Cancer Prediction using Logistic Regression

Andrei Sheremet
IT-Universitetet i København
andsh@itu.dk

Nam Tran
IT-Universitetet i København
hotr@itu.dk

Rahat Alam
IT-Universitetet i København
alam@itu.dk

Abstract

Cancer prediction is a critical area of research in medical diagnostics. In this project, we aim to build a logistic regression model to predict cancer using features extracted from dermoscopic images. We explore various features, including those from image preprocessing and a "hair" feature indicating preprocessing type. You can find the full project on GitHub: github.com/NamTrank5/2025-FYP-Final-C2

1 Introduction

Skin cancer remains one of the most common and treatable cancers when detected early. However, timely diagnosis requires access to specialized equipment and expert dermatological assessment—resources that are not always available, especially in under-resourced or rural regions. Recent advances in computer vision and machine learning have opened new possibilities for developing decision-support tools that aid clinicians by analyzing skin lesion images.

In this project, we explore a lightweight and interpretable approach to automated skin cancer classification using logistic regression. The aim is not to replace dermatologists but to support the early triage process, particularly in settings where full diagnostic tools (like dermatoscopes or biopsy services) may not be immediately available.

Problem Statement

We frame this as a binary classification task: predicting the presence or absence of cancer from image-derived features. Given that a misclassified lesion can delay treatment, our goal is to minimize such errors while maintaining overall accuracy. Logistic regression is chosen due to its interpretability and strong baseline performance in binary tasks. We will dive deeper on why we choose this specific Machine Learning model later.

Objectives

The project has three primary goals:

- Build a logistic regression model to classify lesions as malignant or benign based on extracted features.

- Evaluate model performance using metrics such as accuracy and F1-score.

- Compare a baseline model (with three features) to an extended version using a fourth "hair" feature.

Methodology Overview

After extracting and scaling features from each lesion, we split the dataset (80/20 stratified sampling) into training and test sets to preserve class distribution. We train a logistic regression model using scikit-learn with default settings and L2 regularization. For the extended model, we include the hair feature to test its effect on performance.

We evaluate both models using accuracy and F1-score, and compare confusion matrices to understand error types (false positives/negatives). Feature importance is analyzed via model coefficients.

2 Dataset Description

We base our study on the PAD-UFES-20 datasetthis article, introduced in Pacheco et al. (2020), which is composed of 2,298 clinical skin lesion images acquired using smartphone cameras. Each image is associated with a set of up to 21 patient-level clinical features (age, gender, lesion location, skin type, etc.), and each lesion is labeled as either a skin disease or skin cancer. Notably, 58.4 % of samples are biopsy-proven, and 100% of all malignant cases are confirmed by histopathological examination, making it a reliable source for diagnostic modeling.

For our project, we rely on a subset of this dataset consisting of pre-processed lesion images and segmentation masks, alongside a metadata.csv file containing the ground truth diagnosis. This file also includes demographic and clinical data used during patient appointments, as described in the original paper. We define a binary classification task: all lesions labeled as BCC, SCC, or MEL are treated as malignant; others (NEV, ACK, SEK) are treated as benign.

While deep learning is the dominant method in medical image analysis, we instead focus on manually extracted features (e.g., shape, color, hair presence) from the lesion masks and use them to train a basic yet explainable model. This approach helps evaluate how much diagnostic signal can be derived from simple image-based heuristics without requiring large training datasets or GPU resources.

3 Model Building

3.1 Logistic Regression

Logistic regression is a supervised machine learning algorithm commonly used for binary classification tasks. It models the probability that a given input belongs to a particular class by applying the logistic (sigmoid) function to a linear combination of input features:

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

where $\mathbf{x} = (x_1, \ldots, x_n)$ is the input feature vector, and β_i are the learned coefficients.

Unlike linear regression, which predicts continuous values, logistic regression produces probabilities in the range $[0, 1]$, making it particularly suitable for tasks such as cancer prediction, where the output is binary (malignant vs. benign).

3.2 Model Choice Justification

We chose logistic regression for several reasons:

- **Interpretability:** The coefficients in logistic regression offer information on how each characteristic contributes to the prediction, which is important in medical contexts.

- **Simplicity:** Logistic regression is computationally efficient, easy to implement, and well suited for small to medium-sized data sets.

- **Baseline Comparisons:** It provides a strong baseline model, which makes it easier to evaluate whether more complex models offer substantial performance gains.

While more complex models such as decision trees or neural networks can capture non-linear relationships, our primary goal was to establish a clean, interpretable baseline model. Logistic regression strikes a good balance between performance and explainability.

