Prediction of Skin Cancer Status (Benign vs. Malignant)

Weiling Luo, Joshua Zhang,
Yuyao Huang, Jingjing Yu

Table of contents

01

Introduction

Skin Cancer Context & Dataset Overview

The Skin Cancer Problem

Skin cancer is one of the most common and preventable cancers in the U.S., with more than five million new cases diagnosed each year (American Cancer Society).

UV exposure remains the single largest risk factor, accounting for the majority of mutations, resulting in malignant skin lesions.

Skin Cancer Dataset

Datasets

Training set: **50,000** patients
Testing set: **20,000** patients (no labels)

Target

Benign or Malignant

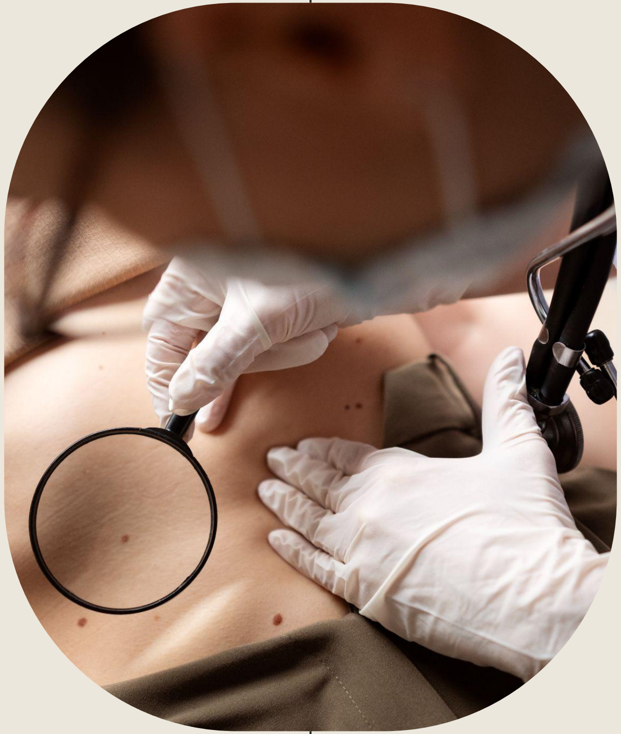
49 predictors (excluding cancer)

Demographic, Environmental, Sun Exposure / Sun Protection, Dermatological features, Lifestyle / Health, Random noise features (explicitly included by instructors)

02

Methodology

Data Cleaning, & Modeling



The Process

01 — Clean Data

Handle missing values, remove irrelevant predictors, and fix multicollinearity

02 — Model Data

Train multiple models on the cleaned dataset to predict malignant vs. benign cases

03 — Analyze Models

Evaluate each model by measuring misclassification rates on testing data

04 — Compare Models

Evaluate each model by measuring misclassification rates on testing data

Checking NA's and Removing Noise Variables

There was approximately 8% missingness on most predictors, so we did median and mode imputation to provide estimates on missing values. And we excluded a number of irrelevant predictors—personal preferences and device-related variables—that did not add anything useful to skin cancer prediction.

Removed:
favorite_color, phone_brand, music_genre, preferred_shoe_type, favorite_cuisine, pets, desk_height_cm, zip_code_last_digit, monthly_screen_time_minutes, uses_smartwatch.

Excluding these predictors allowed us to keep the data clean while enhancing model fit.

Multicollinearity Check

We analyzed groups of predictors that measured similar concepts and observed that sun-exposure, lesion, lifestyle, and environmental variables were strongly correlated. This allowed us to recognize clusters of predictors and remove redundant and low-value predictors.

Sun Exposure Group: avg_daily_uv, sunscreen_freq, sunscreen_spf, skin_photosensitivity, sunburns_last_year, outdoor_job, participates_outdoor_sports, uses_tanning_oil

Lesion Characteristics: lesion_size_mm, lesion_color, lesion_location, number_of_lesions

