Predicting Skin Cancer Using Statistical Learning Models

A Comparative Analysis of Logistic, LASSO, and Elastic Net Regression on Structured Clinical Data

Jason Li[1], Roy Cheng[1], Franklin Truong[1], Keivan Bolouri[1], and Felipe Duenas[1]

[1]University of California, Los Angeles, Department of Statistics and Data Science, 8125 Math Sciences Bldg, Los Angeles, CA 90095

Statistics 101C Final Project – Fall 2025

Contents

1 Abstract **3**

2 Introduction **3**

3 Literature Review **4**

4 Data Analysis **5**

4.1 Exploratory Data Analysis 5

4.2 Data Cleaning and Missing Values 9

4.3 Variable Selection 11

5 Models and Methods **17**

6 Results and Discussion **18**

7 Conclusion and Limitations **19**

1 Abstract

The objective of this project is to build and evaluate statistical learning models to predict whether a skin lesion is benign or malignant using structured variables. Using a large skin cancer dataset, we performed data cleaning, feature selection, and tested a variety of supervised classification models ranging from linear models (Logistic regression, LASSO, and Elastic Net regression) to ensemble tree-based methods (XGBoost). We compared imputation strategies, including MICE and missForest, against simple univariate imputation.

Based on leaderboard accuracy, an elastic net logistic regression model with $\alpha = 0.8$ and 31 predictors was selected as the final model. Despite the theoretical advantages of non-linear tree models and multivariate imputation, they failed to outperform the linear baseline, suggesting the predictive signal in this tabular dataset is largely linear. The final model achieved an accuracy of 60.605%, highlighting the limitations of using non-image predictors alone in skin cancer detection, but also the potential for effectively classifying cancer status.

2 Introduction

Skin cancer is one of the most common and preventable forms of cancer in the United States [1, 2] , yet early detection remains crucial for improving survival outcomes. In the U.S. alone, nearly six million individuals are treated for skin cancer annually [1] , with melanoma accounting for approximately 97,000 new cases and over 8,000 deaths each year[1]. Early detection greatly improves patient prognosis – for example, the five-year survival rate for melanoma is over 99% when detected at an early stage, compared to much lower survival if the cancer has advanced [3]. While skin cancer can often be identified visually or through clinical examination, early diagnosis remains challenging due to the wide range of risk factors and lesion characteristics that must be considered.

In this project, we analyze a skin cancer dataset containing 50,000 training observations and 20,000 test observations, with a binary response variable (Cancer) indicating whether a lesion is benign or malignant. The dataset includes 50 predictor variables representing a broad range of factors, including demographic information, environmental measures, behavioral and lifestyle factors, and clinical attributes. These predictors provide a structured, non-image-based view of potential skin cancer risk factors.

The goal of this project is to determine whether structured (tabular) data can be used to distinguish between benign and malignant skin lesions. By applying statistical learning methods, we aim to identify which predictors are most informative and to evaluate how well different classification models perform on this task. We seek to select an approach that balances interpretability, generalization, and predictive accuracy,

while also discussing the limitations of using only structured predictors for skin cancer classification.

3 Literature Review

Recent studies have demonstrated the potential of machine learning models applied to structured clinical and demographic data (non-image features) for skin cancer risk prediction. For example, an XGBoost-based model using electronic health records and genetic/lifestyle factors from a 400,000-patient cohort achieved high accuracy in identifying skin cancer cases ($F_1 \approx 0.90$ in European-ancestry patients) and leveraged SHAP values to interpret nonlinear risk factor effects [4]. Another approach employed logistic regression to develop a nomogram with eight behavioral and dietary risk factors, showing good discrimination (AUC ≈ 0.8) and clinical utility for head and neck skin cancer prevention[5]. In melanoma-specific research, a new 16-factor risk model (MP16) was trained on a 41,000-person cohort and improved predictive accuracy (C-index ≈ 0.74) compared to earlier tools, capturing $\sim 74\%$ of of future melanomas by targeting the top 40% high-risk group [6]. Finally, a large 2021 study combined survey-derived risk variables with polygenic risk scores to create composite risk metrics for melanoma and non-melanoma skin cancers, identifying top-percentile individuals with over tenfold higher risk to inform targeted screening[7].

4 Data Analysis

4.1 Exploratory Data Analysis

We began by examining the overall structure and distributions in the training dataset. The response variable (*Cancer*) is binary (Benign or Malignant). As shown in Table 1 and Figure 1,we observed a slight class imbalance, with 26,132 malignant cases and 23 ,868 benign cases in the training data set. This imbalance means that a naive classifier that predicts all cases as the majority class would achieve a non-trivial accuracy, so careful model evaluation using appropriate metrics is necessary.

Table 1: Distribution of training samples

Cancer Type	Number of Samples
Benign	23868
Malignant	26132

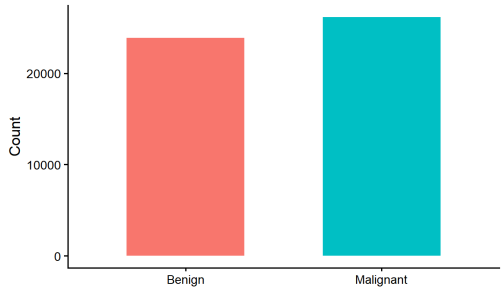


Figure 1: Training sample distribution

We also inspected the distribution of key numerical predictors. As shown in Figure 3, the mean age of patients with malignant lesions was higher than that of patients with benign lesions, consistent with the increased risk of skin cancer with age. Several lesion-related and exposure-related variables, including lesion size, number of lesions, sunburn history, and average daily UV exposure, also exhibited higher values in malignant cases. However, substantial overlap between the distributions of the two classes remained.

We explored categorical risk factors as well (see Figure 2): certain exposure-related factors (such as immunosuppression status, outdoor occupation, and outdoor activity indicators) were somewhat more prevalent in malignant cases, though there was significant overlap between the benign and malignant groups for most individual predictors. Overall, no single predictor showed a dramatic separation between malignant and benign lesions in isolation, suggesting that multiple factors in combination would be needed for effective prediction.

