Prediction of Erythema

Yi Sun
Data Science Institute, Brown University
https://github.com/suchen01/Data1030_final_project.git

1. Introduction

Erythema, a dermatological condition manifesting as noticeable skin reddening, is more common and clinically significant than widely recognized. Beyond its primary symptom of skin redness, erythema can lead to severe complications, including hair loss and scaling on the scalp, which are particularly distressing for patients. This skin response varies from simple irritation to indications of more serious inflammatory diseases. Despite its prevalence, there is a notable lack of public awareness regarding erythema, underscoring an urgent need for focused research and education. The increasing incidence of hair loss and erythema among younger generations, possibly exacerbated by lifestyle factors like extensive cell phone use, signals a shift in dermatological health patterns. This paper aims to tackle these concerns by developing a machine-learning pipeline for predicting erythema based on a dermatology dataset from Kaggle.[1]

The dataset employed in this paper comprises 12 features, with erythema as the target variable. The feature 'family history', is the only binary feature, with a value of 1 indicating a known family history of related diseases, and 0 signifying no such history. 'Age' is the only continuous attribute with a small fraction of missing values. The remaining 10 features are clinical, encompassing various symptoms like scaling, redness, and others. These clinical features and the target variable ‘erythema’ are quantitatively assessed on a scale ranging from 0 to 3. On this scale, 0 denotes the absence of the symptom, 3 reflects its highest observable degree, and 1 and 2 represent intermediate severities. Since the target variable is ordinal, I will approach it as a regression problem.

A recent research conducted by Rahul Ranjan utilized artificial intelligence (AI), specifically machine learning and deep convolutional neural networks (CNNs), for classifying radiation-induced skin reactions (RISRs).[2] Using Scarletred® Vision for digital skin imaging and analyzing 2263 images from 209 patients, the study achieved over 70% accuracy in distinguishing healthy skin from erythema. For a more detailed 3-class severity prediction, accuracies ranged between 60-67%. An ensemble CNN approach further improved the performance, particularly in a 2-class setup, reaching an accuracy of 87%. The study concluded that these AI-based methods could serve as effective pre-screening and decision-support tools for evaluating erythema severity, potentially aiding in the management of skin toxicity in cancer patients. This research represents a pioneering effort in using AI for erythema assessment in radiation dermatitis, offering a promising approach for integrating AI-based erythema scoring into clinical practice.

2. Exploratory Data Analysis

The class distribution of the target variable 'erythema,' which is ordinal in nature, is depicted in Figure 1. This graphical representation provides a clear indication of an imbalanced distribution among the classes. Notably, Class 2 emerges as the majority class, indicating a higher frequency of this severity level in the dataset. Conversely, Class 0 is identified as the minority class, reflecting its lower occurrence.

