Skin Lesion Classification
Télécom Paris - Challenge IMA 205

Giovanni BENEDETTI DA ROSA
May 5, 2024

I. INTRODUCTION

Detecting and classifying skin lesions is crucial in diagnosing various dermatological diseases, including skin cancer, which poses significant public health challenges globally [1]. Computer-aided diagnosis (CAD) systems have emerged as promising tools in this domain, offering efficient and accurate methods for lesion classification. Integrating machine learning algorithms with medical imaging techniques, CAD systems can provide to clinicians valuable decision support, aiding in timely diagnosis and treatment planning.

In this context, this report presents the development and results obtained for Skin lesion classification in the final Challenge of the discipline IMA205. The main goal of this challenge was 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

The provided dataset was generated based on the ISIC(International Skin Imaging Collaboration) dataset, and contains 25,331 samples, divided between train, validation and test set. The training-validation set comprises 18,998 images, which account for 75% of the total dataset. The remaining 6,333 images make up the test set.

The test set was further split into two subsets: 60% for the public ranking and the remaining 40% for the private ranking.

Besides the images themselves, numerical and descriptive data regarding the patients were provided. From now on, these will be referred to as metadata. These include information such as age, gender, lesion location, etc.

II. DATA ANALYSIS AND PRE-PROCESSING

Using the metaData provided we can clearly see that the dataset is unbalanced as can be visualized in Figure 1.

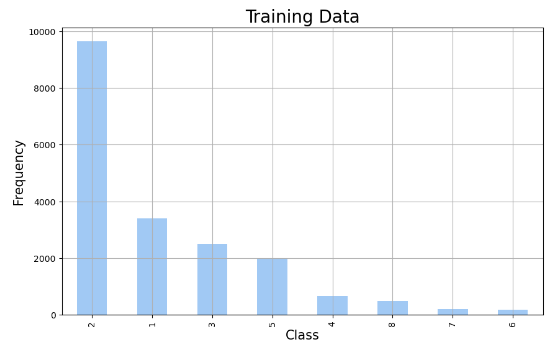

FIG. 1. Visualization of the unbalanced dataset.

Among the 18,998 samples in the training dataset of , 284 are missing SEX, 324 are missing AGE, and 1,970 are missing POSITION. In this sense, for the missing AGE values, it was decided to add the dataset's median. As for the gender and injury body position, one-hot encoding was performed, and columns with NaN values were deleted. An analogous procedure is also performed in the test set.

This type of data can be potentially useful in classifying the lesions, once it was shown in some studies that can exists a correlation[2].

III. METHODS

To identify and classify skin lesions using CAD systems, two approaches have been selected. The first involves segmenting the lesion and extracting appropriate features, which utilized for lesion classification. The second approach directly employs Convolutional Neural Network Architecture, enabling it to learn patterns directly from the images and classify them based on these patterns.

In the following sections, the methods chosen for each of the approaches are explained in more detail.

A. Features Based Classification

To identify the severity of the lesion and to classify the necessity of intervention, doctors generally use two methods: the ABCD criteria and the Glasgow 7-point checklist [5].

In the ABCD criteria, each letter corresponds to a key feature:

- A stands for Asymmetry;
- B stands for Border irregularity;
- C stands for Color variation;
- D stands for Diameter;

To identify this with a CAD system can be seen by the following schema: (i) image pre-processing; (ii) segmentation for lesion location and properties; (iii) feature extraction to characterize the lesion; and (iv) classification of features for decision-making[6].

1. Segmentation

In order to the computer identify where exactly is the lesion in a dermoscopic image first we need to be able to segment it.

In this work, two methods were applied, one using classical methods(otsu) and another one using a U-net.

The method applied to segment the images based on the otsu followed a similar workflow of a project that I and my colleague Paulo Moura have done in IMA201[7].

The procedure basically consists in removing the hair with an inpainting method and morphological operations after identifying it, and selecting a region where we have only the image of the skin and the lesion, excluding the black contours that generally appear in dermoscopic images, using a Conected Components Labelling function; After that the Otsu Method is applied.

For this specific dataset the method did not appear to be robust enough, once it was obtained 0.62 of Dice Score.

So, it was decided to move on to a neural network approach, because according to the literatures it would be able to achieves better scores once we could capture higher scores. The chosen model was a U-net, again because of promising results seen in the literarture[8], to the ISIC dataset.