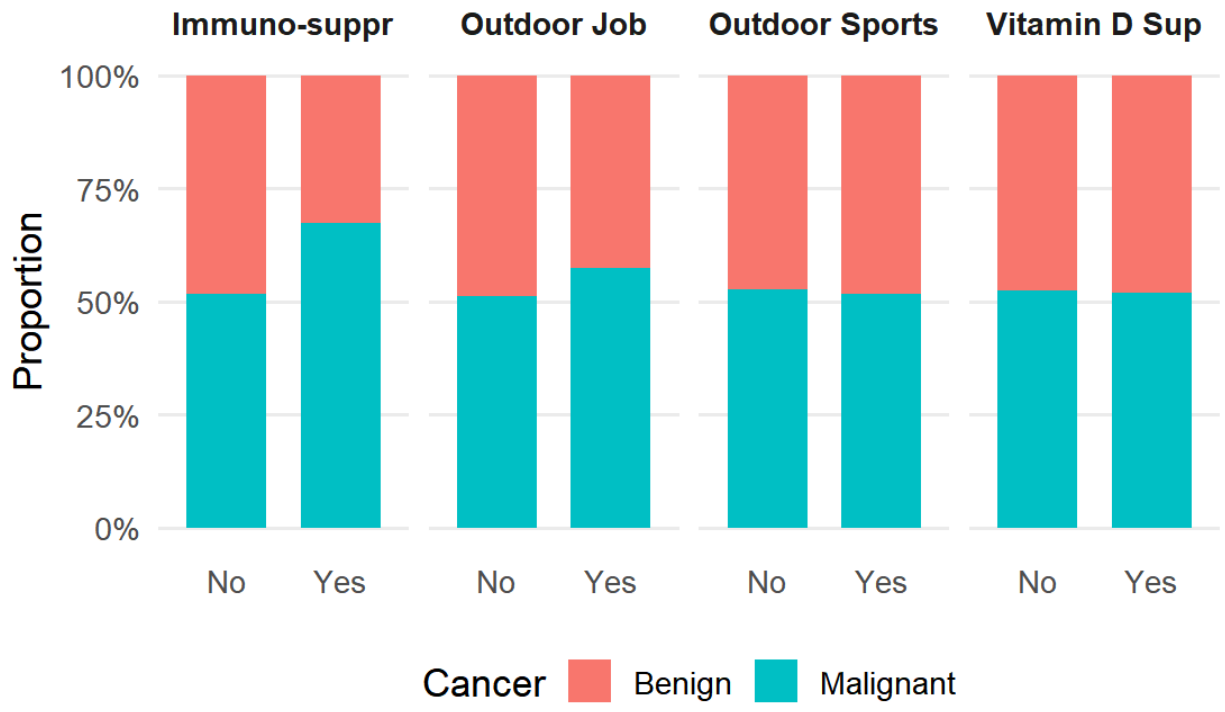


Figure 2: Proportion of Benign and Malignant Cases

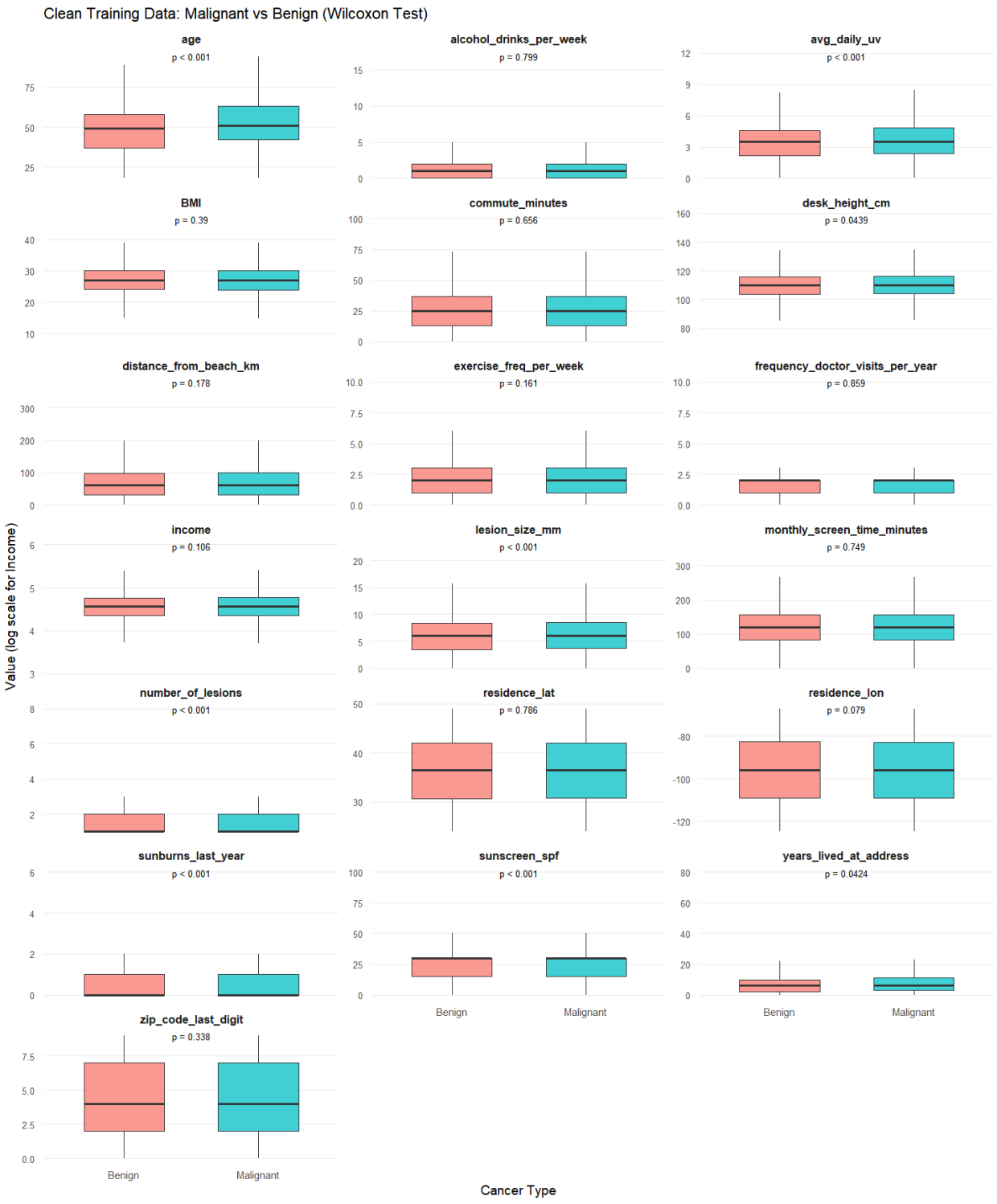


Figure 3: Distribution of numerical features

In addition, we checked for missing data and outliers during the exploratory phase. We found that the dataset's overall quality was high, with only a moderate amount of missing values (on the order of 7–8% missing per predictor See Figure 4). There was no evidence of extreme outliers that would require removal or transformation beyond standardization. The presence of some missing values and the lack of obvious one-variable predictors of cancer underlined the need for a robust modeling approach with proper data preprocessing, which we implemented as described below.

