Projects in Data Science

Flóra Földesi, Hanna Karátson, Luca Martins,
Sara Cetin, Zita Vattamány

GitHub Link

April 2025

1 Introduction

Skin cancer is one of the most common types of cancer, producing more than 300,000 new cases per year. What is more, the rates are increasing every year: the expected number of diagnosed cases in 2025 is going to increase by 5.9 percent. Melanoma is the most invasive type of skin cancer, but the overwhelmed health care systems and the shortage of radiologists make waiting lists long, whereas for cases like melanoma, early detection is key for a successful treatment.
This situation requires a solution that involves modern technology. Applications using Artificial Intelligence (AI) can be useful tools for people who do not have the opportunity to visit a specialist. Furthermore, combining human resources and AI tools could optimize time and provide more accurate predictions. This is what we are investigating in this report; How can machine learning models help to identify melanoma? With our open question; Does redness and swelling around the lesion affect the prediction of melanoma?

2 Data

2.1 Source

For our models, we used the image data set and the metadata set from the Federal University of Espírito Santo (UFES) that has been made available at Mendeley Data. The data set consists of raw images of the skin lesion area of 1373 patients and is being updated every two years. They were taken with smartphones, using an application specially for medical purposes. In addition to the images, the metadata set consists of 26 attributes and includes both clinical and identification features.

2.2 Cleaning

Before going further with our analysis, we needed to gather a separate dataset that only includes the id-s and the ground-truth labels of the images. In our case, the ground truth labels represent the presence of melanoma. For the aim of this project we used the "img_id" and the "diagnostic" columns of the metadata set. We began by excluding the non-relevant columns, and converting the diagnostic column into binary values, where "1" indicates the presence of melanoma and "0" the absence.

2.3 Preparation

Out of 1642 unique lesion IDs, only 31 were melanoma cases. To address this class imbalance, we explored various methods like SMOTE and selective sampling to achieve a higher proportion of melanoma cases. However, these approaches introduced biases that we wished to avoid. Therefore, for our final dataset, we used all 2,299 available images, as a larger dataset improves the reliability and robustness of the results. It also improves the generalizability of the models, which is crucial since we aim to implement our model on a large and diverse patient population.

2.4 Feature extraction

The 'ABCDE' of melanoma is an acronym for key features to identify potential skin cancer. For our baseline project, we focused on the first three: Asymmetry, Border, and Color.

We measured asymmetry by rotating the binary mask at four angles and averaging the asymmetry scores to provide a rotation-invariant measure. A mean score of 1 would mean a perfectly symmetric lesion, and a mean score of 0 would be the most asymmetrical.
The border was evaluated using compactness. We divided the area by the square of the perimeter, then scaled this ratio by 4π (the compactness formula). Values closer to 1 indicate a more circular, regular shape, while lower values suggest irregularity.

The color irregularity was measured by segmenting the image into superpixels and computing the variation in red, green and blue intensities based on their mean values. A high variance across the channels indicates greater color irregularity within the lesion.

In our extended method, we added hair removal and an additional feature, blue veil. Hair removal was applied before mask generation to improve segmentation accuracy, as hair can distort mask borders and negatively impact feature extraction. We detected the hair using a blackhat filter, applied a threshold of 10, and then inpainted to remove them, resulting in cleaner and more reliable feature extraction further along the way.
To measure blue veil, which appears as a bluish or bluish-white hazy area on the lesion, we use a function that specifically counts pixels where the blue value is high and the red and green values are similar.

Together, these features can be used as an indicator of melanoma. A lesion is more likely to be melanoma if it has low asymmetry and border scores, but high color and blue veil scores.

3 Model training and tuning

3.1 Logistic regression

Logistic regression is a classification method that takes several input variables, in our case the scores of the extracted features, and models the probability of the instance being a melanoma lesion or not. It uses the sigmoid function, $\sigma(x) = \frac{1}{1+e^{-x}}$, where z is a linear combination of the input scores. A threshold of choice is applied to determine whether a lesion has melanoma or not. It being easy to implement and interpret, made it a suitable choice for the baseline model.

3.2 Random Forest

Random Forest is a classification method that creates a set number of decision trees, each trained on a bootstrapped subset created from the available samples. During each split in a tree only a random subset of the features is considered. This randomness is used to ensure that the trees created from the process are and only weakly correlated. Once all the trees are created, the algorithm runs a sample through all of them, each time giving individual predictions. Since we are using this method for a classification task, the answers are compiled and the most frequent categorical value is turned into the final prediction. We chose Random Forest for its ability to handle complex, non-linear relationships in the data, and because we expected it to outperform the baseline method on this classification task.