The architecture of U-Net is characterized by its U-shaped layout, which can be seen in the image 2, and it consists of two encoder and decoder units.

Basically, the encoder consists of a series of convolutional layers followed by downsampling operations such as max-pooling.The decoder is designed to gradually recover the spatial information lost during the contracting path, using upsample techniques, with a lot of features chanels which allow the network to propagate context information to higher resolution layers [9].

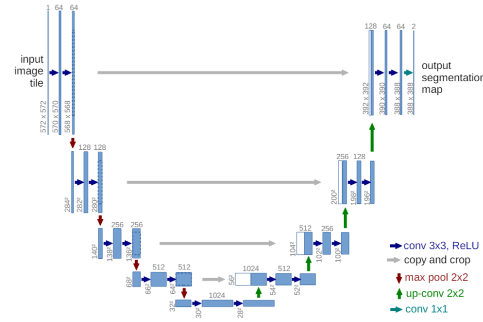

FIG. 2. U-Net Architecture[9]

After training the U-net, the segmentation results were much better, achieving 83% of Dice Score. The training process utilized a binary cross-entropy loss, an Adam optimizer with a $1e^{-5}$ of learning rate, and a scheduler to adjust the learning rate dynamically based on performance, with a factor of 0.1. The training loop iterated over 20 epochs, and some results of the segmentation can be seen in the image 3.

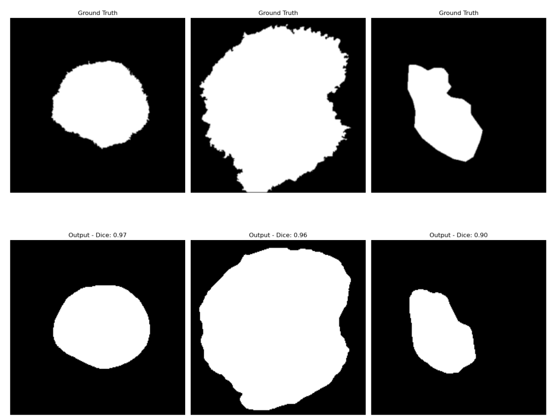

FIG. 3. Results of segmentation with U-net

2. Feature extraction

As already described a common procedure to extract the features in CAD systems is to use the segmentations to extract the ABCD features. The choice of the representative features of asymmetry, border, and color features was made based on the procedure described in Ganster, H., et al[10].

To capture dermatoscopic information, the procedure utilizes textures and local descriptors, inspired by [11]. The chosen features were: Gray level co-occurrence matrix (GLCM), Local binary pattern (LBP), Oriented FAST and rotated BRIEF (ORB) and Hu moment.

In this regard, it is important to clarify two things: firstly, it should be noted that some of the extracted features depend solely on the generated binary mask, while others rely on the application of these masks to the RGB images. Secondly, the quantity of extracted features along with their characteristics is arbitrarily determined and subject to fine-tuning. For example, in the case of ORB descriptors, we could have utilized all the key points of a given image, but we opted for only 32 Means, Standard Deviations and Responses. It's clear that adding features can potentially improve the performance of a given model, but at the same time can make the optimization problem harder to solve.

In the following table, it's possible to see the number of features in each category ABCD and how they were extracted.

Feature Type	**Features**
Asymmetry(A) and Diameter(D)	Area, Perimeter, Circularity, Bulkiness, Solidity, Eccentricity, Fractal Dimension, Hu Moments
Border(B)	Mean and std of the gradient in the border of the segmentation for each RGB channel
Color(C)	Mean and std of the RGB channel of the cropped image with the segmentation
Local Descriptor and Textures	Mean, std and response of 32 ORB descriptors; LBP Histogram values; Contrast, Correlation, Homogeneity

TABLE I. Extracted Features

3. Models to Classication

In the feature-based approach, using the features described in the previous section It was decided to test the performance of some machine learning methods. The chosen methods were Random Forest, Support Vector Machine(SVM) and Multilayer perceptron(MLP).

The Random Forest was chosen, because it was proven to have an excellent performance with non-linear data and its ability to handle overfitting when dealing with high-dimensional datasets, by creating uncorrelated bagged trees. Also, it's been proven to work well in particular to skin lesion classification problems [13].

SVM is selected due to its good generalization in managing high-dimensional spaces and its ability to establish non-linear decision boundaries using various kernel functions, that can potentially be useful in skin lesions as described in (Oliveira et al.)[14].