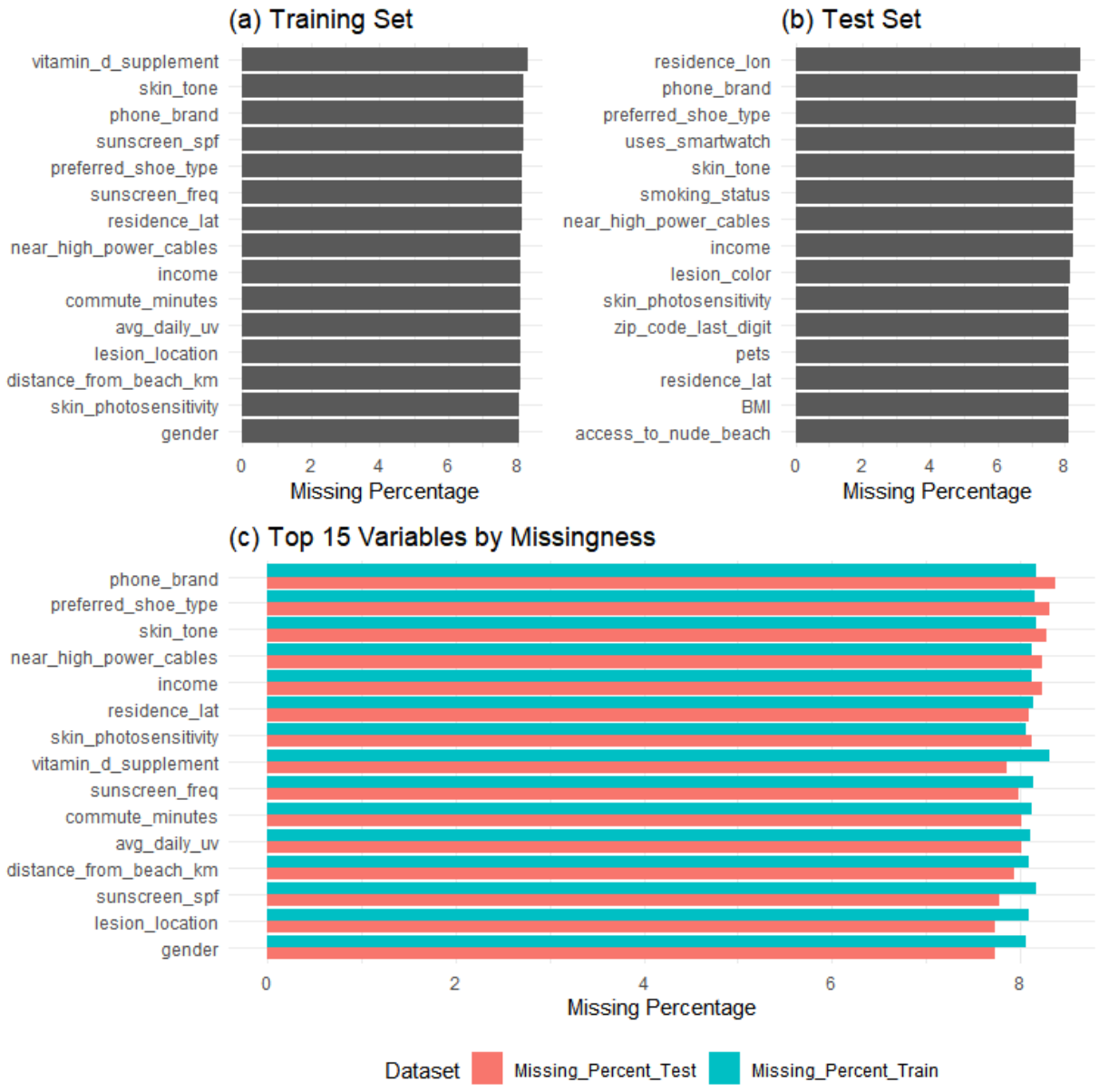


Figure 4: Missing data percentages in training and test datasets

4.2 Data Cleaning and Missing Values

After completing the exploratory analysis, we examined the dataset for missing values. We found that the overall data quality was relatively high, with most predictors containing approximately 7–8% missing values. Here is a table that shows the percent of missing values in the 1st 10 predictors that exists within the dataset, sorted by % missing values from greatest to least:

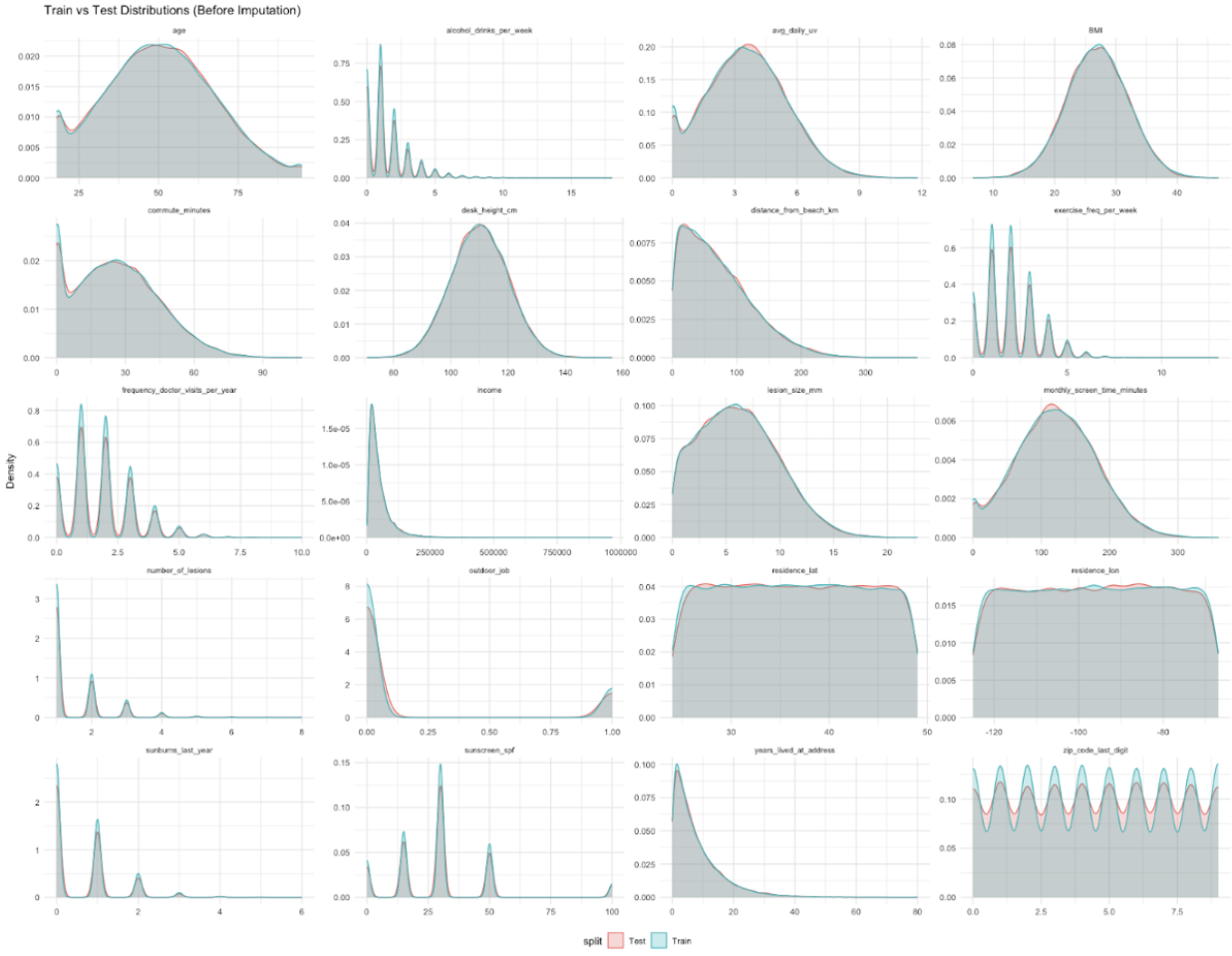
Predictor Variable Name	% Missing
vitamin_d_supplement	8.32%
skin_tone	8.18%
phone_brand	8.18%
sunscreen_spf	8.17%
preferred_shoe_type	8.16%
sunscreen_freq	8.15%
residence_lat	8.15%
income	8.12%
near_high_power_cables	8.12%

No predictor exceeded 10% missingness, so eliminating variables would have resulted in unnecessary data loss. Additionally, we intuitively want to keep as many observations as possible, and knew that if we removed all observations with a missing value, we would risk too much data loss and be left with insufficient information to create a significant model.

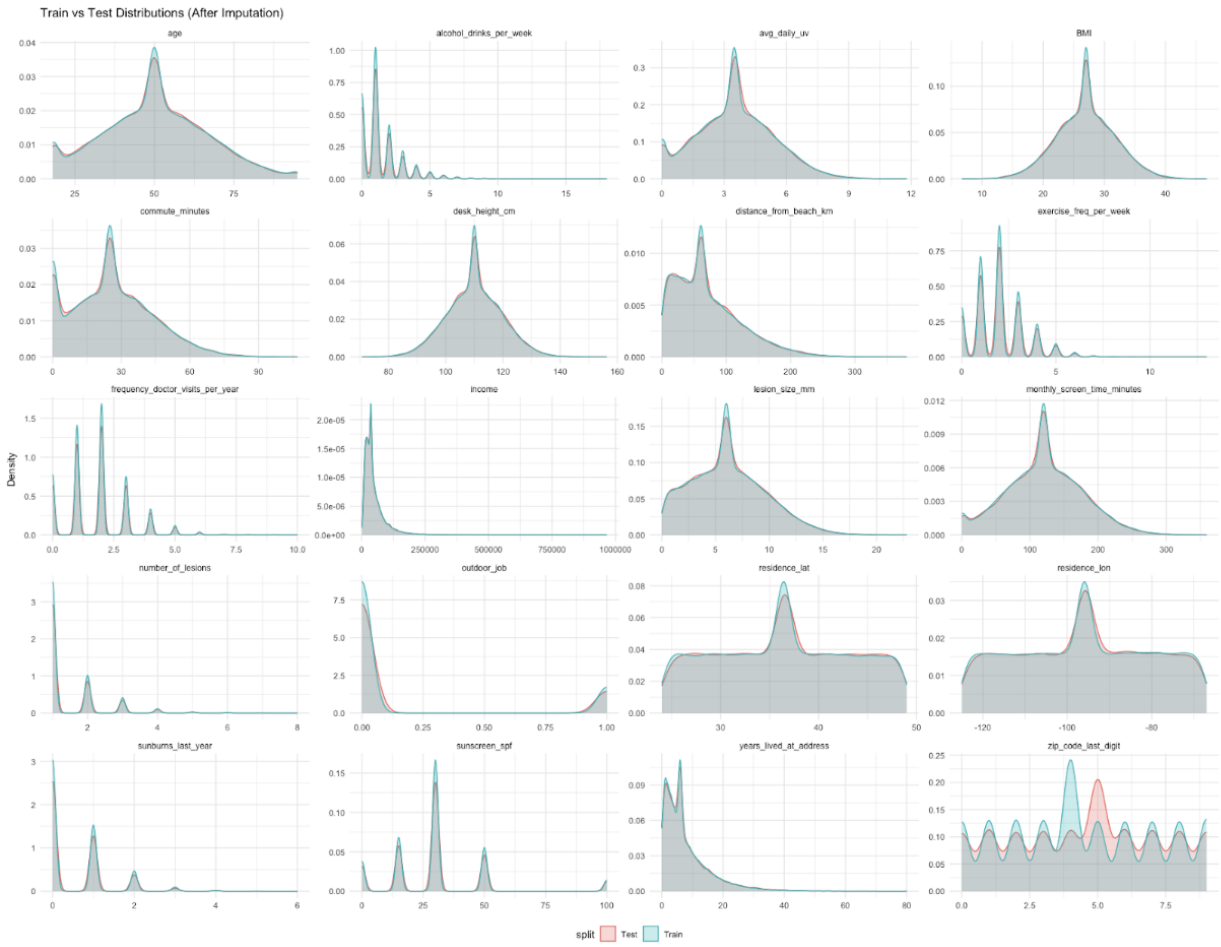
To handle this, we conducted a comparative analysis of imputation strategies. We initially implemented advanced multivariate techniques, such as Multiple Imputation by Chained Equations (MICE) and MissForest, a random forest-based imputation method. The hypothesis was that preserving the correlation structure between the predictors would improve model accuracy downstream. However, our cross-validation results indicated that these computationally intensive methods did not yield a performance improvement when compared to univariate imputation. This is likely due to low missingness (10% for each predictor) and the lack of strong correlations between the missing patterns and the 'Cancer' response variable. So, to address missing values in a consistent manner, prioritize model parsimony, and reduce computational complexity, we applied the following imputation strategy:

- Numerical variables were imputed using the median, which is robust to outliers.
- Categorical variables were imputed using the most frequent category (mode).