4 Results

4.1 Logistic Regression

Following a series of experimentation, the model's performance on the validation set stabilized and we proceeded to evaluate it on the test set. The baseline classifier achieved a 67% accuracy score, similar to the 66% observed on the validation set. The relatively low accuracy stems from several reasons. Firstly, the baseline method only includes 3 features and no hair removal was performed on the images, which may impact the model performance. Additionally, logistic regression assumes a linear relationship between the independent variables and the log-odds of the dependent variable. Statistical tests showed that this assumption is violated for several variables, suggesting that the model is not suitable for our dataset.

The confusion matrix, seen in Figure 1, is generated using a default threshold of 0.5 and displays the model's poor ability to distinguish melanoma from non-melanoma cases. The model falsely predicted a considerable amount of non-melanoma cases as melanoma, which may be expected at this baseline stage

due to the model's limited complexity and tuning.
During the validation process, we aimed to improve the model by considering a different threshold than the default value of 0.5. The goal was to find a threshold that maximizes true positives while minimizing false negatives. Given the critical importance of detecting melanoma, accepting a lower accuracy and an increased number of false positives was necessary. On the validation set a threshold of 0.53 yield the best trade-off. However, to avoid overfitting and to promote better generalization on unseen data, we chose a standard threshold of 0.5 for the final test set.

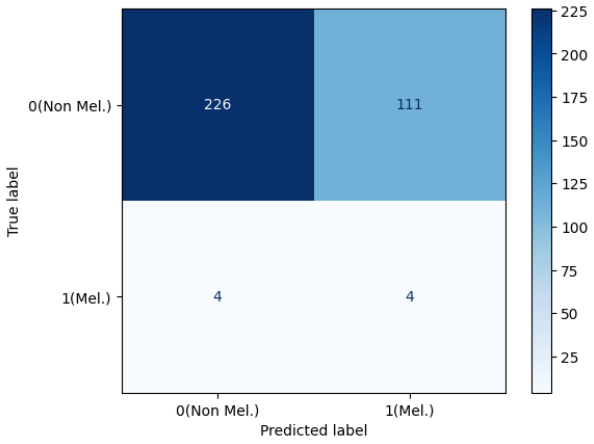

Figure 1: Confusion Matrix for Logistic Regression on Test Data

Due to the imbalance the model struggles to perform well overall. From the classification report, we can read a notably low precision of 3%. This shows that when the model predicts a melanoma case, it is almost always incorrect. We observed that a declined in the model's performance on unseen data, where it missed half of the true melanoma cases, compared to a fourth in the validation set. This gap suggests that the model's ability to generalize may be more restricted than we had initially expected. The low f1-score, supports the fact that the model does not confidently identify instances of the minority class.

	precision	recall	f1-score	support
0	0.98	0.67	0.80	337
1	0.03	0.50	0.07	8
accuracy			0.67	345
macro avg	0.51	0.59	0.43	345
weighted avg	0.96	0.67	0.78	345

Figure 2: Classification Report for Logistic Regression on Test Data

In figure 3, we present a ROC curve for the logistic regression, along side a ROC-AUC score of 0.65. The ROC curve plots the true positive rate against the false positive rate at various thresholds. The AUC score represents the area under the ROC curve, and ranges from 0 to 1. A score of 1 indicates perfect performance of the model, while 0.5 corresponds to random guessing. With a score slightly above 0.5, the model demonstrates minimal improvement over guessing and suggests that there is room for improvement.
Furthermore, the steep shape of the ROC curve indicates that achieving a true positive rate also results in a high false positive rate. However, maintaining a low false positive rate comes with the cost of a low true positive rate. This trade-off makes it challenging to select an optimal threshold that balances

recall, precision and specificity.