Similarly to the other two methods, we hope that a MLP capture complex patterns and relationships within high-dimensional feature spaces, and subsequently obtain a good classification.

For each one of these models the training process has passed by the following pipeline:

- **Data Splitting:** The dataset was split into training and validation sets, with 75% used for training and 25% for validation.

- **Data Scaling :** The extracted features are then scaled, ensuring that each feature has a standard deviation of 1 and a mean of 0.

- **Data Balancing:** As cited before the data is unbalanced, as there are much more samples from a specific class compared to the other ones. To solve this issue, the adaptive synthetic sampling approach (ADASYN) was used[12]. ADASYN calculates the class imbalance ratio and identifies the minority class instances. It then computes the feature differences between each minority instance and its nearest neighbors. Synthetic samples are generated by adding a random fraction of the computed feature differences to the original instances, ensuring diversity while capturing the underlying distribution of the minority class.

- **Hyperparameter Tuning:** In order to determine the best parameters for the data, it was decided to use a hyperparameter strategy.

B. Res-Net

Since its creation in 2015 by Microsft Researchers engineer's the Deep Learning Residual model(Res-Net)[3] has been widely used in many classification problems. The success of this model in accomplishing this task can be explained by the fact that from it the problem of vanishing gradient in the backpropagation has been solved, allowing the construction of increasingly deeper networks, useful in detecting important parameters for image classification.

The Residual Learning is based on the assumption that by adding again the input before applying the activation function can still be solved by the same layers. This assures that even if the gradient becomes very small as it passes through the layers it can still flow to the other layers.

Basically, the structure of the Res-Net is represented in the figure 4 and can be seen as:

-Convolutional Layers: The input image passes through a stack of convolutional layers, each followed by batch normalization and a ReLU activation function. These layers perform initial feature extraction. They are downsampled and passes by a lot of max pooling operations. -Shortcut: During the process of convolution, it's established some shortcuts, which turn the network into residual version [3]. – Output Layer: The network ends with a global average pooling layer and a 1000-way fully-connected layer with softmax activation function to generate a probability map for each class. It returns the the class with the highest probability. The total number of weighted layers is 34.

The Res-Net architecture was proven to achieve good results in skin Lesion Classification[4]. In this work, it was decided to use a Res-Net101, using the implementation defined in the Torch library with a pre-weights set.

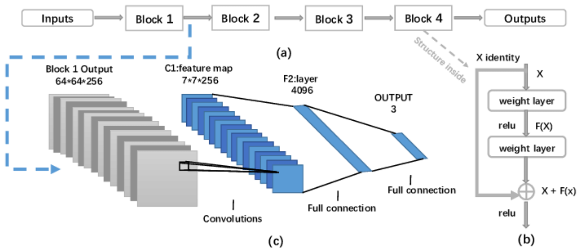

FIG. 4. Resnet architecture [15]

As expected all the images have different sizes. So, during the training process, the images were resized from its center to be squared (224×224), which was defined as described in the original Resnet paper. The used training loop was done with 20 epochs, and again an adaptive learning rate that started with a learning rate of $1e^{-5}$. In this training process, it was used an Adam optimizer and a cross entropy loss using the weights provided in the Kaggle, to try to handle better with the unbalanced class. Using the *Dataset*, of torch, it was decided to use 32 of batch size. In addition, it was decided to use earling stopping to prevent overfitting.

IV. RESULTS

During the challenge, several attempts following the processes described earlier with different hyperparameter values were made. In this regard, as expected, it became evident that the more complex model, ResNet, achieved better results. The performance of models using the extracted features was extremely similar.

Here again, it is worth noting that due to the nature of the challenge, we did not have access to the labels of the test data, and when submitting the possible predictions, we only had access to the accuracy relative to 40% of the test data. Therefore, the analysis below will be conducted considering all these obtained scores.

Method	Validation	Public Test	Private Test
SVM	0.64	0.38	0.36
Random Forest	0.62	0.40	0.38
MLP	0.57	0.38	0.36
Resnet	0.76	0.67	0.63

TABLE II. Accuracy results

As can be seen in the table II, the Resnet has been confirmed to be the best model, while the results of the features extracted models are comparable between themselves. It is also noticeable that there is a considerable decline in the results of the feature-based models compared to the validation results for the Kaggle submissions. This can be due to an overfitting problem not detected during the grid search.

