First Year Project Report: Image Analysis

IT University of Copenhagen

Anna Weronika Lekston awle@itu.dk **Dara Dankova Georgieva** dage@itu.dk **Julia Otto** juot@itu.dk

Sofia Stanimirova Sotirova ssot@itu.dk **Yossef Alla Eldin Salem** yoal@itu.dk

Abstract

This report presents a project carried out as part of the first-year curriculum at the IT University of Copenhagen, focused on medical image analysis. The study aims to answer the following question: Can machine learning models, trained on medically-inspired features extracted from clinical images, accurately identify melanoma? An open question explored in this study is whether synthetic minority oversampling can effectively enhance model performance in reducing misdiagnosis through the mitigation of class imbalance.

The findings of the study show that Random Forest combined with the SMOTEENN method achieves the best results in terms of recall a critical metric for reducing underdiagnosis of melanoma. The outcomes underscore the potential of image analysis techniques to support more accurate melanoma detection, while also highlighting areas for further improvement in addressing challenges within dermatological healthcare.

1 Introduction

Melanoma is a malignancy of melanocytes, pigment-producing cells located in the basal layer of the epidermis. Although it accounts for only about 1% of all skin cancers, it causes more than 80% of skin cancer-related deaths ((Saginala et al., 2021)). In the early stages, melanoma is highly treatable with surgery, but outcomes decline dramatically once the disease metastasizes ((Davis et al., 2019)).

A major challenge in the detection of melanoma is its frequent visual resemblance to benign skin lesions, which often leads to misdiagnoses. Underdiagnosis can delay lifesaving treatment, while overdiagnosis - the identification of tumors that would not progress - can lead to unnecessary biopsies, patient anxiety, and draining healthcare resources ((Muzumdar et al., 2021)). These diagnostic challenges are compounded by increasing incidence rates, which may reflect a mix of true rise, heightened screening, and expanded histopathologic criteria. At a global scale, the public health burden of melanoma continues to grow. In 2020, there were an estimated 325,000 new melanoma cases and 57,000 deaths worldwide. If current trends persist, incidence may increase to 510,000 cases and deaths to 96,000 annually by 2040—a 50% and 68% increase, respectively ((Arnold et al., 2022)). These projections highlight an urgent need for improved diagnostic strategies.

The methodology used in this study involved preprocessing steps, including hair removal and data cleaning, followed by the extraction of clinically relevant features, using image processing pipelines built with Python and OpenCV. Using these features, we chose to train and evaluate Logistic Regression and Random Forest classifiers. Performance was assessed using standard metrics, including accuracy, precision, recall, F1 score, AUC, and confusion matrices. To address data imbalance present in the dataset - particularly the underrepresentation of melanoma cases — we applied the bootstrapping-based resampling strategy SMOTEENN to enhance the model's medical accuracy and sensitivity.

2 Data

2.1 Source

The data used in this project is sourced from ((Science Direct at el., 2020)) - a PAD-UFES-20 dataset, which contains clinical images of skin lesions collected via smartphones during patient ap-

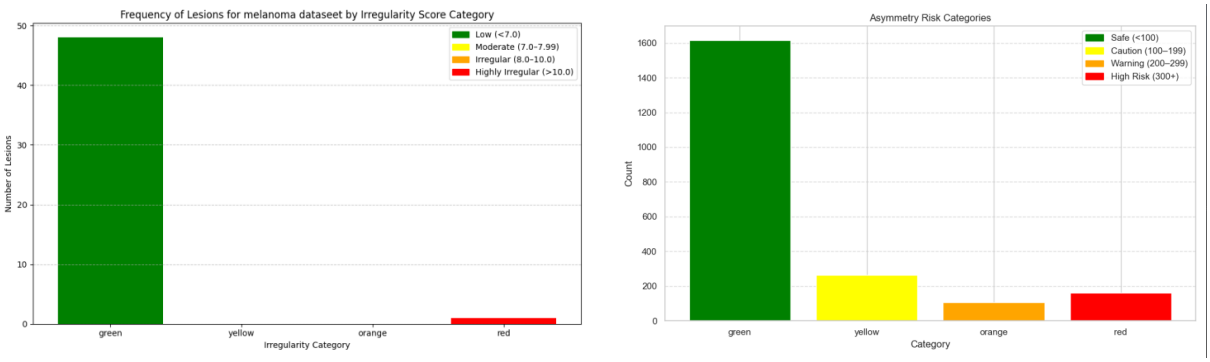

Figure 1: Plot of distribution of images based on Asymmetry and Border Irregularity features.