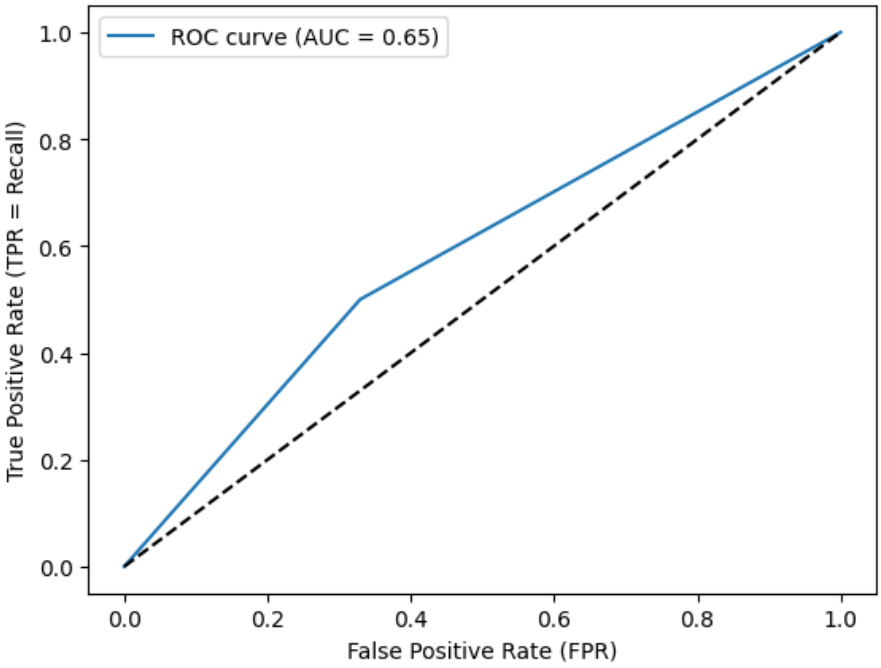

Figure 3: Logistic Regression ROC curve on test data

4.2 Random Forest

During the validation process for the extended method, we aimed to determine the most effective threshold, similarly to the baseline method. Given the clinical relevance of our task, having a high true positive rate is more crucial than minimizing the false positive rate. Thus, we opted for correctly identifying more melanoma cases. As shown in the ROC curve plot with thresholds (Figure 4), it is clear that the threshold needed to be between 0.0 and 0.1 to meet this criteria. However, with a threshold of 0.05, only half of the melanoma cases were detected. We therefore adjusted the threshold and agreed on 0.01. Supported by a promising AUC score on the validation data, we proceeded to evaluate on the test data.

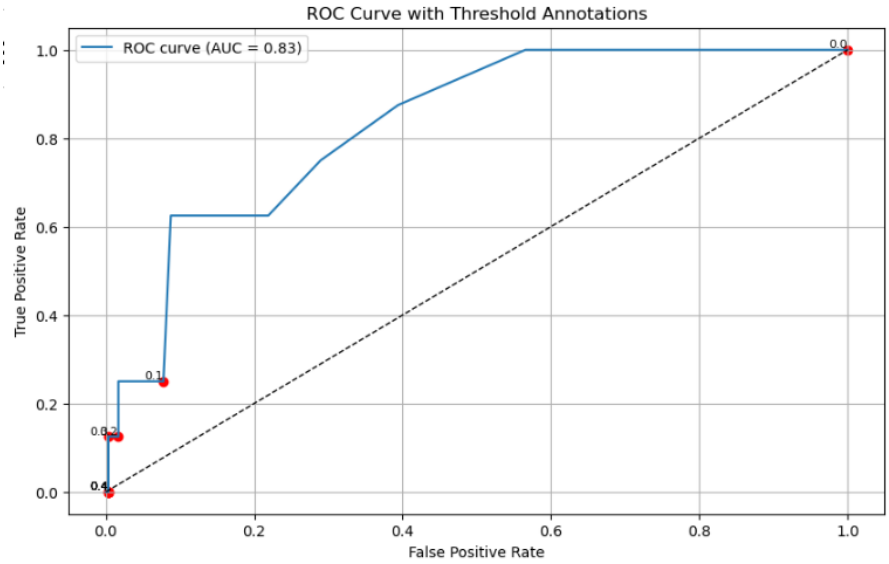

Figure 4: ROC curve with thresholds, random forest, validation data

With a threshold of 0.01 the random forest classifier had an accuracy of 0.50, which is inevitably not desirable but the confusion matrix (see Figure 5) on the test set can be used to further investigate the model's performance.
The tradeoff we chose to express is that a high true positive rate also drastically increases a high false positive rate. As shown in the confusion matrix, a large proportion melanoma cases were correctly classified, and there were a few more false positives than true negatives. From a clinical perspective, the model would have to ensure that critical cases, such as melanoma, are not overlooked. The model's performance should therefore be evaluated based on more than just the accuracy score.

