma and Bowen instances.

Table 3: Models evaluation metrics on related project dataset

model	accuracy	precision	recall	F1
base model	0.46	0.42	0.77	0.54
GA1	0.44	0.40	0.77	0.53
GA2	0.45	0.42	0.84	0.56

9.5 Explainability Analysis

For each model, we selected ten random instances from each set (test, Carcinoma, Bowen, and related project set). In five of these instances, the model made incorrect predictions, whereas in the other five cases, the model made correct predictions. Figure 17 shows the predictions on some test examples, the others are available on the GitHub repo.

10 DATA ANALYSIS - ADDRESSING THE RQS

The methodical steps described in Section 4 allowed us to address the research questions, namely:

- The survey helped us to address the following research questions: RQ1, RQ2 and RQ4;
- The image processing techniques helped us to improve our model performance. These techniques have been chosen based on the survey results and helped us to address the research questions RQ3 and RQ4;
- The new found datasets, helped us to understand the limit of the chosen approach, therefore to address the research question RQ1;
- Genetic algorithms helped us to improve the architecture and the configuration of the training algorithm hyperparameters, therefore to address the research question RQ3.

10.1 Addressing RQ_1

The small number of samples in the dataset on which the AI model can be trained is one of the problems encountered in the related project. As a result, the model does not have enough samples to learn the useful pattern that can be used in prediction time. Indeed, the dataset has 70 positive samples and 100 negative samples. As highlighted in the Data Quality section, the second problem encountered with the associated project dataset is the poor quality of the samples, which are low in brightness and contrast. Moreover, they do not involve the hyperparameters configuration in the genetic algorithm, which plays a crucial role in defining the best model and allows us to converge faster during training. Finally, due to the unavailability of the related project model, we could not compare the evaluation metrics of our model with theirs but we can affirm that relying solely on improving accuracy in the genetic algorithm can result in defining a bad model. This is because accuracy alone does not consider the balance between positive and negative samples. In fact, it is possible to create an AI model that predicts the

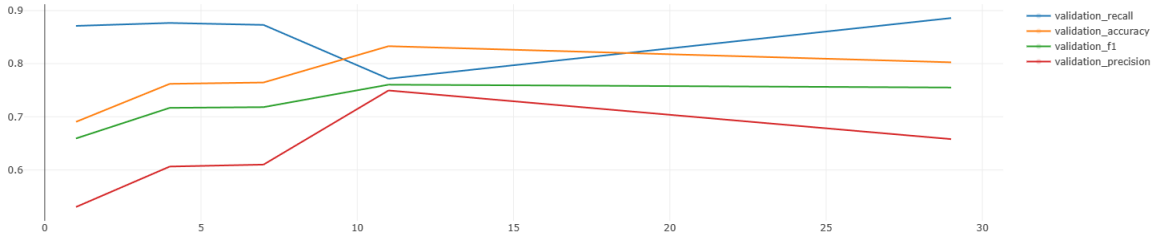


Figure 11: Evaluation metrics trend on validation set by base model

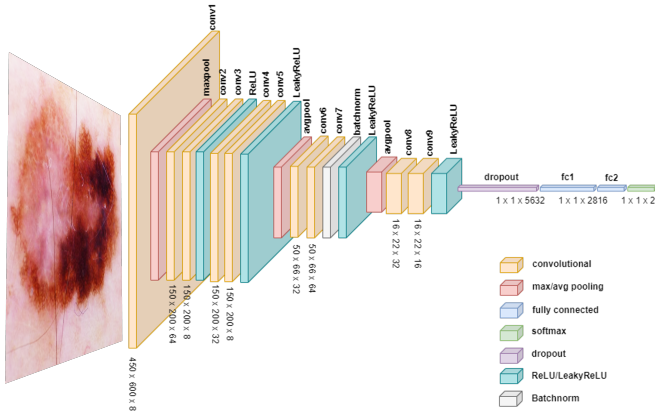


Figure 12: GA2 architecture

negative samples very well but still makes a lot of errors on the positive samples.