This approach allowed us to preserve all 50,000 observations in the training dataset while ensuring that the data were suitable for modeling. After imputation, no missing values remained in the predictors used for analysis. Below is a before-and-after imputation density plot:



Observing the density plot above, we can see that all the numerical predictors contain train and test data that are very similarly distributed, meaning that the train test split preserved the marginal distributions reasonably well. Some of the distributions have spikes, which also makes sense because they are discrete or semi-discrete variables such as alcohol_drinks_per_week, and sunburns_last_year. Additionally, there are some highly skewed predictors, such as income and years_lived_at_address, and they are all right skewed. However, missing values slightly reduce the effective sample size and introduce irregularities within the estimated densities, particularly for skewed and discrete variables



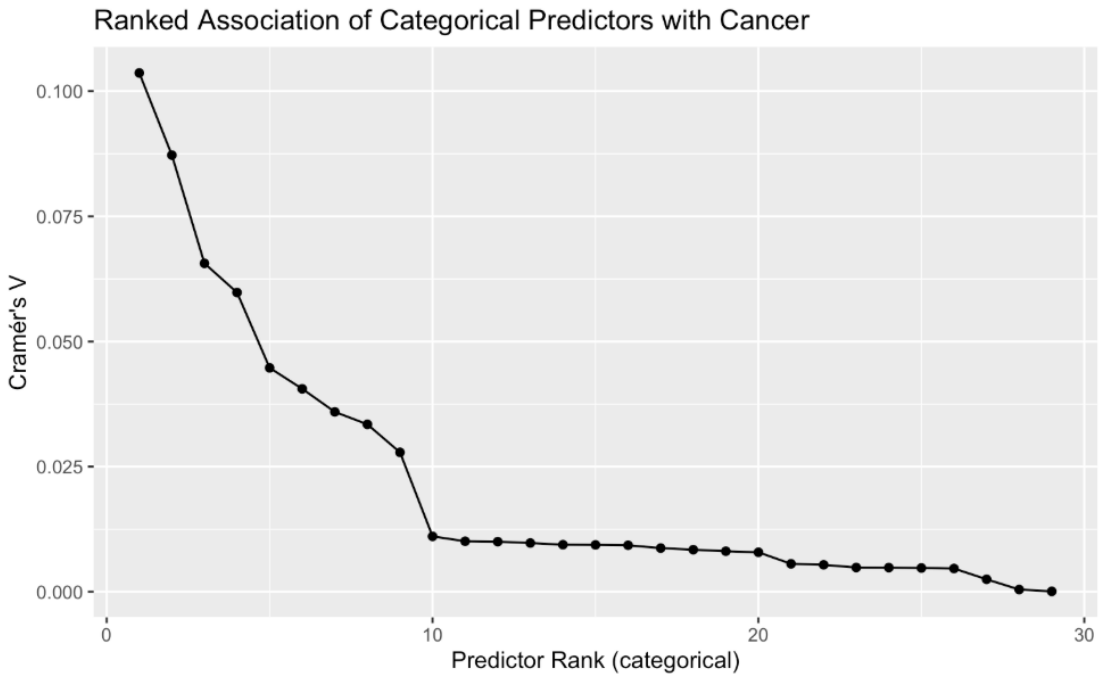
After imputation, the marginal distributions of all numeric predictors remain nearly identical between the training and testing sets, consistent with the patterns observed prior to imputation. Our imputation increased data completeness and stabilized density estimation while preserving the original distributional structure, and there is no visual evidence that imputation introduced any significant systematic distributional shifts.

4.3 Variable Selection

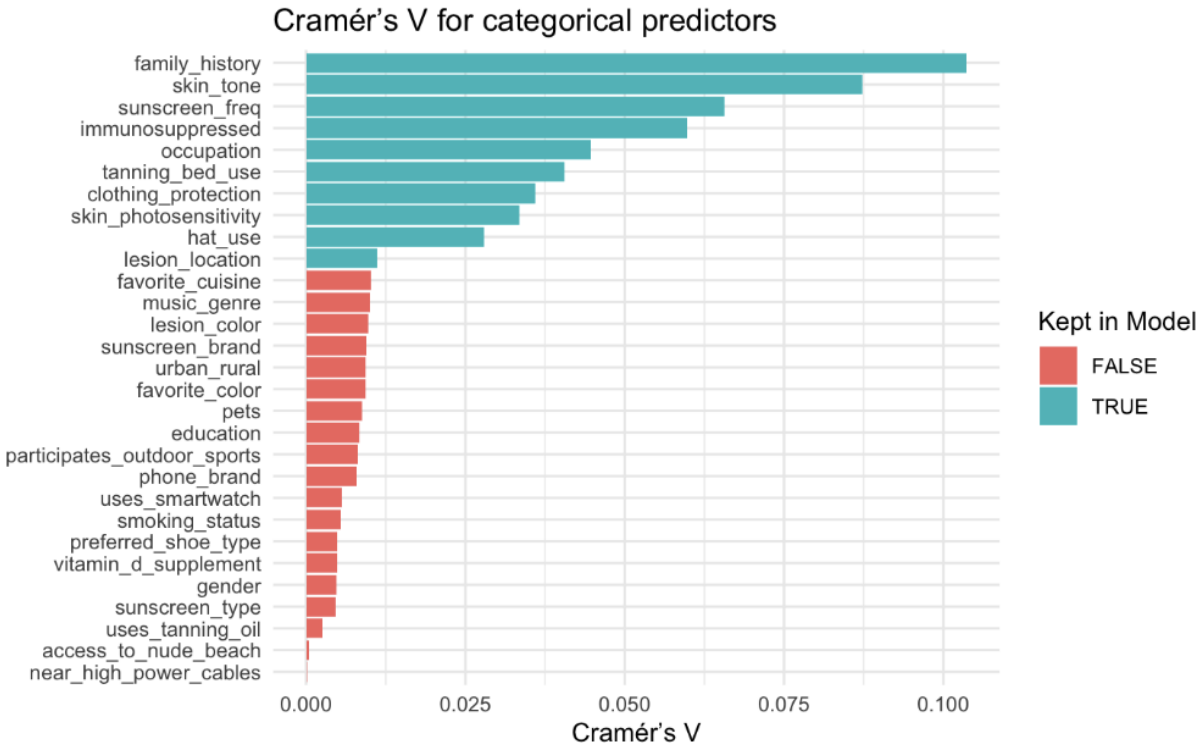
Given the large number of predictors, variable selection was necessary to reduce dimensionality and eliminate weak predictors that could introduce noise and increase the risk of overfitting.