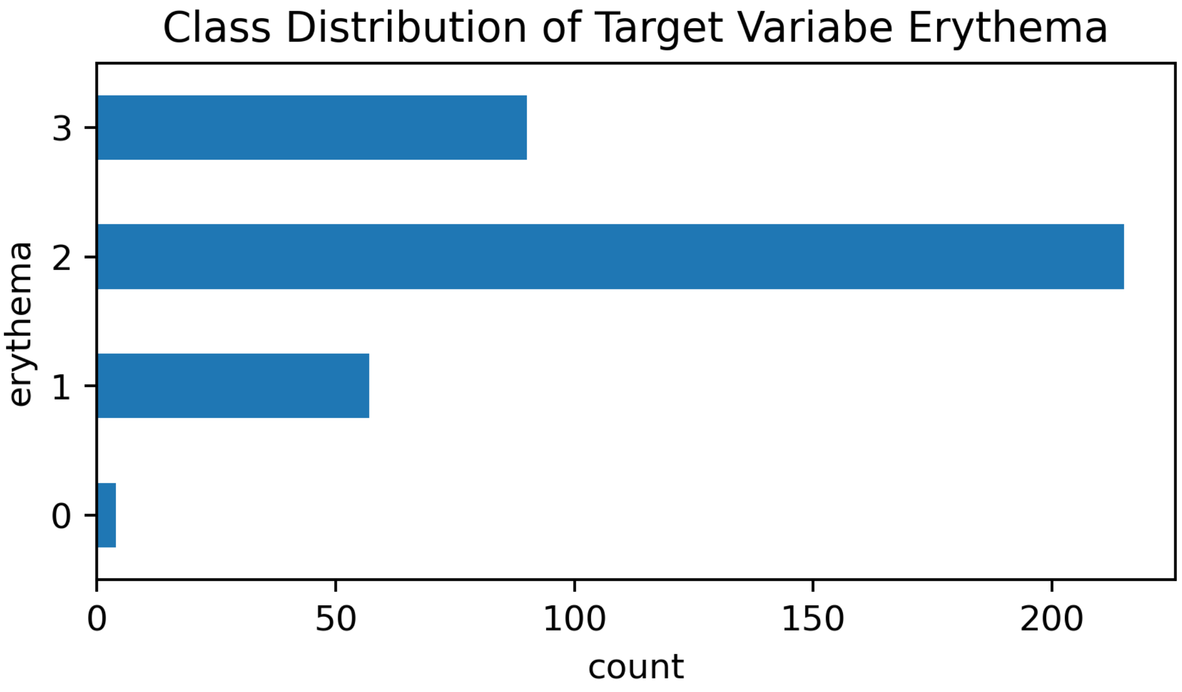


Figure 1: This barplot shows the class distribution of the target variable. Class 2 is the majority class and class 0 is the minority class.

In addition to analyzing the target variable, this paper also examined the class distributions of two other ordinal features: scaling and itching, as depicted in Figure 2 and Figure 3, respectively. This exploration is crucial given their ordinal nature and potential impact on the overall analysis. The class distribution for the ‘scaling’ feature exhibits a significant imbalance. Class 2 is predominant, serving as the majority class, while Class 0 is notably the least represented, establishing it as the minority class. On the other hand, the distribution of the’ itching’ feature presents a contrasting scenario. It appears to be approximately balanced across its classes, suggesting a more uniform occurrence of this symptom in the dataset.

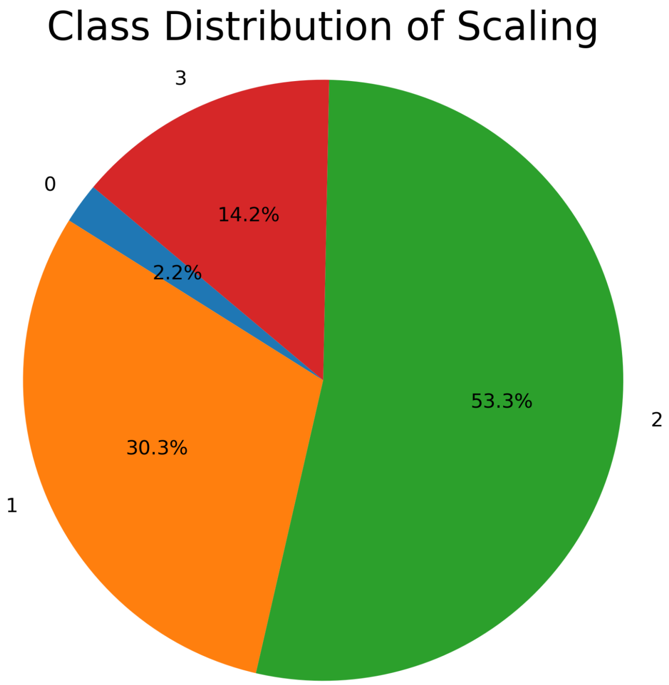


Figure 2: A pie chart of the class distribution of ‘scaling’. Class 2 is the majority class and class 0 is the minority class.