10.2 Addressing RQ_2

As highlighted by the survey, histological images or both histological and dermoscopic images are required for each patient to make a good diagnosis. Furthermore, as shown in questions Q_9 and Q_{10}, the most common problems observed with skin images are the incorrect angle of dermatoscope, patient movement, body hair and poor lighting for dermoscopic images and non-uniform cutting of the tissue, noise in the tissue (air bubble, crease, etc.) and poor tissue preparation for histological images. In fact, these problems were found both in the related project dataset and in ours.

10.3 Addressing RQ_3

To improve the performance of AI models for skin cancer detection, we acted on the various pipeline processes. In particular, we have:

- Improved data quality in the data preprocessing phase;
- Used genetic algorithms to optimize an objective function that aims to improve the evaluation metrics of models. Particularly, precision and recall, as they are medical data, we do not want to have a high false positive rate given the importance of the disease and therefore do not have a high false negative rate;
- Weighted loss function in the training phase so that the model pays more attention to minority class instances.

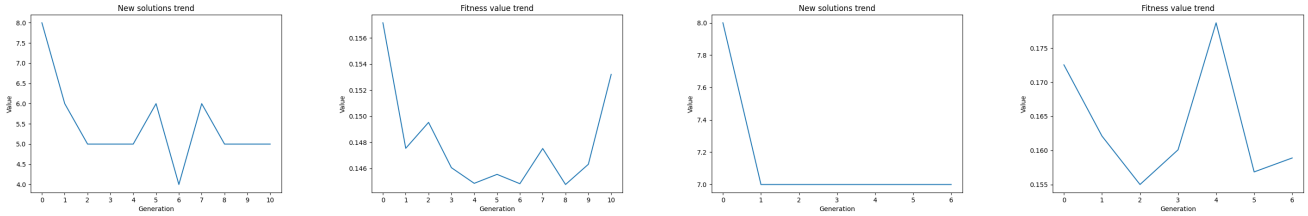
10.4 Addressing RQ_4

To improve data quality, we acted on the data preprocessing phase. We have removed black borders from images as it made them very noisy and would affect the model predictions. Figure 7 shows that we have achieved an increase in metrics compared to the uncleaned data, but we have achieved a decrease in image contrast due to the black borders remotion. Furthermore, the collected data has low sharpness, and during the training phase, we randomly improved sharpness and contrast using PyTorch techniques described in Section 6.3 to improve it.

11 IMPLICATIONS OF THE RESULTS

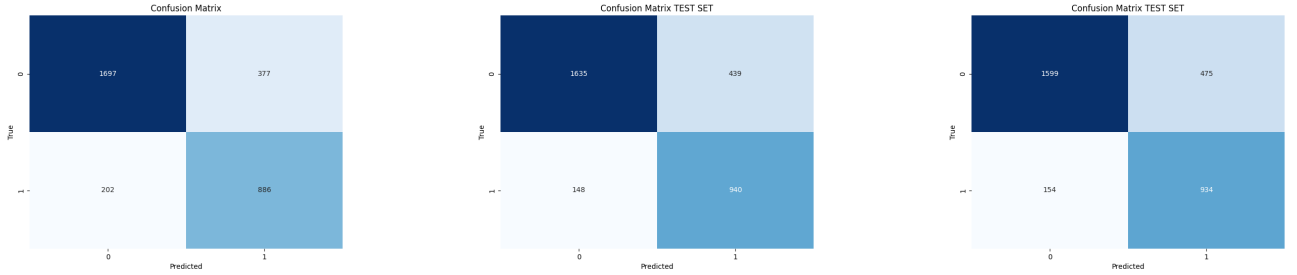
As mentioned in Section 5.1, we can confirm that the common problems that can occur in dermoscopic images were also found in the collected data. Moreover, as shown in Figure 7, the data preprocessing step allowed us to remove the noise from the images, thus improving the data quality. As highlighted in the data quality step, the quality of our dataset is higher than the corresponding project dataset on more metrics.
Based on the survey results, we cannot have much confidence in the model predictions if we only have the dermoscopic images for the melanoma detection task (3/5 on the Likert scale Q_6) therefore, we also need the histological image for each dataset sample to achieve a high reliability of model outcomes.
As mentioned in Section 9, we can observe:

- The base model was able to reach good results despite the default hyperparameters configuration;
- The GA1 has reached the 10 iterations and obtained the best solution at the 5th generation. Figure 13b shows the fitness trend over the generations of the genetic algorithm and we can see that:
  - The curve initially declines due to the population improvement;



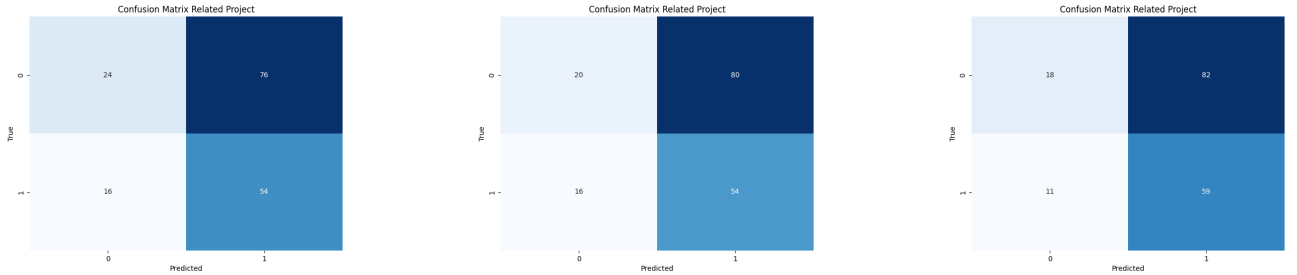
(a) GA1 new solutions trend **(b) GA1 fitness trend** **(c) GA2 new solutions trend** **(d) GA2 fitness trend**

Figure 13: GAs trends



(a) Base model confusion matrix on test set **(b) GA1 confusion matrix on test set** **(c) GA2 confusion matrix on test set**

Figure 14: Confusion matrices on test set



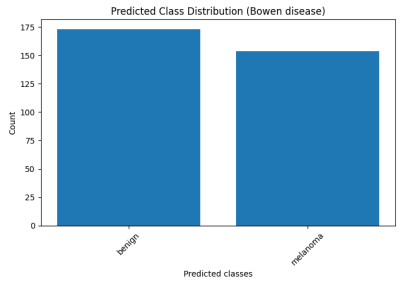
(a) Base model confusion matrix on related project dataset **(b) GA1 confusion matrix on related project dataset** **(c) GA2 confusion matrix on related project dataset**

Figure 15: Confusion matrices on related project dataset

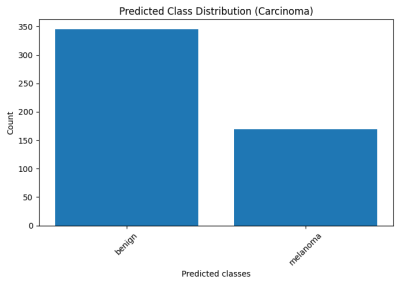
  - In the central part of the curve, we can see a peak that represents the best solution found by the GA1 that survives until the 10 generations;
  - In the last part, the curve growth is due to the deterioration of the population with respect to the best solution.
- GA2 did not reach the 10th iteration due to premature convergence indeed, it obtained the best solution in the initial population and was stopped at the 6th generation. The fitness function trend shown in Figure 13d provides further confirmation, namely:
  - The curve initially improves for the first two generations;
  - In the middle part of the curve, we can see a deterioration of the population because the peak reached in the 4th generation represents the fitness value of the best solution, which is higher than its initial value. This shows how much the population has worsened;
  - In the last part, we can see a small improvement in fitness, but only for one generation.

  Moreover, during the execution of GA2, we noticed a high rate of invalid solutions which were randomly replaced. This may be a good reason to explain why the GA2 did not reach a good solution due to the high number of random substitutions. Indeed, the confusion matrices show a high false positive rate.