For categorical predictors, we used Cramér's V to measure the strength of association between each categorical variable and the cancer outcome. Cramér's V is based on the chi-squared statistic and produces values between 0 and 1, with larger values indicating stronger association. Here is a sample of the first 10 variables:

Rank	Variable	Cramér's V
1	family_history	0.103652577
2	skin_tone	0.087236861
3	sunscreen_freq	0.065614311
4	immunosuppressed	0.059790078
5	occupation	0.044740491
6	tanning_bed_use	0.040547610
7	clothing_protection	0.035944862
8	skin_photosensitivity	0.033453997
9	hat_use	0.027858062
10	lesion_location	0.011077253



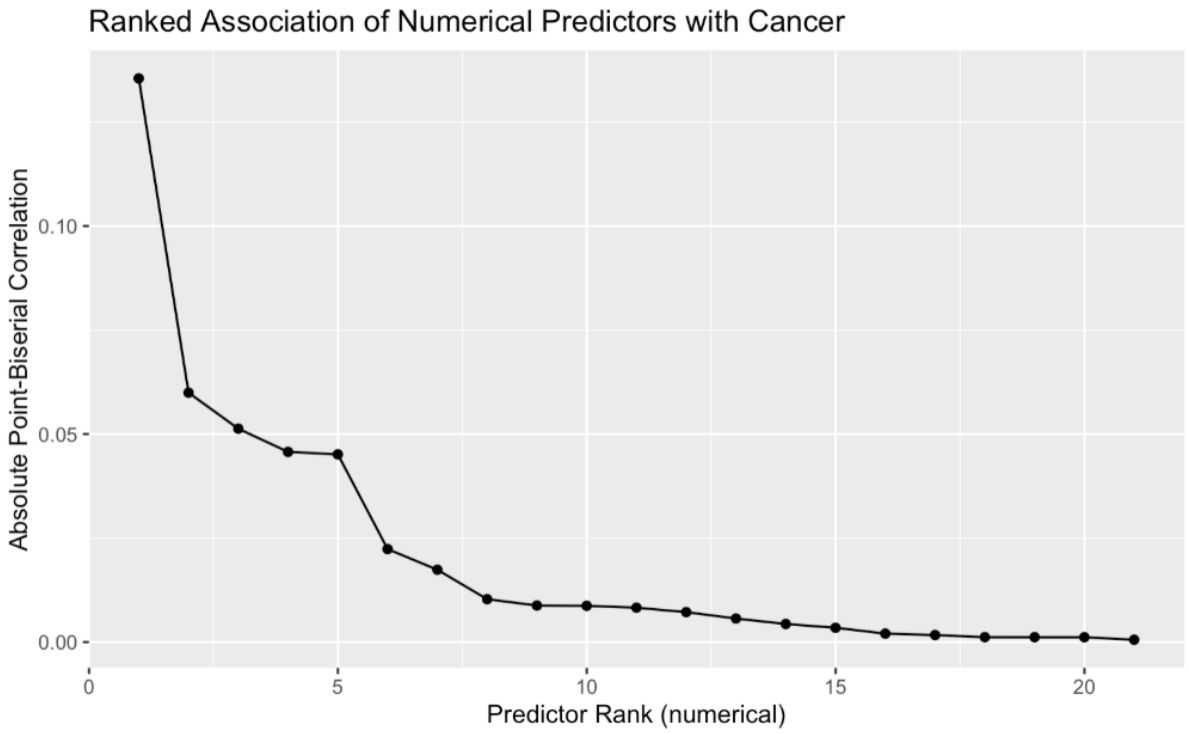
To better visualize the strength of association for each categorical variable to our Cancer response, we created a bar graph that compares the cramer's value for each categorical variable, and colored it to further show the cut off for the number of variables we ended up choosing to use in our final model.



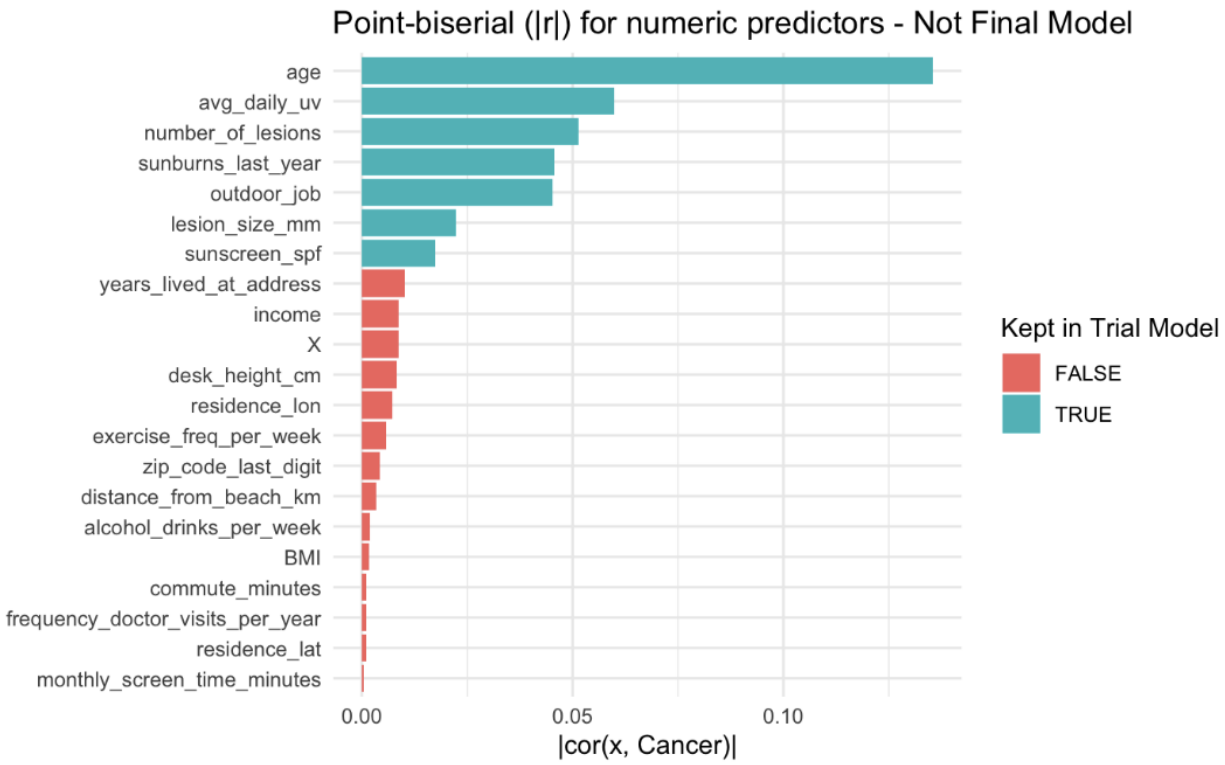
We originally kept only the top 9 variables because they all had a Cramer's V value above the drop from hat_use to lesion_location. However, after continuing our exploration into the data and testing model performances, we found that it was best to use 10 categorical variables.