pointments by the Dermatological and Surgical Assistance Program at the Federal University of Esprito Santo. Each sample in the dataset includes at least one image of a skin lesion along with a corresponding binary mask that highlights the pixels belonging to the lesion. Lesions were labeled by expert doctors based on visual inspection, and in some cases, their diagnoses were confirmed by biopsy and laboratory tests. The dataset is designed to support the development of Computer-Aided Diagnosis (CAD) systems for skin cancer detection without relying on dermoscopy images, which makes it particularly valuable for applications in low-resource or remote settings.

In total, the dataset comprises:

- 1373 patients
- 1641 unique skin lesions
- 2298 images, covering six different diagnostic categories (three skin diseases and three types of skin cancer)

2.2 Data Cleaning

The data cleaning process involves manually reviewing each lesion mask to remove those that appear unusual or do not match typical skin disease patterns. All discarded lesion masks are compared with their original images to confirm that no valid lesion was accidentally removed. We then created a table with the names of all melanoma pictures and compared them with the deleted images to make sure that no melanoma cases had been removed. During the process, we identified three images of melanoma that did not have the corresponding lesion masks. Therefore, we manually create the missing masks using the online platform 'Photopea'.

3 Feature Extraction

We have decided to follow the ABCD rule (Asymmetry, Border Irregularity, Color, Diameter) when creating our features, as it is the main method dermatologists use to detect skin cancer and has been tested clinically for decades. We also include a fifth feature - Blue-White Veil — which is another dermoscopic structure often associated with melanoma. We decided not to extract Diameter as a feature because the images in our dataset come from heterogeneous sources, and therefore differ significantly in scale and resolution.

Asymmetry was measured by a function that calculates asymmetry by rotating the lesion mask, comparing XOR differences between the original and rotated masks, and averaging the results across multiple angles. A higher score indicates greater asymmetry, which is often associated with malignancy.

Border Irregularity was quantified using the circularity formula applied to the contour of the lesion.

$$compactness = \frac{parameter^2}{4\pi area} \quad (1)$$

This metric captures how much the lesion deviates from a regular circular shape.

As shown in Figure 1, the distributions of both asymmetry and border irregularity features reflect the class imbalance present in the dataset. The predominance of low-risk scores (smooth borders and symmetric shapes) corresponds to

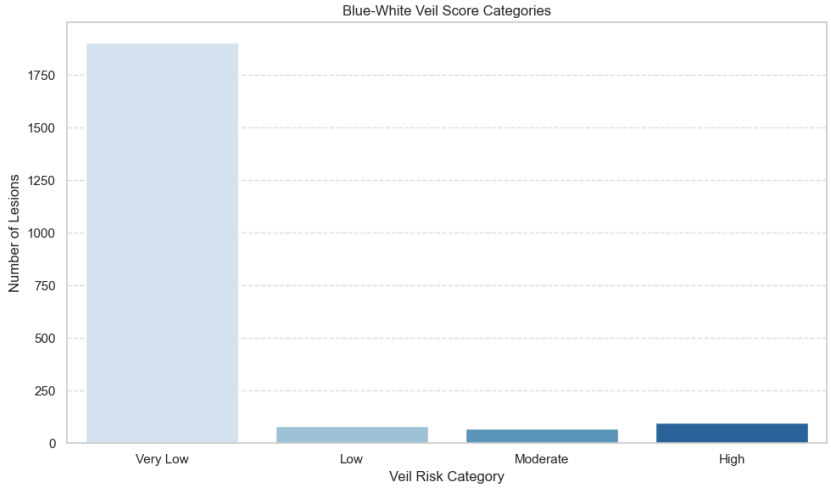

Figure 2: Plot of distribution of images after applying Blue-White Veil feature.