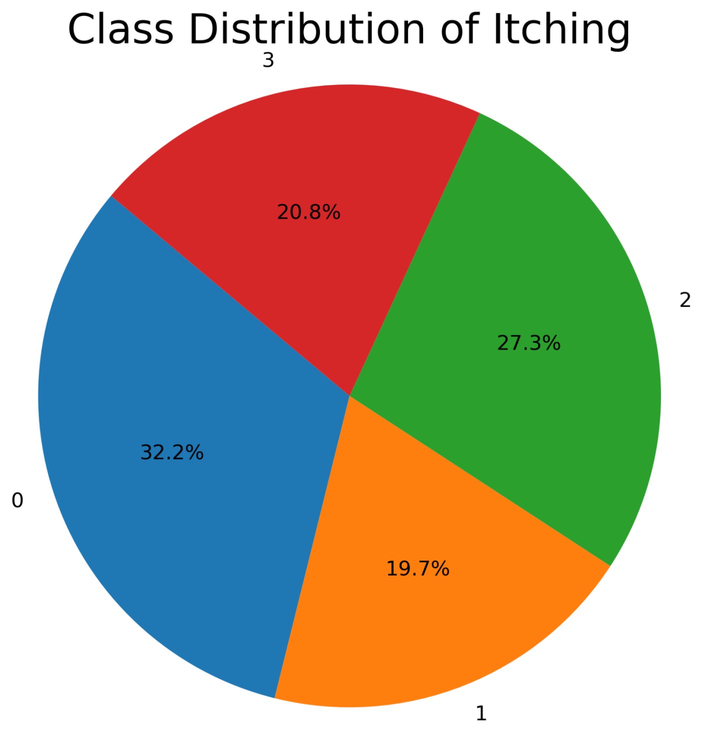


Figure 3: A pie chart of the class distribution of ‘itching’. It is approximately balanced.

To further understand the interplay between different clinical symptoms, a bar plot was constructed in Figure 4 to elucidate this association, with each bar representing a degree of scaling (from 0 to 3). Within these bars, the proportions of erythema classes (ranging from 0 to 3) are distinctly visualized. The plot reveals a noteworthy trend: as the degree of scaling intensifies, there is a corresponding increase in the instances of Class 3 erythema. This pattern suggests a significant relationship where more severe scaling is indicative of more severe erythema.

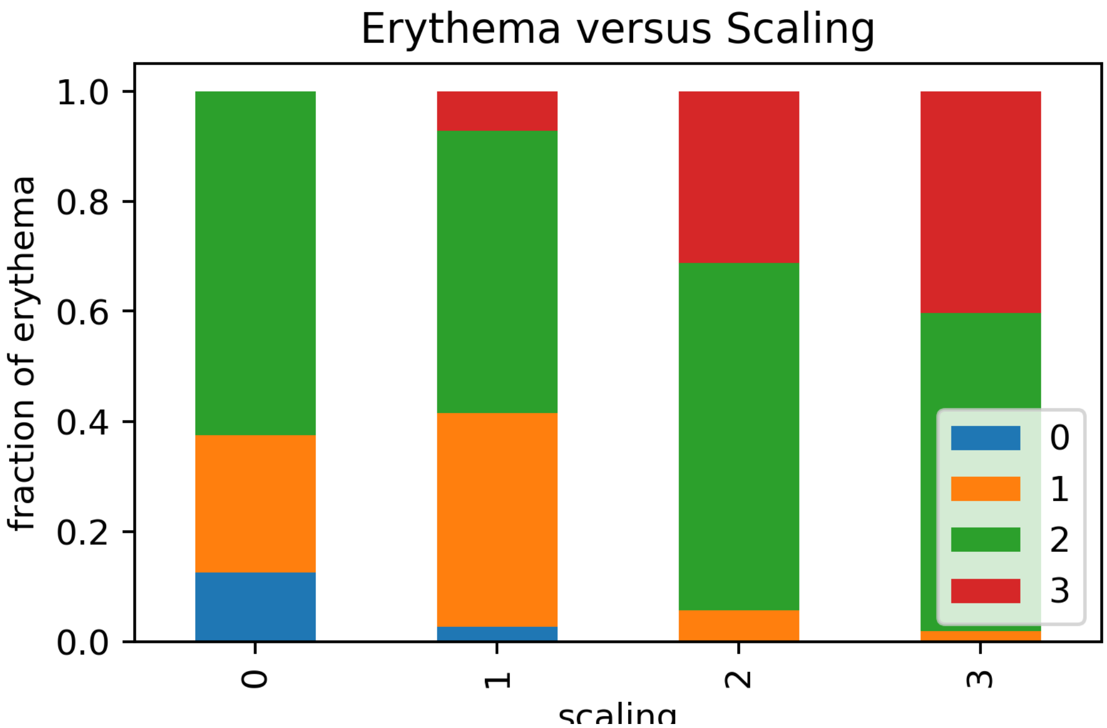


Figure 4: A barplot of Erythema vs scaling. X-axis represents the degrees fo scaling, y-axis represents the fraction of erythema. Each color represents a different class of erythema.

The 'Age' feature is the only attribute with missing values, which constitute a minimal fraction of approximately 0.18%.

3.Methods

3.1 splitting

A stratified k-fold approach was employed when splitting the dataset since it is independent, identically distributed, and imbalanced. This technique partitions the dataset into 'k' folds, ensuring each fold accurately mirrors the overall class distribution. This approach increases the likelihood of a balanced subset representation, enhancing model evaluation accuracy and reliability.