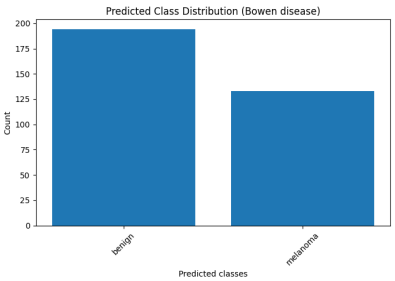
Based on the results obtained on the related project dataset, we can affirm that the base model and GA1 achieved a similar behavior;



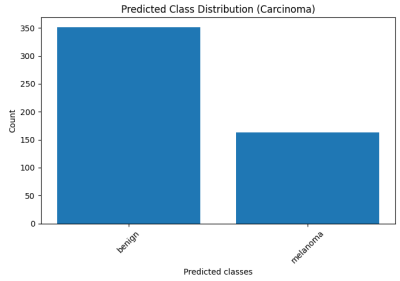
(a) Base model predictions on Bowen instances



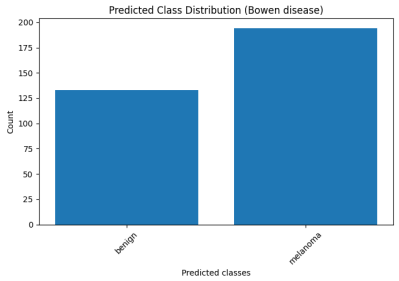
(b) Base model predictions on Carcinoma instances