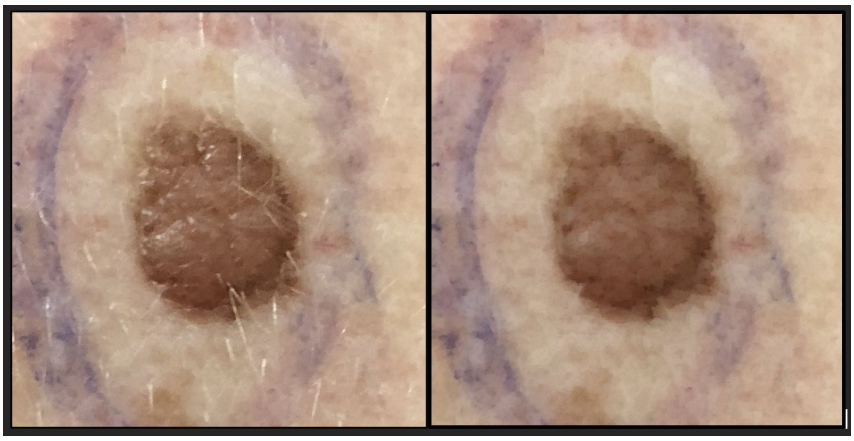

Figure 3: Hair Removal: Before and after hair removal using blackhat and whitehat.

the overwhelming majority of non-melanoma cases, while the sparse high-risk values (irregular borders and asymmetric shapes) align with the rarity of melanoma images. This visualization quantitatively confirms the dataset's skew toward non-melanoma lesions, which underscores the need for techniques like SMOTEENN to address this imbalance during model training.

Color Variation was assessed by extracting all lesion pixels (via masking), and computing the standard deviation in each RGB channel over the masked area:

$$\sigma_c = \text{std}(\text{pixels}_c), \quad c \in \{R, G, B\}$$

Then, we compute the mean of the three channel standard deviations:

$$\text{ColorVar}_{\text{RGB}} = \frac{\sigma_R + \sigma_G + \sigma_B}{3} \qquad (2)$$

A high value reflects visual complexity or irregular pigmentation, which may be indicative of malignant lesions.

Blue-White Veil corresponding pixels were identified by selecting those exhibiting either characteristic blue tones or bright, near-white coloration. The blue pixels were defined based on hue, saturation, and brightness ranges that capture typical blue appearance while excluding pale or overly bright pixels, helping to minimize false of blue pen marks around lesions, present in many of the images in the dataset. The white pixels were selected for low saturation and high brightness to represent the veil's whitish regions commonly seen in dermoscopic images.

The final blue-white veil score ($\text{BWV}_{\%}$) was calculated as the percentage of lesion pixels that satisfy either the blue or white pixel criteria:

$$\text{BWV}_{\%} = \frac{\text{blue/white pixels in lesion}}{\text{total lesion pixels}} \times 100 \quad (3)$$

Figure 2 shows the distribution of images across predefined Blue-White Veil categories. As expected, the vast majority of lesions fall into the "Very Low" category, which is consistent with the the class distribution in the data set as well as with the fact that blue-white veil is a relatively rare dermoscopic feature.

Hair Coverage Ratio feature quantifies the proportion of the lesion area occluded by hair. This is achieved by first applying morphological operations - blackhat and whitehat transformations - to detect hair regions on the lesion images. The detected hair pixels are combined into a binary mask, which is then used both to calculate the hair coverage ratio (the fraction of lesion pixels covered by hair) and to perform hair inpainting, producing hair-removed images for downstream analysis.

Figure 3 demonstrates the efficacy of the hair removal pipeline, showcasing representative results before and after processing. The left panel displays the original lesion image with obstructive hair filaments, while the right panel presents the cleaned output where hair structures are selectively removed. Critically, the lesion's diagnostic features—such as pigment networks and border structures—remain intact, confirming the method's suitability for downstream tasks like feature scoring.