3.2 Preprocessing

Distinct encoding and scaling techniques were adapted to the dataset's varied feature types. A one-hot encoder was applied to the binary 'family_history' variable, while a Standard Scaler normalized the continuous 'age' variable, which had missing values. These missing values in 'age' were addressed using an Iterative Imputer, chosen for its effectiveness over methods like constant imputer. Since the rest of the features are ordinal, an Ordinal Encoder was used. Notably, for models like XGBoost that do not require explicit handling of missing values, the imputation step was bypassed.

3.3 ML Pipeline

A function, MLpipe_KFold_RMSE, was constructed to systematically evaluates and optimizes machine learning models using a combination of K-Fold cross-validation and GridSearchCV. It operates by iterating over a series of random states (defaulting to ten if none are provided) to ensure diverse model training. Within each iteration, it employs a 4-fold cross-validation to split the training data, and a pipeline is created that integrates the preprocessors and the chosen algorithm. If applicable, the algorithm's random state is set for reproducibility. The function then conducts hyperparameter tuning through GridSearchCV, aiming to minimize the negative mean squared error. The best model from each iteration is used to predict the test dataset, and its performance is evaluated using the Root Mean Squared Error (RMSE). The function concludes by aggregating and reporting the mean and standard deviation of these RMSE scores, returning a comprehensive list of the best models, their corresponding test scores, and all predictions on the test set.

RMSE was chosen as the evaluation metric for the models because of its suitability in measuring the average differences between predicted and true values in this regression problem.

3.4 Algorithms and Hyperparameter Tuning

There were six models used in the pipeline. For Lasso and Ridge regression, the alpha parameter was adjusted with values [0.001, 0.01, 0.1, 1, 10, 20] for Lasso and [0.001, 0.01, 0.1, 5, 20, 30] for Ridge. Elastic-Net's parameters alpha and l1_ratio were tuned with [0.005, 0.01, 0.015, 0.02, 10, 100] and [0.3, 0.5, 0.9], respectively. Random Forest involved tuning n_estimators ([100, 300, 400]), max_depth ([None, 20, 40]), and min_samples_split ([10, 30, 60]). The hyperparameters of SVR, C, kernel, degree, and epsilon, were adjusted with ranges [0.001, 0.01, 0.1, 0.15, 10, 100], ['linear', 'rbf', 'poly'], [2, 4, 5], and [0.01, 0.1, 3] respectively. For XGBoost, n_estimators ([20, 30, 50]), learning_rate ([0.1, 0.12, 0.15]), and max_depth ([1, 2, 3])

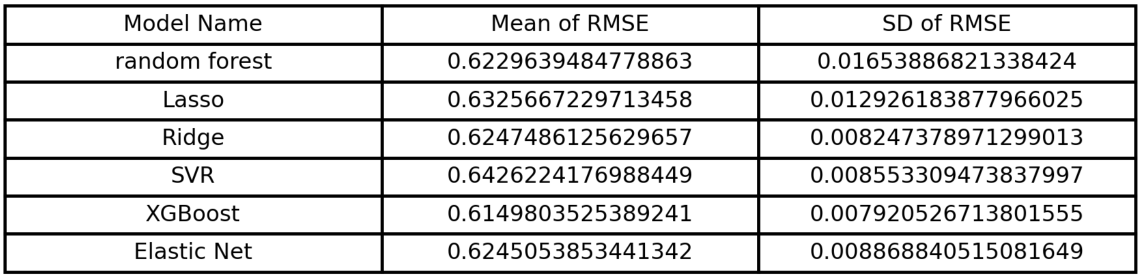
were the focus of tuning. This strategic approach aimed to enhance each model's predictive accuracy.

4. Results

4.1 Test Scores and Overall Performances

The baseline Root Mean Squared Error (RMSE) was approximately 0.67567567, with an associated standard deviation of 0.0251.

The models and their test scors are summarisrd in Table 1.



Model Name	Mean of RMSE	SD of RMSE
random forest	0.6229639484778863	0.01653886821338424
Lasso	0.6325667229713458	0.012926183877966025
Ridge	0.6247486125629657	0.008247378971299013
SVR	0.6426224176988449	0.008553309473837997
XGBoost	0.6149803525389241	0.007920526713801555
Elastic Net	0.6245053853441342	0.008868840515081649

Table 1: Models and their test scores.