3.3 Feature Description

To enable effective classification of skin lesions, we extracted a set of handcrafted features from lesion masks and images. These features were designed to capture shape, texture, and color properties relevant to cancer detection. The main categories are:

- **Asymmetry:** Calculated by rotating and comparing flipped segments of the mask. High asymmetry suggests malignancy.

- **Compactness:** Measures how close the shape is to a circle using the formula

$$\text{Compactness} = 1 - \frac{4\pi \cdot \text{Area}}{\text{Perimeter}^2}$$

Less compact lesions may indicate irregular, cancerous borders.

- **Color Variation (Multicolor Rate):** Computed by clustering lesion pixels (using KMeans) and measuring the maximum distance between color cluster centers in RGB space. High variation often correlates with malignancy.

- **Hair Feature:** Using morphological filtering and inpainting (OpenCV's blackhat + inpaint), we computed the proportion of detected hair pixels in each image. This was included as a binary feature to indicate whether hair was present and processed.

All features were normalized to the $[0, 1]$ range to ensure compatibility with logistic regression. We trained two model versions — one with the hair-related feature included, and one without — to assess its impact on performance.

The dataset was split into 80% training and 20% testing using stratified sampling to preserve class balance.

4 Results and Evaluation

4.1 Evaluation Metrics

We evaluated model performance using accuracy, F1 score, and the confusion matrix. Accuracy provides a general sense of correctness, but given the potential class imbalance in our dataset, the F1 score is more informative. Balances both precision and recall, making it a better indicator of how well the model handles both classes.
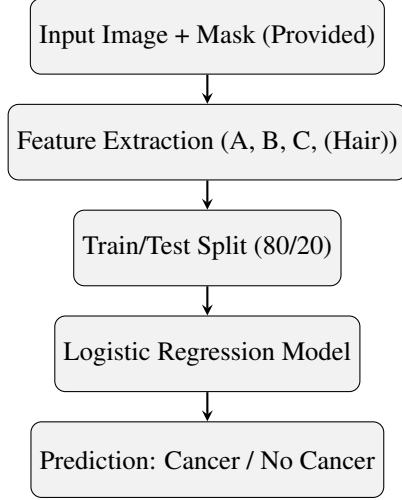
Figure 1: Overview of the model pipeline.

4.2 Baseline Model

The baseline model was trained using three features extracted from each image: Asymmetry (A), Border Compactness (B), and Color Variation (C). We used logistic regression to classify skin lesions as benign or malignant.

On the test set, the model achieved **an accuracy of 0.587** and **an F1 score of 0.600**. While the accuracy was modest, the F1 score indicates reasonable performance in handling both false positives and false negatives.
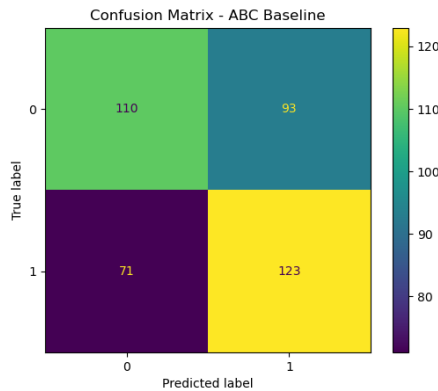

Figure 2: Confusion Matrix – ABC Baseline Model

The confusion matrix (Figure 2) shows that the model correctly identified 123 malignant cases and 110 benign cases, but misclassified 71 malignant and 93 benign samples.

4.3 Extended Model

The extended model added a fourth feature (H), representing the proportion of hair pixels in each image. The same logistic regression algorithm was used for consistency. This model achieved a slightly higher accuracy of 0.602 and an F1 score of 0.609.
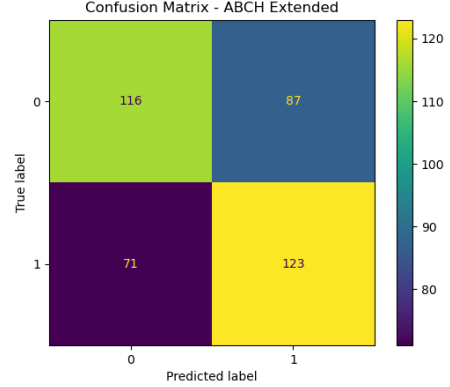

Figure 3: Confusion Matrix – ABCH Extended Model

Compared to the baseline, the extended model slightly improved correct benign classifications (116 vs. 110) while keeping malignant prediction performance the same (123 true positives). False negatives remained unchanged at 71, but false positives were reduced from 93 to 87. These small improvements suggest that **the hair feature (H) contributed useful but limited additional signal.**

4.4 Insights

While both models used a simple logistic regression classifier, we observed that even small changes in feature engineering impacted the results. Adding the hair-based feature (H) slightly improved the accuracy and F1 score, suggesting it provides some additional discriminative value.