The optimal configuration for random forests yielded a max depth of 16, min samples leaf of 2, min samples split of 2, and n estimators of 175. In the case of SVM, the best performing SVM model was configured with a C value of 2, gamma set to 'auto', and a radial basis function (RBF) kernel.

V. DICSUSSION

Analyzing the results, it seems clear that the feature based models did not capture well the patterns of the classification. This can be due to the following reasons:

- **Segmentation:** Despite achieving segmentation results of over 80% of average dice in some masks the overall performance was poor. This could lead to a cascading effect where unrealistic feature extraction negatively impacts the model's performance.
- **Adasyn:** The strategy solve to balance the classes maybe it was not the best option. As described in Beinecke and Heider[16], Adasyn can add some noise and thus introduces some variance to the synthetic data points generated from the k-nearest neighbors. Also, when dealing with Random Forest this problem can be amplified once we create a bunch of decision trees that have already high variance [17].
- **Extracted Features** Maybe the chosen features to be extracted were not enough representative for the machine learning methods be able to approximate well the relationship between the data and the classes. Instead of extracting extracting mean and stds, other representations of the measures could have been employed. Ganster, H., et al described more complex features that were not implemented in in this project.

With respect to the results obtained with ResNet, the drop in values from training to testing was much less abrupt, although no attempt was made to address the issue of imbalanced classes.

VI. CONCLUSION

This report addresses the challenge of skin lesion classification, crucial for diagnosing various dermatological diseases, including skin cancer, using computer-aided diagnosis (CAD) systems. Two approaches were explored: feature-based classification and ResNet-based classification. The feature-based approach involved lesion segmentation and feature extraction based on the ABCD criteria and texture descriptors, with machine learning models like Random Forest, Support Vector Machine (SVM), and Multilayer Perceptron (MLP) used for classification. However, the performance of these models was limited, likely due to challenges in segmentation accuracy, imbalanced data, and feature representativeness.

In contrast, ResNet101, a deep convolutional neural network (CNN), was employed for direct image classification. ResNet101, known for its effectiveness in handling deep networks, processed images resized to 224 $\times$ 224 pixels. Results indicated that ResNet101 outperformed the feature-based models, achieving the highest accuracy. However, there was a notable decline in the performance of feature-based models when tested on Kaggle submissions compared to validation results, suggesting potential generalization issues.

In conclusion, while both feature-based and deep learning approaches were explored, ResNet101 demonstrated superior performance. Future work could focus on improving segmentation accuracy, exploring more representative features, and addressing class imbalance to further enhance CAD systems for skin lesion classification.

VII. FUTURE IMPROVEMENTS

As mentioned previously, there are opportunities to enhance the accuracy of the models.

In the ResNet approach, optimizing certain network parameters such as batch size and epoch numbers, and exploring effective criteria for early stopping could be beneficial. Moreover, implementing an oversampling method for images, including techniques like rotation, flipping, translation, scaling, shear, noise addition, and adjusting brightness and contrast, could further elevate the model's performance by augmenting the dataset with artificial data. This type of approach has been proven to increase a lot the accuracy in (Perez et al) [18].

To address the challenges encountered with the feature-based approach, several strategies can be pursued. These include refining the segmentation method to enhance accuracy, using specific nets to skin lesions like MFSNet[19]. Also, revising the class balancing strategy, and exploring the generation of more sophisticated features to capture nuanced patterns in the data.

Another idea that could be implemented is to combine deep learning features with handcrafted features as it was done in (Ali et al.)[20]. In this way we could be able to capture the best of both worlds, interpretable features, and non-interperatble features that can increase the accuracy.

VIII. APPENDIX

During a final code review, it was found that the feature-based models had such a sudden drop because they were overfitting. This would explain the problems already reported, and it is obvious that the choice of hyperparameters should have been made in order to resolve this.

[1] National Cancer Institute (2018) Melanoma of the Skin - Cancer Stat Facts, SEER. Available at: https://seer.cancer.gov/statfacts/html/melan.html.

[2] Ribero, Simone, et al. "Effect of Age on Melanoma Risk, Prognosis and Treatment Response." Acta Dermato-Venereologica, vol. 98, no. 7, 11 July 2018, pp. 624–629, www.ncbi.nlm.nih.gov/pubmed/29648671, https://doi.org/10.2340/00015555-2944.