The mean RMSE for all models under consideration consistently fell below the established baseline RMSE of 0.67567567.The XGBoost model showed the best predictive performance with the lowest mean RMSE among all evaluated models in this paper.The XGBoost model outperforms the baseline by about 2.43 standard deviations, highlighting its superior efficacy in predicting erythema.

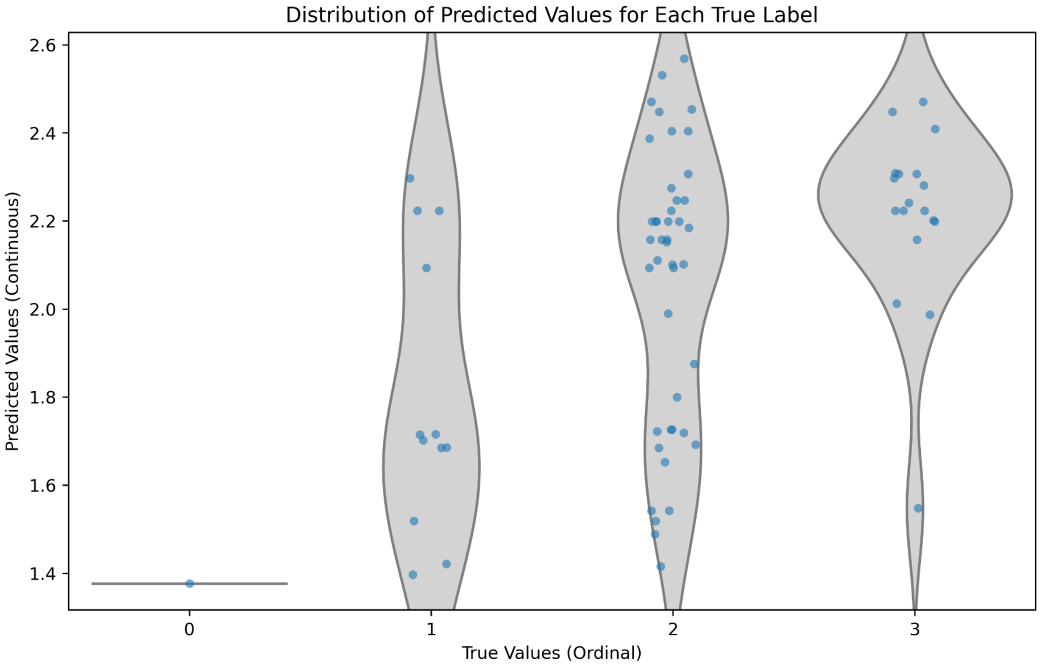


Figure 5: A violin plot of true vs. predicted values. The x-axis represents the four classes of the target. The y-axis represents the predicted values.

In Figure 5, it is evident that there exists substantial room for enhancement in the model's predictive performance. The predicted values predominantly concentrate within class 2, which is coherent since class 2 constitutes the majority class. Due to the relatively small difference between the baseline and mean RMSE of this model, this result is not unexpected. Data imbalance may have contributed to this suboptimal result.