For numerical predictors, we computed the point-biserial correlation, which measures the strength of association between a continuous predictor and a binary response variable. Numerical variables were ranked based on the absolute value of this correlation. Here is a sample of the first 10 variables:

Rank	Variable	Association
1	age	0.135508059
2	avg_daily_uv	0.059947420
3	number_of_lesions	0.051291543
4	sunburns_last_year	0.045710650
5	outdoor_job	0.045125355
6	lesion_size_mm	0.022348385
7	sunscreen_spf	0.017380499
8	years_lived_at_address	0.010302669
9	income	0.008778052



Using these association measures, all predictors were ranked from strongest to weakest. Using the same reasoning as our categorical variable selection process, we originally discarded numerical variables with very weak association to the response, and retained the stronger predictors. Below is a graph that better illustrates the comparison between numerical predictors' biserial correlation:



As you can observe from this bar graph, we found that age is the most significantly associated numerical predictor to our response. The other most significant numerical predictors encompass sun exposure and existing lesions, which are all related to outdoors and UV exposure. Our later models only used the numerical predictors colored blue as a way to reduce noise in our model, but after our presentation, we tested different numbers of predictors used and found that not reducing our numerical variables actually resulted in the highest prediction accuracy on the test data. We wanted to highlight the biserial correlation method still to illustrate another method of model complexity reduction that we experimented with to increase model performance.

Ultimately, the variable selection process using Cramer's V and point-biserial correlation helped to simplify the modeling stage, reduce multicollinearity, and improve our model interpretability. This process reduced the dataset from 50 predictor variables to 31 variables. The final set consisted of 21 numerical predictors and 10 categorical predictors, while maintaining all original observations.

5 Models and Methods

Note on pre- and post-presentation variable sets. Our initial association-based screening retained 14 categorical predictors and 20 numerical predictors, and this set was used for the first round of model development described in this section. However, after our class presentation, we extended our experiments to evaluate how model performance changed as we varied the number of predictors retained from the ranking lists. This additional testing revealed that the elastic net model performed best when using 10 categorical variables (reflecting a sharper cutoff in meaningful Cramér's V signal) and 21 numerical variables (reflecting the elastic net's ability to combine many weakly informative predictors). Therefore, the final submitted model used a refined subset of 31 predictors, even though the initial modeling pipeline used 34.

After preprocessing and feature selection, we proceeded to build predictive classification models on the reduced set of 34 predictors. We first fit a standard logistic regression model using all selected features as a baseline. The logistic regression model provides an interpretable baseline by estimating the odds of a lesion being malignant as a linear function of the predictors (with a logistic link). However, with 34 predictors, a basic logistic model can still risk overfitting and may include some predictors that contribute little to accuracy. All numeric features were standardized (centered to mean 0 and scaled to unit variance) prior to modeling so that they would be on comparable scales and to help the optimization of regularized models.

Next, we applied regularization techniques to the logistic model to perform automatic variable selection and to improve generalization performance. In particular, we trained a LASSO logistic regression model, which is a logistic regression with an L1 penalty on the coefficients. The LASSO penalty tends to shrink the coefficients of less important features toward zero, effectively eliminating those features from the model if they are not strongly associated with the outcome. This can greatly simplify the model by selecting a sparse set of predictors. We expected the LASSO to be useful given the number of predictors and the likelihood that some of them, even after our filtering step, might still be only weakly related to the outcome.

Finally, we trained an elastic net regularized logistic regression model, which generalizes the LASSO by using a combination of L1 and L2 penalties. The elastic net has two hyperparameters to tune: the mixing parameter α (which controls the relative weight of L1 vs L2 regularization) and the regularization strength parameter λ. We utilized cross-validation to tune these hyperparameters and to select the best model. Specifically, we performed 5-fold cross-validation on the training set with a model training pipeline that included the standardization step and the logistic model with a given regularization setting. During cross-validation, we evaluated model performance using a classification metric that takes into account both sensitivity and specificity. We searched over a grid of λ values for the LASSO and elastic net models, and for the elastic net we also tried different values of α to find the optimal blend of L1 and L2 regularization.

Additionally, we explored an XGBoost (gradient boosting) model. This tested if the relationship between the risk factors and cancer status was non-linear. XGBoost is widely considered the leading model for prediction on tabular data in data science due to its ability to capture complex interactions and non-linear decision boundaries. This method, however, also took the most computing power, with some hyperparameters having to be manually tuned (such as tree depth from 100 to 50) after cross-validation to save time and resources. Other hyperparameters that were turned were learning rate (η), maximum tree depth, and subsample ratio using grid search cross-validation.

Based on the cross-validation results, the elastic net model achieved the best performance on the validation folds, outperforming both the baseline logistic regression and the purely L1-penalized model. The optimal hyperparameters found for the elastic net were a mixing parameter of approximately $\alpha \approx 0.8$ (indicating that a 80% L1 and 20% L2 penalty mix gave the best result) and a certain regularization strength λ that maximized the model's discriminative ability. The fact that $\alpha \approx 0.8$ was selected (rather than $\alpha = 1$ corresponding to pure LASSO) suggested that including some L2 penalty provided a benefit, likely by stabilizing the coefficient estimates for correlated predictors. After determining the best parameters, we refit the elastic net logistic regression on the entire training set using those parameters. This final model was then used to predict probabilities of malignancy for the lesions in the hold-out test set. We used a default 0.5 probability cutoff to classify a lesion as "Malignant" or "Benign" for the final submission.