4 Classification and Evaluation

For the classification task, we employed binary classification, where class 0 represents images identified as non-melanoma and class 1 corresponds to melanoma. Prior to selecting classifiers, we conducted a visual inspection of the class distributions and analyzed model performance using accuracy, precision, recall, F1 score, AUC score, and confusion matrices as evaluation metrics. They evaluated the trade-off between sensitivity (true positive rate) and specificity (true negative rate), ensuring a balanced handling of false positives and false negatives. These metrics varied depending on the model and preprocessing techniques applied.

We selected Logistic Regression and Random Forest as our final models due to their complementary strengths. These are considered supervised learning algorithms, meaning they require true labels during the training process to learn from the data. Other models were excluded due to instability, complexity, or poor fit with our data characteristics. In clinical decision support, it's important to balance:

- Model interpretability (for physician trust).
- High recall (to avoid missing malignant cases).
- Robustness to imbalance (melanoma being rare).

After observing the data, we noticed a significant class imbalance: the proportion of images diagnosed as melanoma (52 images) was much lower than non-melanoma cases (2 211 images). This imbalance posed challenges for our classifiers, which tended to favor the majority class, reducing sensitivity to actual melanoma cases. To address this, we applied a bootstrapping-based resampling strategy using SMOTEENN (Synthetic Minority Over-sampling Technique combined with Edited Nearest Neighbors). This technique first generates synthetic samples for the minority class x_i and then removes ambiguous or noisy examples using nearest-neighbor rules. The goal was to improve class balance and data quality.

When creating a new synthetic data point x_{new} this technique uses the following formula:

$$x_{\text{new}} = x_i + \delta \cdot (x_{zi} - x_i) \quad (4)$$

where δ is a random number in the range $[0, 1]$.

This augmentation was incorporated into our modeling workflow using a flag (`use_bootstrap=True`), allowing us to compare model performance with and without SMOTEENN augmentation. We hypothesized that this method would enhance recall by improving the model's ability to detect minority class instances without excessively compromising precision.

4.1 Logistic Regression

Logistic regression is a parametric classifier that takes the output of the linear regression function as input and uses a sigmoid function to estimate the probability for the given class. In our case, we applied L1 regularization (Lasso) to promote sparsity in the model's coefficients, effectively carrying out feature selection and helping prevent

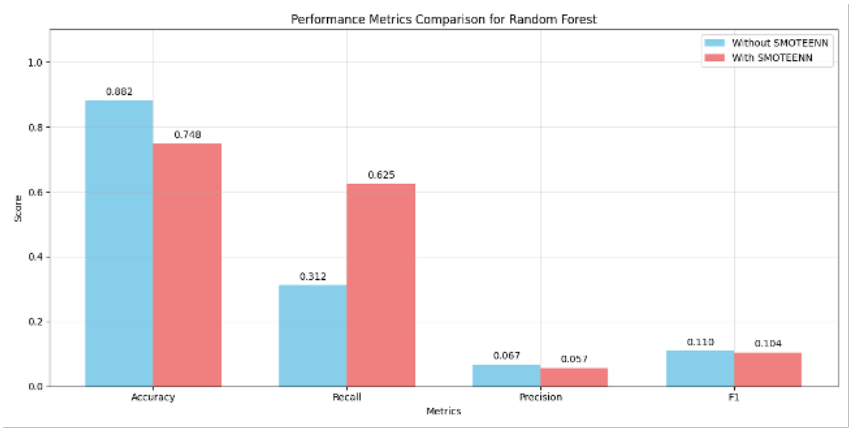

Figure 4: Plot of results before and after bootstrapping for Random Forest Classification