4.2 Global Feature Importance

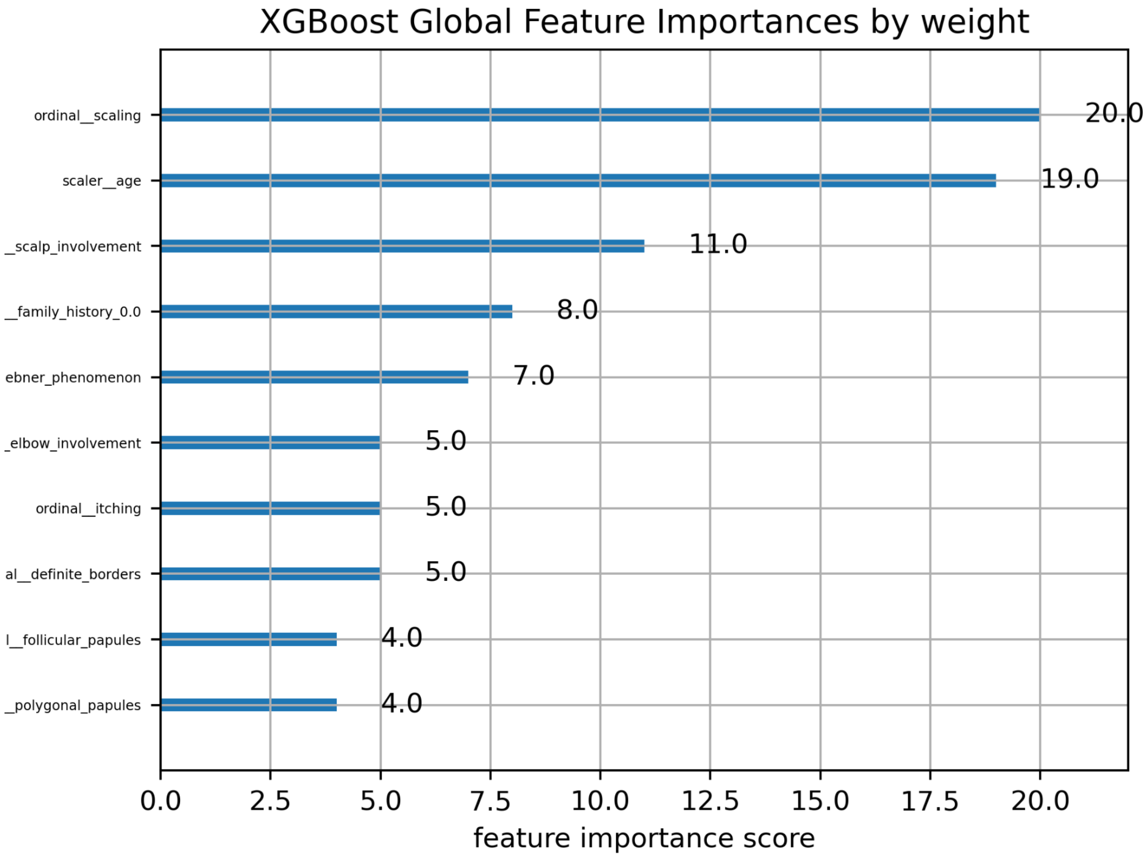


Figure 6: A barplot of feature importances by weight. The x-axis represents feature importance score, the y-axis represents the features.