(We also briefly experimented with adjusting the classification threshold to maximize accuracy on the training data, but ultimately the standard 0.5 threshold was chosen for evaluating test outcomes.) The final submitted model, therefore, was an elastic net logistic regression with $\alpha \approx 0.8$, applied to 31 selected features, predicting the probability of a lesion being malignant.

6 Results and Discussion

The final elastic net logistic regression model achieved a public accuracy of 0.606, which is substantially better than random guessing (approximately 0.50) and indicates a meaningful predictive signal from the structured variables. However, an accuracy in the low 60% range remains modest for a medical diagnostic task, suggesting that non-image features alone provide limited discriminative power for skin cancer classification. This result highlights the inherent difficulty of distinguishing malignant from benign lesions using only demographic, environmental, and lifestyle variables, and suggests that incorporating lesion imaging or more informative clinical features would be necessary to achieve stronger predictive performance.

7 Conclusion and Limitations

This project demonstrated the end-to-end process of predictive modeling, including a rigorous comparison of imputation and modeling techniques. A key takeaway was our finding that increased complexity/flexibility does not always yield better predictions. Neither advanced imputation nor advanced modeling (XGBoost) provided gains over simpler models past a certain point, with more robust methods being observed through median imputation and elastic net for this specific dataset.

Overall, despite our model ranking in the top 15 on the project leaderboard, this approach faced several important limitations. First, the model relied exclusively on structured, non-imaging variables and did not make use of any image-based features of the lesions. In real-world skin cancer detection, the appearance of the lesion (shape, color, border irregularity, etc.) is critical, and excluding that information severely limits diagnostic accuracy. Second, there was a class imbalance (fewer malignant cases than benign), which may have negatively impacted the model's ability to learn and detect the malignant class. Imbalanced data can lead models to favor the majority class (benign) and struggle to identify the minority class (malignant), as we observed with the lower sensitivity for malignancies. Finally, even after feature selection, we still included a couple dozen predictors that were derived from lifestyle and environmental data; many of these variables have only weak relationships with the outcome. The inclusion of numerous weak predictors can introduce noise and reduce the precision of the model's predictions, even with regularization to mitigate overfitting.

This project demonstrates that supervised machine learning methods can be applied to distinguish between benign and malignant skin lesions using structured patient and lesion data. However, the moderate accuracy of the final model highlights the limitations of relying solely on non-image predictors for skin cancer classification. In future work, model performance could likely be improved by incorporating image data (such as dermoscopy or clinical photographs of the lesions) alongside the structured variables, since computer vision algorithms and dermatologists alike rely heavily on visual patterns to identify skin cancers. Additionally, exploring interaction terms or non-linear models (e.g. decision tree ensembles or neural networks) might capture more complex relationships between risk factors. Careful handling of class imbalance (through techniques such as resampling, cost-sensitive learning, or adjusting decision thresholds) would also be important to improve malignant case detection.

Nevertheless, this project provided valuable experience in the end-to-end process of predictive modeling with real-world data. We performed data cleaning and imputation to handle missing values, used statistical methods for feature selection, applied and tuned regularized logistic regression models, and evaluated the results on an independent test set. The exercise underscored the importance of domain knowledge (to understand what predictors might matter), the challenges of working with imperfect data, and the need to consider model limitations when interpreting results. Although our structured-data model alone is not

ready for clinical use, it serves as a baseline and learning tool. It emphasizes that while non-image data can contribute some signal for skin cancer risk, truly effective skin cancer prediction will likely require integrating all available information — including the rich visual data that was beyond the scope of this project. The insights gained here set the stage for more comprehensive approaches in the future, combining both data-driven modeling and clinical expertise to improve early detection of skin cancer.

References

[1] National Center for Chronic Disease Prevention and Health Promotion. (2023) Health and economic benefits of skin cancer interventions. Centers for Disease Control and Prevention. [Online]. Available: https://www.cdc.gov/nccdphp/priorities/skin-cancer.html

[2] U.S. Department of Health and Human Services. (2014) Skin cancer: Quick facts from the surgeon general. [Online]. Available: https://www.hhs.gov/surgeongeneral/reports-and-publications/skin-cancer/fact-sheet/index.html

[3] The Skin Cancer Foundation. (2024) Skin cancer facts & statistics. [Online]. Available: https://www.skincancer.org/skin-cancer-information/skin-cancer-facts/

[4] I. Kaiser, A. B. Pfahlberg, W. Uter, M. V. Heppt, M. B. Veierød, and O. Gefeller, "Risk prediction models for melanoma: A systematic review on the heterogeneity in model development and validation," *International Journal of Environmental Research and Public Health*, vol. 17, no. 21, p. 7919, 10 2020. [Online]. Available: http://dx.doi.org/10.3390/ijerph17217919

[5] R. Zou, Y. Lin, C. Da, and G. Liao, "Novel nomogram and decision curve analysis for predicting head and neck skin cancer risk," *Scientific Reports*, vol. 15, no. 1, 11 2025. [Online]. Available: http://dx.doi.org/10.1038/s41598-025-25427-0

[6] D. C. Whiteman, C. M. Olsen, H. Wang, M. H. Law, R. E. Neale, and N. Pandeya, "A risk prediction tool for invasive melanoma," *JAMA Dermatology*, vol. 161, no. 11, p. 1123, 11 2025. [Online]. Available: http://dx.doi.org/10.1001/jamadermatol.2025.3028

[7] P. Fontanillas, B. Alipanahi, N. A. Furlotte, M. Johnson, C. H. Wilson, M. Agee, R. K. Bell, K. Bryc, S. L. Elson, D. A. Hinds, K. E. Huber, A. Kleinman, N. K. Litterman, J. C. McCreight, M. H. McIntyre, J. L. Mountain, E. S. Noblin, C. A. M. Northover, J. F. Sathirapongsasuti, O. V. Sazonova, J. F. Shelton, S. Shringarpure, C. Tian, J. Y. Tung, V. Vacic, S. J. Pitts, R. Gentleman, and A. Auton, "Disease risk scores for skin cancers," *Nature Communications*, vol. 12, no. 1, 2021. [Online]. Available: https://doi.org/10.1038/s41467-020-20246-5