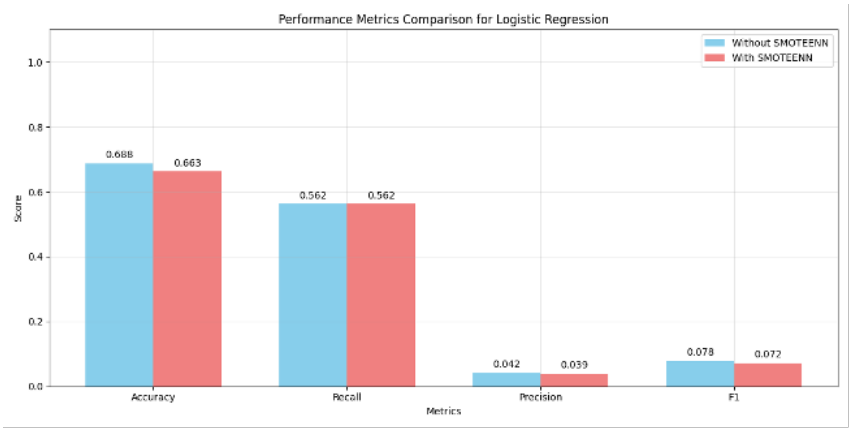

Figure 5: Plot of results before and after bootstrapping for Logistic Regression Classification

overfitting. L1 regularization is used by adding a penalty to the model based on the absolute values of the coefficients. This is especially beneficial when handling high-dimensional data, as it can streamline the model by disregarding less important features. It also has weights for each predictor feature, which show how the predictors and outcome variable relate to one another. The classifier was called 100 times in a row. Logistic Regression offers high interpretability, which is valuable in clinical settings.

4.2 Random Forest

Random forest classification is a non-parametric ensemble learning method used for categorical outcome prediction. It operates by constructing multiple decision trees during training. It then outputs the class that is the mode of individual tree predictions. Each tree is built using a random subset of the training data and a random subset of features, which helps reduce overfitting and improve model stability. Unlike logistic regression, it does not use coefficients. It evaluates feature importance based on how much each of them improves node purity. In our study the model was called 100 times, just like Logistic Regression. We selected Random Forest due to its effectiveness in handling imbalanced data, especially when used in conjunction with resampling methods like SMOTEENN.

4.3 Results

For evaluation, we used 5-fold Stratified K-Fold Cross-Validation to preserve class proportions in each fold. We trained both Logistic Regression and Random Forest classifiers on regular and augmented data in each of the five folds. This approach provided a robust estimate of generalization performance across varying data splits. From the performance, we've provided plots that give insights into how SMOTEENN impacts the used classifiers.

In fig. 4 for Random Forest, the results are shown:

- The accuracy score dropped by approxi-

mately 0.134.

- The recall improved by 0.313.
- Precision decreased slightly by 0.01.
- F1-score remained relatively unchanged, with a 0.006 decrease.

In fig. 5 for Logistic Regression, the following results were observed:

- The accuracy score dropped slightly by approximately 0.025.
- The recall remained unchanged at 0.562.
- Precision decreased slightly by 0.003.
- The F1-score showed a slight decrease of 0.006.

5 Performance Analysis

Prior to augmentation, both classifiers exhibited high specificity but low sensitivity, reflecting a tendency to misclassify melanoma cases as non-melanoma. Logistic Regression favored precision, while Random Forest offered slightly better recall, though still insufficient for clinical use.

The application of SMOTEENN had different impacts of the two classifiers. In Random Forest, resampling led to an improve of recall, meaning the model became better at identifying melanoma cases. That is a postitive change, where missing a positive case can have serious consequences. This gain in sensitivity came at the expense of overall accuracy and a slight rise in false positives, as reflected by the drop in precision. However, in the medical context, especially for early-stage cancer detection, this is acceptable, as it prioritizes patient safety by ensuring fewer true cases are missed. The F1-score remained largely stable, indicating that the model's balance between precision and recall did not change dramatically, even though its ability to detect positives improved.

In contrast, for Logistic Regression there was no improvement in recall, and slight decreases were observed in accuracy, precision, and F1-score. This suggests that the linear model struggled to take advantage of the synthetic and cleaned data provided by the resampling method.

Therefore, Random Forest outperformed Logistic Regression in handling the augmented data and provided more robust, clinically useful predictions—supporting its suitability for tasks where minimizing missed malignant cases is essential.

6 Limitations

One of the primary limitations encountered in this study was the limited size and diversity of the dataset, particularly the underrepresentation of melanoma images. This limitation significantly reduced the model's ability to learn distinguishing features of melanoma and the final performance.