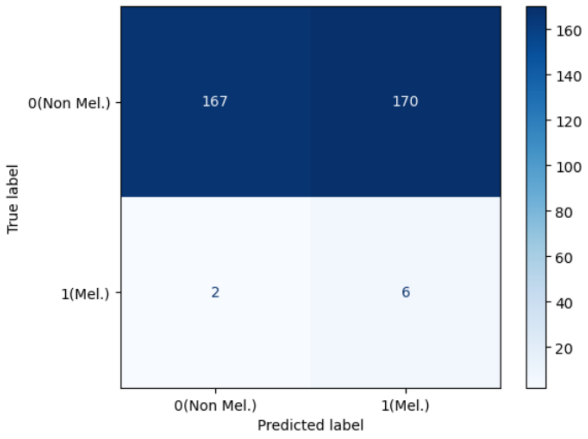

Figure 5: Confusion Matrix, random forest, test data

The AUC score, as shown in Figure 6, is lower for the random forest than for the logistic regression, however that does not imply uniformly better performance by the latter. Essentially both of the models struggled to distinguish between the majority and minority classes, resulting in no threshold that could

guarantee a high true positive and a low false positive rate.

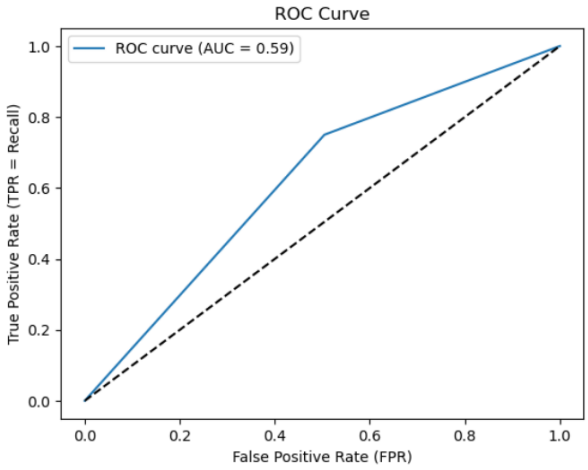

Figure 6: ROC curve, random forest, test data

In Figure 7 we have the classification report for random forest on the test data. While the precision for melanoma (class 1) is the same as in logistic regression, the recall is much higher. Thus making the model better at detecting true melanoma cases. This suggests that the high-stakes nature of the task is better addressed by the random forest.
Overall, both of the models performed worse than we had originally presumed and can benefit from certain improvements. That being said, the random forest gains the upper hand in this specific clinical context, where the patients with melanoma require urgent care.

	precision	recall	f1-score	support
0	0.99	0.50	0.66	337
1	0.03	0.75	0.07	8
accuracy			0.50	345
macro avg	0.51	0.62	0.36	345
weighted avg	0.97	0.50	0.65	345

Figure 7: Classification report, random forest, test data

5 Limitations

In our research we had some obstacles because of the limited amount of information, which is why we could only analyze Asymmetry, Border and Color from the recognized ABCDE features. Not knowing the distance the images were taken from we cannot measure diameter, and since the time intervals between images from the same person are unknown, we are unable to analyze the lesions' evolution. Thus the features Diameter and Evolving cannot be extracted.
Another constraint is that the masking does not work equally well on all pictures. The quality differs overall, but particularly completely black or white lesions go overlooked. Additionally, the mask function is not always able to handle two lesions in the same image. The function selects one seed point, the darkest pixel in the image, and grows a region around it based on that point. As a result, only one lesion is segmented, and any additional lesions are missed because they are not connected to the initial region and no other seed points are considered (seen on Figure 8).

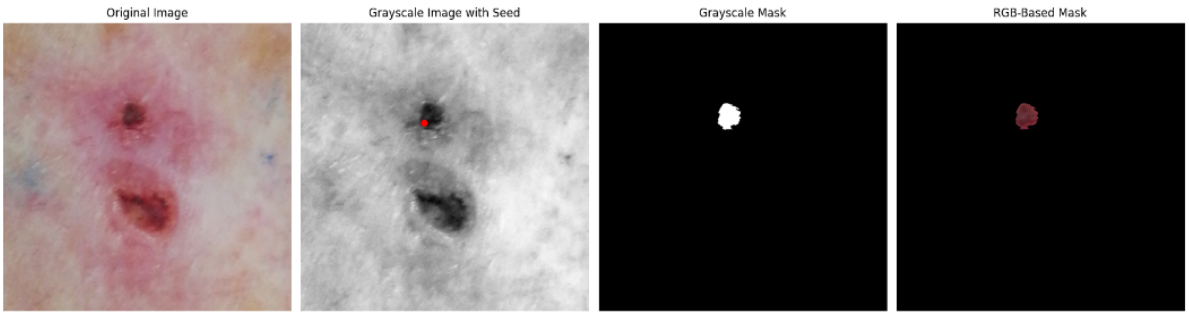

Figure 8: Two lesions where only one is detected

6 Future work

With more information about the patients we could make even more precise prediction about their lesions. One factor we could research is their geographical location, since the distance from the equator influences the skin fairness and sun exposure. We could investigate how the effect of ultraviolet radiation varies between people with different skin types, considering the number of skin layers damaged. Another interesting thing to look at is the socioeconomic status of patients, which influences not only lifestyle and health, but also awareness of diseases such as skin cancer.