Interestingly, we also experimented with normalizing all feature values to the $[0, 1]$ range. Contrary to expectations, the performance worsened: accuracy dropped to 0.588 and F1 score to 0.579. This suggests that the logistic regression model may have benefited from the original raw scale of some features, particularly the asymmetry (A) score.

These observations highlight that even basic modeling decisions—such as whether or not to normalize—can significantly affect outcomes and should be guided by empirical testing rather than assumptions.

4.5 Open Question: Feature Comparison

While our extended model showed slight improvements, a deeper analysis of individual feature contributions re-

vealed a more nuanced picture. As shown in Figure 4, features B (border compactness) and C (color variation) provided the highest discriminative power, while the asymmetry feature (A), despite being intuitively useful, contributed less than expected.
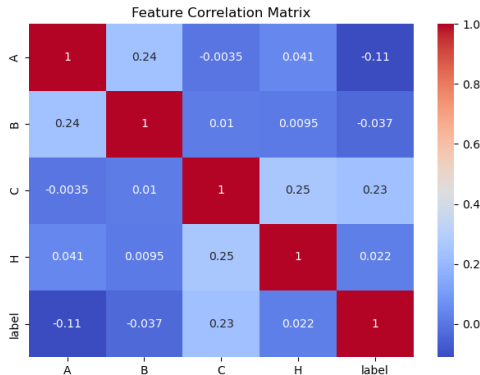

Figure 4: Feature importance comparison based on logistic regression weights.

This raises an open question: to what extent are our handcrafted features complementary, and could feature correlation or redundancy be limiting model performance? Future work could explore feature selection techniques or employ model interpretation tools such as SHAP or the importance of permutation of permutations to better understand these dynamics.

4.6 Limitations

Several limitations affected the model's overall performance. First, the handcrafted features—while intuitive—may not capture the full complexity of lesion characteristics, especially compared to deep learning-based approaches. The asymmetry score, for example, was sensitive to how the lesion was cropped and rotated, potentially introducing noise. Additionally, while the hair detection feature (H) added slight value, its accuracy is uncertain. As shown in Figure 5, the correlation between our automatic hair score and manual hair labels is weak and noisy, suggesting our detection method may need refinement. Finally, since the dataset already provided masks, we did not assess performance on raw images without preprocessing, which limits real-world applicability.

5 Conclusion

We successfully built a simple skin cancer prediction model using logistic regression and handcrafted image features. Despite the model's simplicity, it achieved reasonable accuracy and F1 scores, with the extended ver-
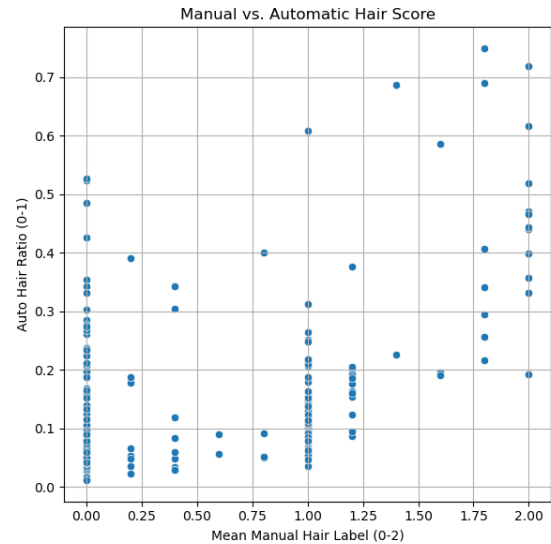

Figure 5: Scatter plot comparing manual hair labels (x-axis) and automatically detected hair ratio (y-axis). Weak correlation suggests limited reliability of the hair detection algorithm.

sion slightly outperforming the baseline. Our experiments showed that even small changes—such as adding a hair detection feature or normalizing input values—can impact model performance.

Interestingly, normalization and asymmetry normalization did not improve results, highlighting that empirical testing should guide design decisions over assumptions. The strongest predictors were border compactness and color variation, while asymmetry, despite its intuitive appeal, contributed less than expected.

While the model has limitations—including potentially noisy features, reliance on provided masks, and weak correlation in hair detection—this project demonstrates that meaningful predictions can still be extracted using basic tools. Future improvements could involve feature selection, more robust hair analysis, or moving toward deep learning techniques for end-to-end training on raw images.