Limited melanoma examples not only restrict model training but also reduce generalizability. As some research show, inadequate datasets, especially those lacking high-quality melanoma images, can severely affect machine learning outcomes. For example, in one study, only 36.6 percent of the provided melanoma images were detected suitable for analysis, while the remainder were excluded due to issues such as the absence of normal surrounding skin or the absence of pigmentation ((Rinner et al., 2020)). This finding aligns with our experience, where the limited and imperfect data quality reduced model performance.

Another specific challenge regards amelanotic melanoma, which accounts of approximately 2–8 percent of all melanomas. These lesions often lack pigmentation and evidence symmetry, making them difficult for current machine learning systems to detect ((Kern et al., 2017)). Future models would benefit from datasets that include a greater number of such cases.

Moreover, the small dataset raises concerns about the model's robustness when applied to new, diverse, or lower-quality images. As noted in some reasearches, limited training data can lead to overfitting, where the model performs well on training samples but fails to generalize to unseen data ((Digit Health., 2024)).

To reduce the unwanted consequences of limited data availability, we applied bootstrapping as a strategy to improve the performance of our model.

Another limitation of our study involves the inability to accurately determine the physical diameter of skin lesions. The images in our dataset were captured using a variety of smartphone devices, resulting in differences in scale. In theory, to obtain needed diameter scores, we would have to convert pixels into centimeters using the formula:

$$\text{Diameter (mm)} = \frac{\text{Object size (pixels)} \times \text{Sensor size (mm)}}{\text{Image resolution (pixels)}} \tag{5}$$

However, this calculation requires detailed data in-

formation that are not provided in our case which makes obtaining the result not feasible. Although the dataset includes a metadata.csv file with diameter values, many entries are missing, making the feature unreliable for consistent model input.

Despite the described limitation, lesion diameter is not considered a critical feature in melanoma classification, especially when compared to others such as asymmetry. Studies using the Total Dermatoscopy Score (TDS) often assign a lower weight to diameter (0.5) compared to asymmetry (1.3), suggesting a relatively minor influence on diagnostic outcomes. Additionally, while a diameter threshold of more than 6 mm has traditionally been used as a diagnostic criterion, recent research highlights that many melanomas are now detected at smaller sizes due to increased awareness and early screening ((Science Direct., 2021)).

Given the above, we decided not to include the diamter feature in our model. Our focus was on reducing false positives and false negatives, enhancing the detection of clinically significant features, particularly those prioritized in the ABCD rule framework. While diameter remains a component of this rule, it is generally less informative for detecting early-stage melanomas compared to asymmetry, border, and color features.

7 Conclusion and Future Work

This project explored the use of clinically inspired features—asymmetry, border irregularity, color variation, and blue-white veil—for melanoma classification. We implemented a pipeline combining these features with Logistic Regression and Random Forest classifiers, evaluated under SMOTEENN resampling to mitigate class imbalance, which resulted in findings that confirm that non-melanoma lesions can be detected with reasonable accuracy using our features. However, the models struggle to detect malignancies reliably, resulting in false negatives and therefore low scores for the recall evaluation metrics. Key find of the study is that Random Forest classifier combined with the SMOTEENN method achieves the best results in terms of recall therefore reducing underdiagnosis of melanoma. Overall, the study reveals a key limitation: while our features capture important dermatological patterns, they are insufficient alone for robust melanoma detection in real-world settings.

Looking ahead, improving data quality is essential. We hope that in the future we can utilize a dataset composed primarily of medically acquired dermoscopic images, which would provide a more reliable foundation for model training and evaluation. Furthermore, combining machine learning models with expert input from dermatologists—through a human-in-the-loop approach—can enhance diagnostic accuracy. Another valuable direction is tracking lesion evolution over time, which would enable modeling the full ABCDE framework by incorporating the "E" for evolution criterion. Finally, enriching the dataset with clinical metadata such as patient history, lesion location, and professional annotations could support a more holistic and accurate assessment of malignancy risk.