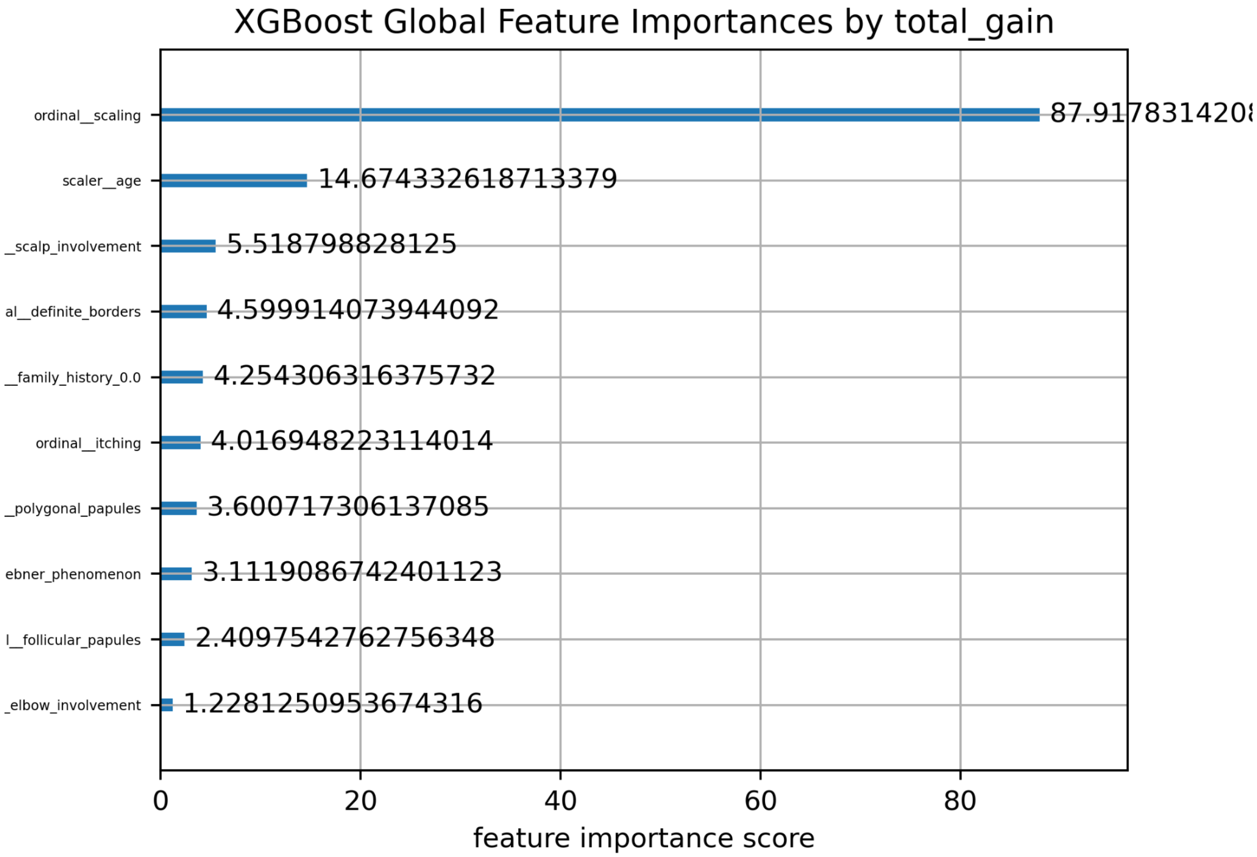


Figure 7: A barplot of feature importances by total_gain. The x-axis represents feature importance score, the y-axis represents the features.

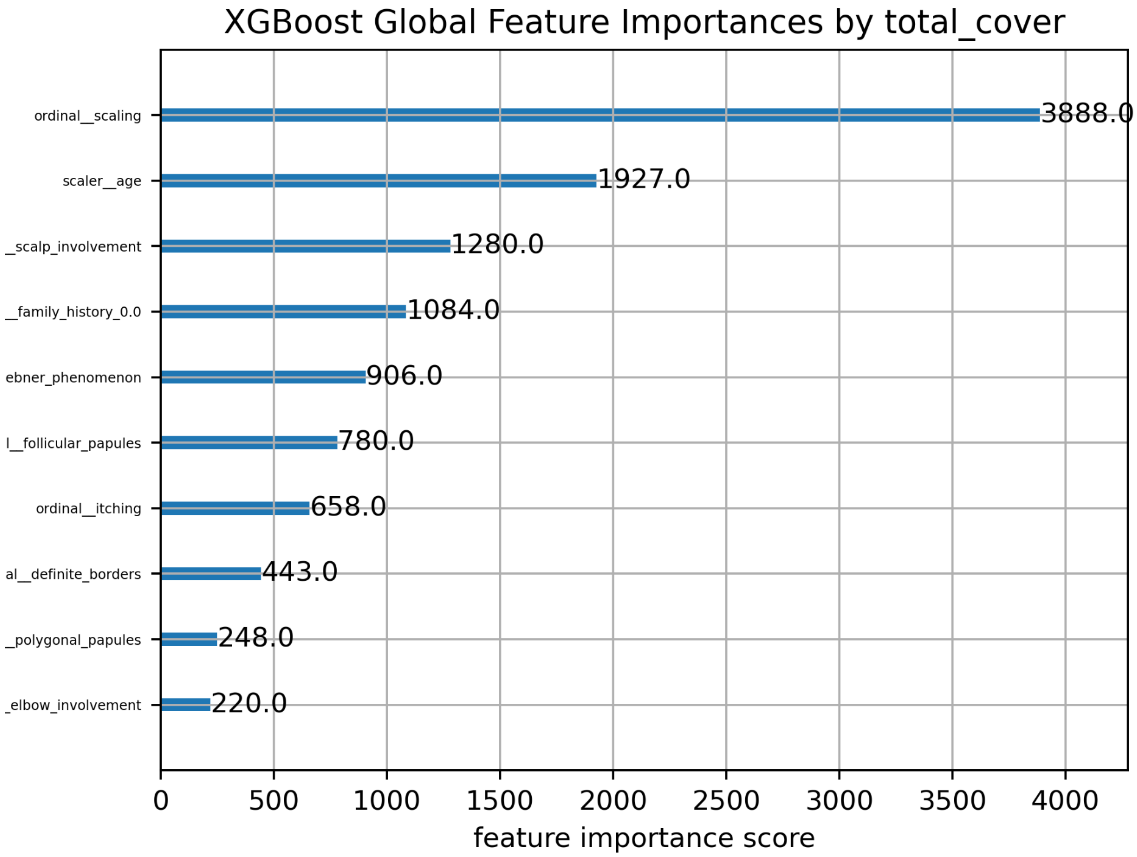


Figure 8: A barplot of feature importances by total_cover. The x-axis represents feature importance score, the y-axis represents the features.