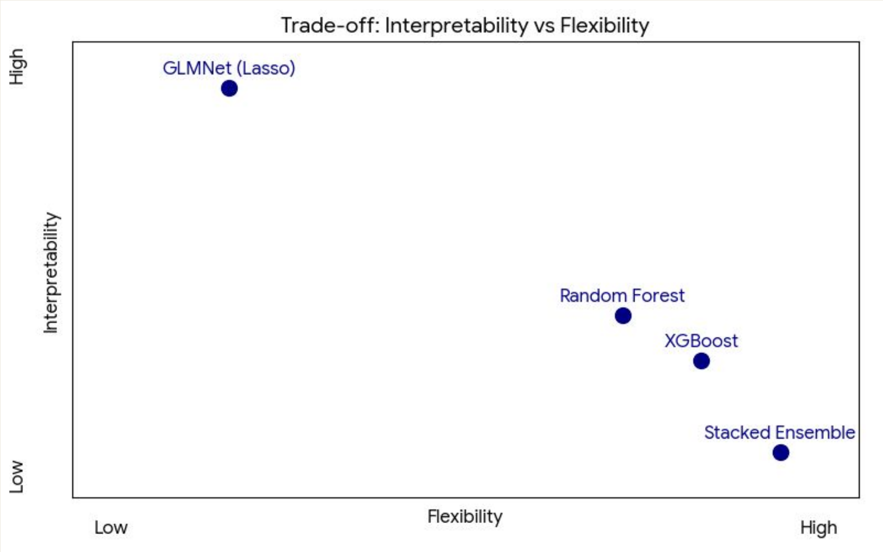
Lifestyle & Health: smoking_status, alcohol_drinks_per_week, BMI, exercise_freq_per_week, vitamin_d_supplement

Environmental: urban_rural, distance_from_beach_km, residence_lat, residence_lon, near_high_power_cables

Summary of Cleaning

- Observation
  - keep 100% of the original observations from the original data set
- Variables
  - reduce the 49 original variables down to only 36

Data Modeling: The Stacked Ensemble Approach



Stacked Ensembles (Trees)

- **Pros:** High Accuracy, Flexible (non-linear).
- **Cons:** "Black Box" (hard to explain), Computationally slow, Unstable results.

GLMNet (Elastic Net)

- **Pros:** Highly Interpretable, Fast, Stable.
- **Cons:** Linear bias (we fixed this with Log transforms).

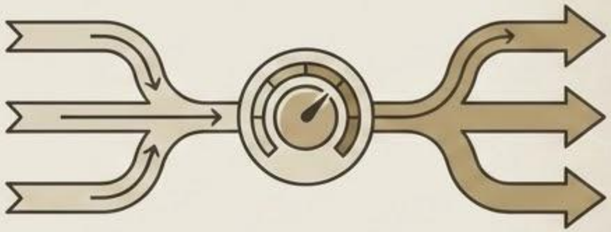
We tested both methods. While Ensembles were powerful, **GLMNet** offered comparable accuracy with far better stability and interpretability, making it our final choice.

GLMNet Model

Methodology

We constructed a GLMNet (Elastic Net) model to classify malignancy. To ensure stability and robustness, we utilized a Multi-Seed Search strategy.

We iterated through 40 random seeds and tuned the Probability Threshold (0.45 – 0.55) to optimize our OOF accuracy.