References

[Science Direct at el.2020] Andre G.C. Pacheco, Gustavo R. Lima, Amanda S. Salomão, Breno Krohling, Igor P. Biral, Gabriel G. de Angelo, Fábio C.R. Alves, José G.M. Esgario, Alana C. Simora, Pedro B.C. Castro, Felipe B. Rodrigues, Patricia H.L. Frasson, Renato A. Krohling, Helder Knidel, Maria C.S. Santos, Rachel B. do Espírito Santo, Telma L.S.G. Macedo, Tania R.P. Canuto, Luíz F.S. de Barros 2020. PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. https://www.sciencedirect.com/science/article/pii/S235234092031115X

[Saginala et al.2021] Kartikeya Saginala, Adam Barsouk, Jonelle S. Aluru, Prashanth Rawla, and Artur Barsouk. 2021. Epidemiology of Melanoma. *Medical Sciences*, 9(4):63. https://www.mdpi.com/2076-3271/9/4/63

[Marghoob et al.2009] Ashfaq A. Marghoob, Linus Changchien, Jennifer DeFazio, William Slue, Allan C. Halpern, and Alfred W. Kopf. 2009. The most common challenges in melanoma diagnosis and how to avoid them. *Australasian Journal of Dermatology*, 50(1):1–13. https://doi.org/10.1111/j.1440-0960.2008.00496_1.x

[Davis et al.2019] L. E. Davis, S. C. Shalin, and A. J. Tackett. 2019. Current state of melanoma diagnosis and treatment. *Cancer Biology & Therapy*, 20(11):1366–1379. https://doi.org/10.1080/15384047.2019.1640032

[Muzumdar et al.2021] Sonal Muzumdar, Gloria Lin, Philip Kerr, and Jane M. Grant-Kels. 2021. Evidence concerning the accusation that melanoma is overdiagnosed. *Journal of the American Academy of Dermatology*, 85(4):841–846. https://doi.org/10.1016/j.jaad.2021.06.010

[Arnold et al.2022] Melina Arnold, Deependra Singh, Mathieu Laversanne, Jerome Vignat, Salvatore Vaccarella, Filip Meheus, Anne E. Cust, Esther de Vries, David C. Whiteman, and Freddie Bray. 2022. Global Burden of Cutaneous Melanoma in 2020 and Projections to 2040. *JAMA Dermatology*, 158(5):495–503. https://doi.org/10.1001/jamadermatol.2022.0160

[Rinner et al.2020] C. González-Cruz, M.A. Jofre, S. Podlipnik, M. Combalia, D. Gareau, M. Gamboa, M.G. Vallone, Z. Faride Barragán-Estudillo, A.L. Tamez-Peña, J. Montoya, M. América Jesús-Silva, C. Carrera, J. Malvehy, S. Puig 2020. Machine Learning in Melanoma Diagnosis. Limitations About to be Overcome. *Actas Dermo-Sifiliográficas*, 313-316. https://www.sciencedirect.com/science/article/pii/S1578219020300846

[Kern et al.2017] M.A. Pizzichetta , H. Kittler , I. Stanganelli , G. Ghigliotti , M.T. Corradin , P. Rubegni , S. Cavicchini , V. De Giorgi , R. Bono , M. Alaibac , S. Astorino , F. Ayala , P. Quaglino , G. Pellacani , G. Argenziano , D. Guardoli , F. Specchio , D. Serraino , R. Talamini 2017. Dermoscopic diagnosis of amelanotic/hypomelanotic melanoma. *British Journal of Dermatology*, 538-540. https://academic.oup.com/bjd/article-abstract/177/2/538/6668684

[Digit Health.2024] Maram F Almufareh, Noshina Tariq, Mamoona Humayun, Farrukh Aslam Khan 2024. Melanoma identification and classification model based on fine-tuned convolutional neural network. [https://pmc.ncbi.nlm.nih.gov/articles/PMC11119457/

[Science Direct.2021] Ebrahim Mohammed Senan a, Mukti E Jadhav. 2021. Analysis of dermoscopy images by using ABCD rule for early detection of skin cancer. *Global Transitions Proceedings*, 1-7. https://www.sciencedirect.com/science/article/pii/S2666285X21000017