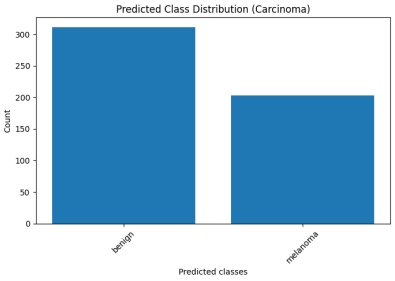
(c) GA1 predictions on Bowen instances



(d) GA1 predictions on Carcinoma instances

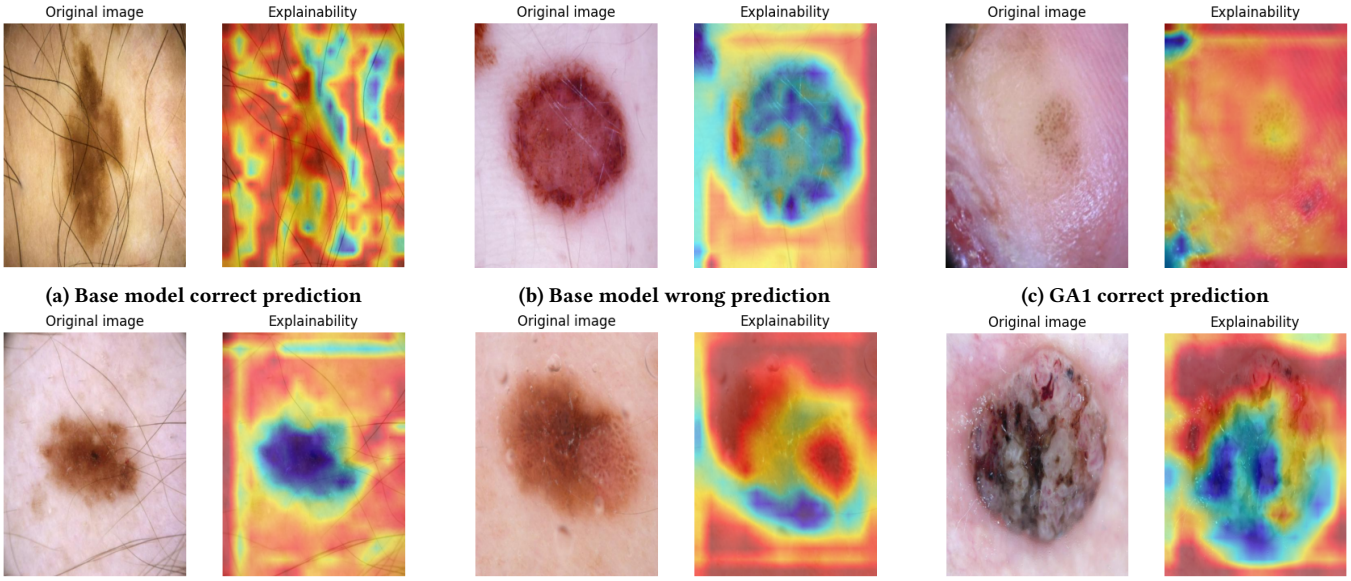


(e) GA2 predictions on Bowen instances



(f) GA2 predictions on Carcinoma instances

Figure 16: Models predictions on similar diseases



(a) Base model correct prediction

(b) Base model wrong prediction

(c) GA1 correct prediction

(d) GA1 wrong prediction

(e) GA2 correct prediction

(f) GA2 wrong prediction

Figure 17: Explainability on some predictions

instead, GA2, despite having a high recall, had a higher false positive rate. In general, all three models obtained a high false positive rate, which may be due to poor lighting and low contrast of images, which makes it more difficult to detect the stain. Instead, based on the obtained results on Carcinoma and Bowen instances, we can affirm that the base model achieved better results than the other two models despite GA2 having predicted the positive class many times.

Therefore, based on the results obtained by the models on the test set and related project dataset, we can affirm that the best solution is the output of the GA1 because it represents a good trade-off between accuracy, precision and recall. Indeed, as shown in Figure 14b, the GA1 best solution made the smallest number of false negatives, while on the corresponding project dataset obtained the same results as the base model but made a few more errors on false positives. Moreover, the Explainability section shows that all models show abnormal behaviour on certain samples, both on the test, Carcinoma, Bowen and related project dataset. Indeed, as shown in Figure 17, during the prediction, the models did not focus on the region of interest but on the image background.

12 CONCLUSION

Based on the previous sections, we can affirm that the collected data has some of the common problems highlighted by the survey, but the pre-processing step helped us to achieve a higher data quality than the related project dataset. Looking at the data collected, we can suppose that our models may be affected by fairness issues due to the lack of samples of different ethnicities, as well as in the corresponding project dataset. All the obtained models have achieved an accuracy value close to the maximum value of the related project model. Moreover, the first genetic algorithm provided the solution with a good trade-off between accuracy, precision and recall. Based on the results obtained by the genetic algorithms, we can observe that GA2 obtained a premature convergence most likely due to the high randomness substitution rate, whereas GA1 achieved its best solution in the 5th generation. In addition, all the models obtained showed abnormal behaviour on certain samples, i.e. they did not focus on the region of interest but on the image background. Based on the previous discussion, the feature work will be:

(1) Collect new data, especially samples of different ethnicities, to address potential fairness issues;
(2) Increase the population size and the number of generations of genetic algorithms to explore the search space further. This increases the probability of finding a better solution;
(3) Change the substitution techniques in genetic algorithms to reduce the randomness;
(4) Use a cloud platform with more powerful hardware components to speed up the training time and satisfy the second point on the list;
(5) Consider energy consumption during training to improve the sustainability of the model, using tools such as CodeCarbon.

REFERENCES

[1] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. 2017. Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET)*. Ieee, 1–6.
[2] Erik Bochinski, Tobias Senst, and Thomas Sikora. 2017. Hyper-parameter optimization for convolutional neural network committees based on evolutionary algorithms. In *2017 IEEE international conference on image processing (ICIP)*. IEEE, 3924–3928.
[3] Luigi Di Biasi, Alessia Auriemma Citarella, Fabiola De Marco, Michele Risi, Genoveffa Tortora, and Stefano Piotto. 2023. Exploration of Genetic Algorithms and CNN for Melanoma Classification. In *Artificial Life and Evolutionary Computation: 15th Italian Workshop, WIVACE 2021, Winterthur, Switzerland, September 15–17, 2021, Revised Selected Papers*. Springer, 135–138.
[4] Thomas P Quinn, Manisha Senadeera, Stephan Jacobs, Simon Coghlan, and Vuong Le. 2020. *Trust and medical AI: the challenges we face and the expertise needed to overcome them.* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7973477/
[5] Dirk Schadendorf, David E Fisher, Claus Garbe, Jeffrey E Gershenwald, Jean-Jacques Grob, Allan Halpern, Meenhard Herlyn, Michael A Marchetti, Grant McArthur, Antoni Ribas, et al. 2015. Melanoma. *Nature reviews Disease primers* 1, 1 (2015), 1–20.
[6] Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. 2018. Artificial intelligence in healthcare. *Nature biomedical engineering* 2, 10 (2018), 719–731.