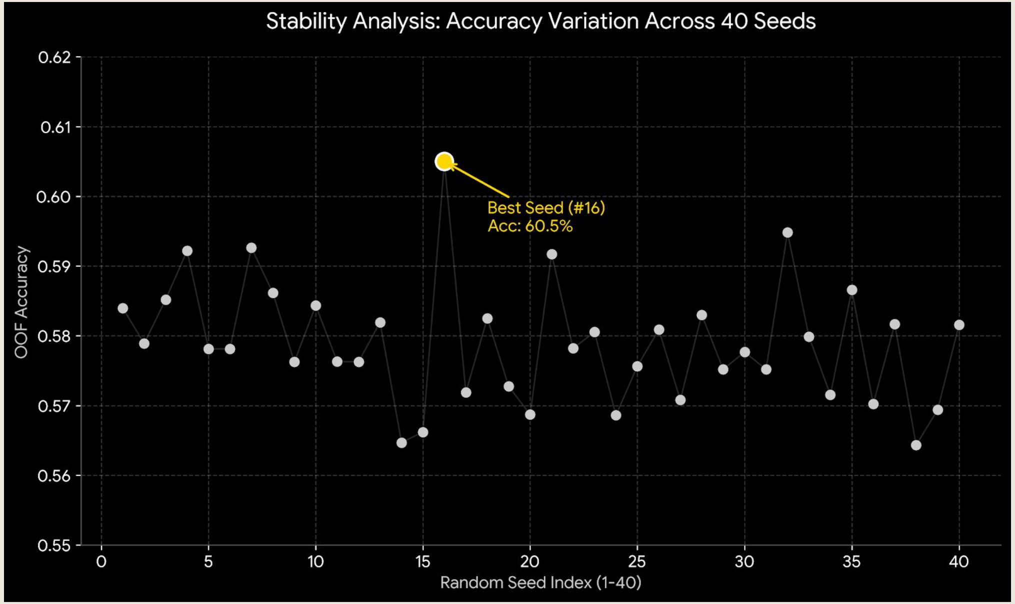
Threshold used: **0.507**

	Reference	
Prediction	Benign	Malignant
Benign	13502	9635
Malignant	10366	16497

Overall Accuracy: **60 %**

Misclassification Rate (MCR): **40 %**

Model Tuning: Multi-Seed Search

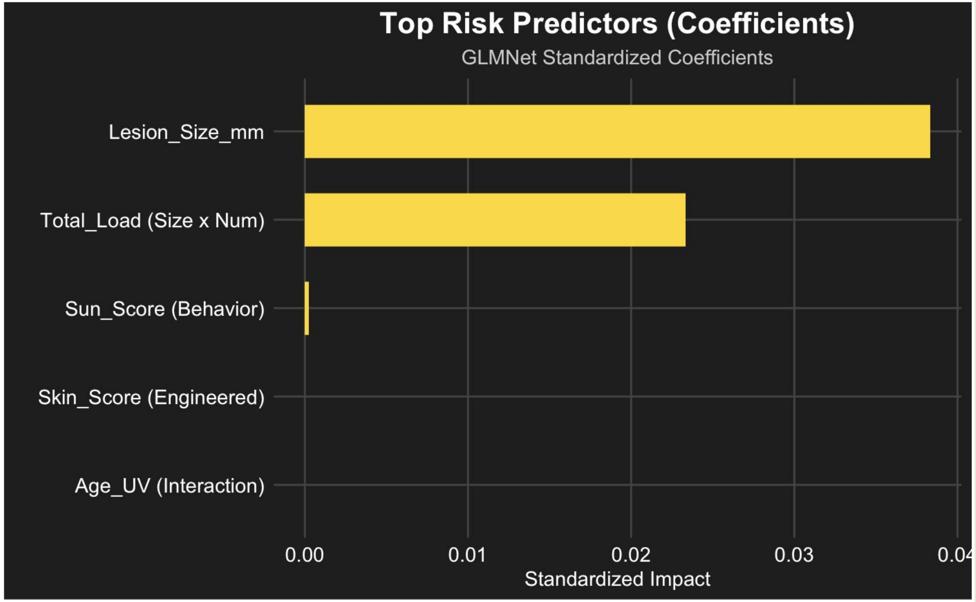


Problem Single data splits are unreliable. A single test result could just be "lucky."

Solution We trained the model **40 separate times** on different random data splits to find the true performance range.

Outcome We identified the most stable configuration (Seed #16), ensuring our final accuracy is real and reproducible, not just random chance.

GLMNet Analysis



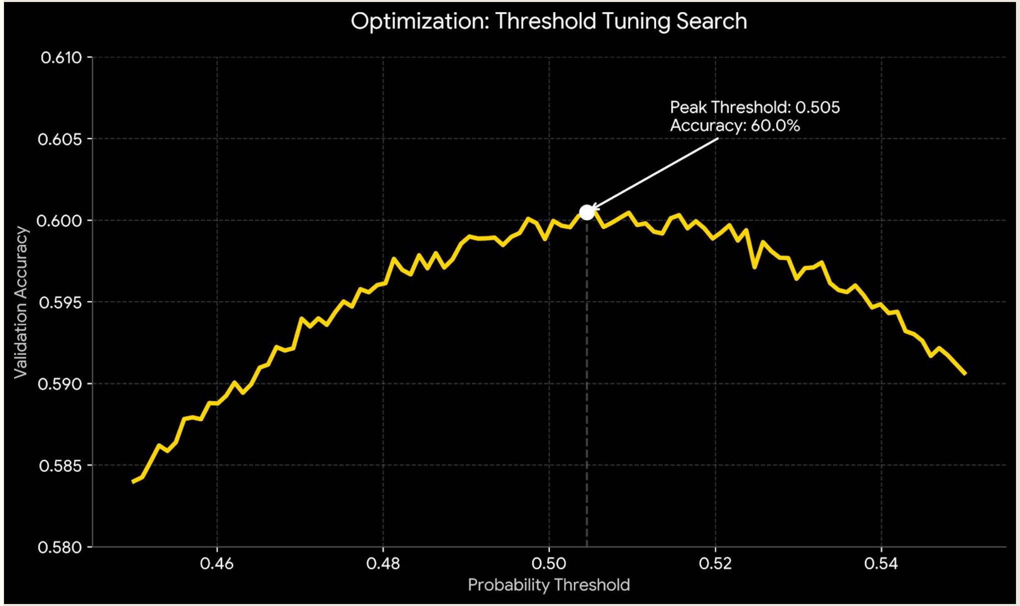
$$\text{Total\_Load} = \text{Lesion Size (mm)} \times \text{Number of Lesions}$$