What we chose to further investigate was whether redness and swelling around the lesion would improve the prediction of melanoma. As seen in Figure 9 and 10, the region growing function does not always take into account the red area that appears around some lesions in the dataset. This red area may still hold valuable information for the model, as redness and swelling can be signs of inflammation. Inflammation is the body's response to various triggers, such as infections, allergic reactions, autoimmune disorders, or physical injuries. Although redness and swelling are not the primary diagnostic features of melanoma, in some cases they can indicate its presence.
In an open-access case report published by Cureus, redness and swelling were the key factors in detecting melanoma. According to the paper, "the patient presented to the emergency department with a 24-hour history of pain, redness, and swelling of the right thigh and inguinal region". Later, they diagnosed the patient with metastatic melanoma, which means that skin cancer spreads inside the body and the lesion might even disappear. In this case, the absence of a visible lesion led to a poor prognosis, and the condition rapidly progressed to a fatal outcome. This case highlights the importance of considering redness and swelling as potential indicators in melanoma detection.
By all means, it should be taken into account that the symptom redness and melanoma are not always connected. Redness can indicate many other things, mostly related to sun exposure, chemical exposure, or injuries. However, if other symptoms appear as well, it is likely that the redness is caused by melanoma. As a potential solution, a separate mask could be applied exclusively to the surrounding area of the lesion to determine whether the redness is similar to that seen in other melanoma cases. In conclusion, during the classification process, checking for redness combined with the other factors can give a more supported output.

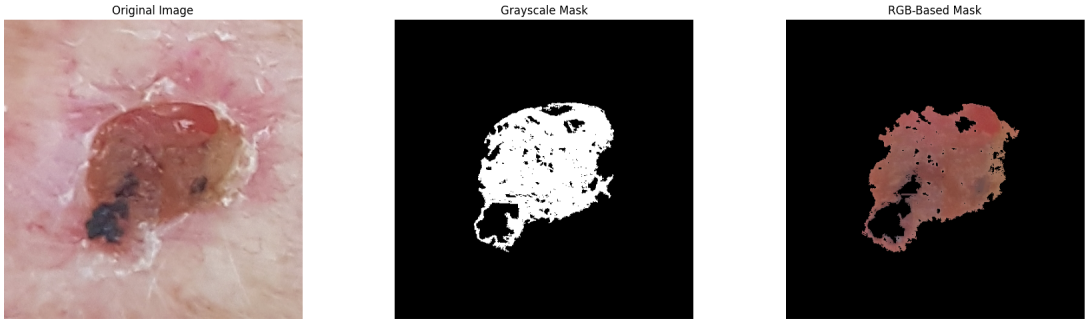

Figure 9: Redness not picked up by segmentation mask

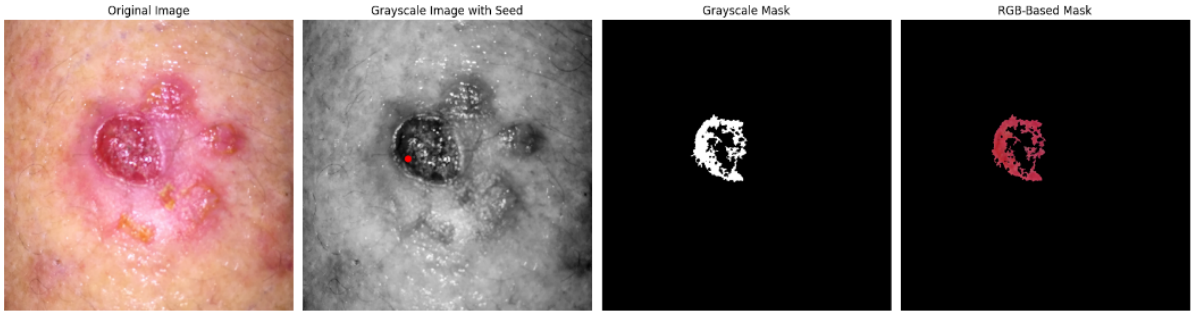

Figure 10: Redness not picked up by segmentation mask

7 References

Cancer Research UK - A study measuring the risks and benefits of exposure to sunlight
Melanoma: Multiple Presentations of an Invisible Illness
The Lancet: Melanoma
Socioeconomic and lifestyle factors and melanoma: a systematic review