The feature importances have been delineated in Figure 6, 7, and 8, utilizing the weight, total_gain, and total_cover metrics of the XGBoost model, respectively. While the order of features in these figures may vary, the top three features remain consistent: scaling, age, and scalp involvement. It is evident that these three features hold significant importance to predictions.

4.3 Local Feature Importance

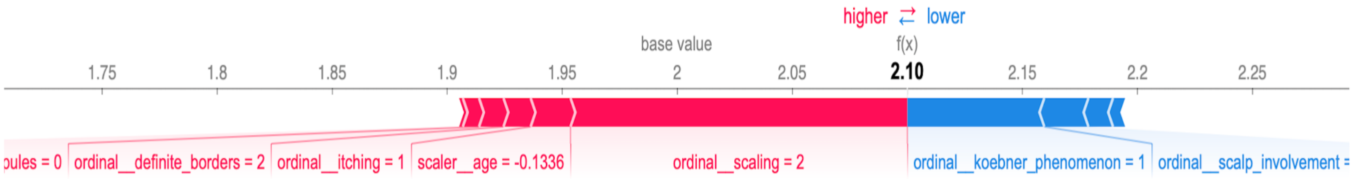


Figure 9: SHAP local value for index 0

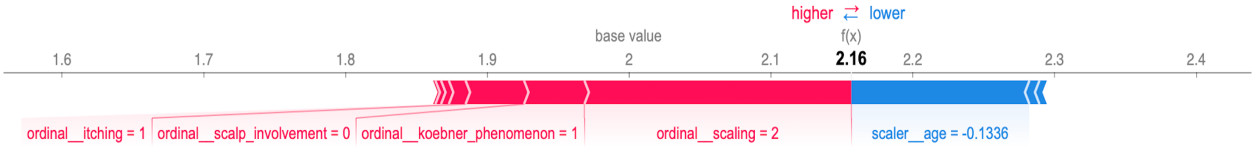


Figure 10: SHAP local value for index 10

In both Figure 9 and 10, it is evident that scaling exhibits a robust and positive contribution to the predictive model. This observation, in conjunction with the global feature importance analysis conducted earlier, underscores the significance of the scaling feature in the context of predictions.

4.4 Interpretaion

The most important feature is scaling. Elbow-involvement ranks lowest in importance according to Figures 7 and 8. Scaling serves as a primary indicator of erythema occurrence. It is intriguing to note that itching has minimal effect on erythema, despite being a predominant symptom associated with it, as shown in Figures 6, 7, and 8.

5. Outlook

A weakness of the modeling approach is the way it handled the extreme imbalanced dataset. Therefore, one method to improve the model is to oversample the minority class and undersample the majority class.[3] Additionally, expanding the dataset by collecting more data from patients with erythema would be beneficial due to its current limited size. It would be helpful to focus on more symptoms associated with erythema. Furthermore, given more time, exploring the histopathological features that were updated by the original author may also improve the model. Lastly, utilizing more global interpretability techniques like Partial

Dependence Plots (PDP) and Accumulated Local Effects (ALE) plots could provide a more comprehensive understanding of the model's behavior across various feature values, collectively enhancing the erythema prediction model's performance and interpretability.[4]

6. Reference

[1]OLCAY BOLAT, Dermatology Dataset (Multi-class classification), Kaggle, 2022, https://www.kaggle.com/datasets/olcaybolat1/dermatology-dataset-classification/data

[2] Ranjan, R., Partl, R., Erhart, R., Kurup, N., & Schnidar, H. (2021). The mathematics of erythema: Development of machine learning models for artificial intelligence assisted measurement and severity scoring of radiation induced dermatitis. *Computers in Biology and Medicine*, *139*, 104952. https://doi.org/10.1016/j.compbiomed.2021.104952

[3]Bach, M., Werner, A., & Palt, M. (2019). The Proposal of Undersampling Method for Learning from Imbalanced Datasets. *Procedia Computer Science*, *159*, 125-134. https://doi.org/10.1016/j.procs.2019.09.167

[4]Danesh, T., Ouaret, R., Floquet, P., & Negny, S. (2022). Interpretability of neural networks predictions using Accumulated Local Effects as a model-agnostic method. *Computer Aided Chemical Engineering*, *51*, 1501-1506. https://doi.org/10.1016/B978-0-323-95879-0.50251-4