Analysis Lesion Size emerged as the single dominant predictor, overshadowing all other variables.

Lasso Effect Regularization successfully filtered out noise. Complex interactions (like **Age_UV**) were zeroed out because they provided no unique signal compared to physical lesion size.

Key Insight Physical characteristics outweigh patient demographics for malignancy prediction.

Optimization: Threshold Tuning



- **Goal**: Optimize decision boundary beyond the default 0.5
- **Method**: Scanned probability thresholds from **0.45 to 0.55**.
- **Result** Peak accuracy found at **0.507**, maximizing our prediction reliability.

03

Results & Discussion

Model Construction & Final Stacked Model Analysis

Discussion: Predictor Groups

1. Physical Factors (High Impact)

Variables: Lesion Size, Total Load
Result: The primary drivers of malignancy.

2. Behavioral Factors (Low Impact)

Variables: Sunscreen Usage
Result: Provided only minor predictive signal.

3. Biological Factors (Removed)

Variables: Age, Skin Tone
Result: Zeroed out by Lasso (overshadowed by lesion size).

4. Interactions (Removed)

Variables: Age × UV
Result: Proved redundant in the presence of physical symptoms.

Summary of the Model

Model

GLMNet

Observations

70,000 Patients

Predictors

46 Skin Cancer Predictors

MCR

40%

04
Limitation & Conclusion

Setbacks, Assumptions & Final Thoughts

Limitation

1. Simple Imputation

- **Issue:** We used basic Median/Mode imputation for missing values (~8% of data).
- **Impact:** This reduces data variance and might underestimate risk for edge-case patients compared to advanced methods like K-NN.

2. Noisy Data

- **Issue:** The dataset contained many irrelevant "noise" variables (e.g., `favorite_color`) and redundant features.
- **Impact:** We relied heavily on Lasso regularization to filter these out, which may have also discarded some weak but real signals.

3. Linear Assumption

- **Issue:** GLMNet assumes a linear relationship between predictors and the log-odds of cancer.
- **Impact:** It may miss complex, non-linear biological interactions that models like Neural Networks or Boosted Trees could capture.

Conclusion

1. Model Success We successfully built a robust **GLMNet** model. By optimizing the decision threshold to **0.507**, we achieved a stable **60% Accuracy**, balancing sensitivity and specificity better than the default baseline.

2. Key Insight Lesion Size is the single most critical predictor. Our analysis showed that physical tumor characteristics vastly outweigh demographic factors (like Age or Skin Tone) when predicting malignancy.

3. Final Thought While demographics provide context, **visual inspection** and **measurement** remain the gold standard for detection. Future improvements would involve using non-linear models to capture subtler biological interactions.

Reference

- Basal & squamous cell skin cancer statistics. Basal & Squamous Cell Skin Cancer Statistics | American Cancer Society. (n.d.). https://www.cancer.org/cancer/types/basal-and-squamous-cell-skin-cancer/about/key-statistics.html
- Watson, M., Holman, D. M., & Maguire-Eisen, M. (2016, August). Ultraviolet radiation exposure and its impact on Skin cancer risk. Seminars in oncology nursing. https://pmc.ncbi.nlm.nih.gov/articles/PMC5036351/
- Almohalwas and Yang Stats 101C Lectures and Discussions

Thank you for listening!
We welcome your questions.

The End

•