[3] He, Kaiming, et al. "Deep Residual Learning for Image Recognition." ArXiv.org, 10 Dec. 2015, arxiv.org/abs/1512.03385.

[4] "Dermoscopic Image Classification Method Using an Ensemble of Fine-Tuned Convolutional Neural Networks." Sensors, vol. 22, no. 11, 30 May 2022, p. 4147, https://doi.org/10.3390/s22114147.

[5] Korotkov, Konstantin, and Rafael Garcia. "Computerized Analysis of Pigmented Skin Lesions: A Review." Artificial Intelligence in Medicine, vol. 56, no. 2, Oct. 2012, pp. 69–90, https://doi.org/10.1016/j.artmed.2012.08.002.

[6] Mojdeh Rastgoo, et al. "Classification of Melanoma Lesions Using Sparse Coded Features and Random Forests." Proceedings of SPIE, 24 Mar. 2016, https://doi.org/10.1117/12.2216973. Accessed 29 Oct. 2023.

[7] GitHub, 2023, `https://github.com/giovanni-br/Telecom_Paris/blob/main/IMAGE%20PROCESSING/IMA201/Project/ReportISkinLesionSeg.ipynb`

[8] Liu, Lina, et al. Skin Lesion Segmentation Based on Improved U-Net. 5 May 2019, https://doi.org/10.1109/ccece.2019.8861848. Accessed 23 July 2023.

[9] Ronneberger, Olaf, et al. U-Net: Convolutional Networks for Biomedical Image Segmentation.

[10] Ganster, H., et al. "Automated Melanoma Recognition." IEEE Transactions on Medical Imaging, vol. 20, no. 3, Mar. 2001, pp. 233–239, https://doi.org/10.1109/42.918473.

[11] Bansal, Priti, et al. "Improving the Classification Accuracy of Melanoma Detection by Performing Feature Selection Using Binary Harris Hawks Optimization Algorithm." Soft Computing, vol. 26, no. 17, 21 June 2022, pp. 8163–8181, https://doi.org/10.1007/s00500-022-07234-1. Accessed 3 May 2024.

[12] Haibo He, et al. "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning." IEEE Xplore, 1 June 2008, `ieeexplore.ieee.org/abstract/document/4633969?casa_token=J_CENnbbg04AAAAA:H66LkaQgQseWdiBmYNy3Puy0nrHpFfZ7OA3Io7ZXVSCE-0_bXw-pmblGkrE7HoIISMjkQqG7Ng`.

[13] Rastgoo, Mojdeh, et al. "Classifica-Tion of Melanoma Lesions Using Sparse Coded Features and Random Forests." 2016.

[14] Oliveira, Roberta B., et al. "A Computational Approach for Detecting Pigmented Skin Lesions in Macroscopic Images." Expert Systems with Applications, vol. 61, Nov. 2016, pp. 53–63,

https://doi.org/10.1016/j.eswa.2016.05.017. Accessed in 04 May 2024.

[15] Tao, Siyan, et al. Highly Efficient Follicular Segmentation in Thyroid Cytopathological Whole Slide Image. 1 Jan. 2020, pp. 149–157. Accessed 4 May 2024.

[16] Beinecke, Jacqueline, and Dominik Heider. “Gaussian Noise Up-Sampling Is Better Suited than SMOTE and ADASYN for Clinical Decision Making.” BioData Mining, vol. 14, no. 1, 29 Nov. 2021, https://doi.org/10.1186/s13040-021-00283-6.

[17] Brandt, Jakob, and Emil Lanzén. A Comparative Review of SMOTE and ADASYN in Imbalanced Data Classification.

[18] —. “Data Augmentation for Skin Lesion Analysis.” Lecture Notes in Computer Science, 2018, pp. 303–311, https://doi.org/10.1007/978-3-030-01201-$4_3$3.

[19] Hritam Basak, Rohit Kundu, and Ram Sarkar. Mfsnet: A multi focus segmentation network for skin lesion segmentation. Pattern Recognition, page 108673, 2022.

[20] Ali, Redha, et al. “Skin Lesion Segmentation and Classification for ISIC 2018 by Combining Deep CNN and Handcrafted Features.” ArXiv:1908.05730 [Cs, Eess, Stat], 13 Aug. 2019, arxiv.org/abs/1908.05730.