Uncover Patterns in Dermatological Data

# CLASSIFICATION OF ERYTHEMATO-SQUAMOUS DISEASES

ADVANCED ANALYSIS

**GROUP 04:**

NAETHREE(15552)
DINULA(15369)
ASLAM(15338)

## Abstract

This report presents a comprehensive analysis of a dermatology dataset, which includes various clinical and histopathological features related to skin conditions. The dataset comprises a range of ordinal variables, reflecting the severity or presence of symptoms, as well as a ‘family_history’ and an 'age' variable. In addition to basic descriptive statistics and visualization of feature relationships, the analysis includes model fitting to identify key predictors. A random forest model was found to be the best-performing model, utilizing the top 16 features. The insights gained from this analysis provide a strong foundation for classifying erythemato-squamous diseases, using the identified patterns and associations among the features.

## Table of Contents

## List of Figures

## List of Tables

## 1. Introduction

The differential diagnosis of erythemato-squamous diseases is recognized as a challenging problem in dermatology due to the shared clinical features of erythema and scaling [2]. The disorders within this group include psoriasis, seborrheic dermatitis, lichen planus, pityriasis rosea, chronic dermatitis, and pityriasis rubra pilaris [2].

The descriptions of each are given below:

- **Psoriasis - Chronic autoimmune condition causing red, scaly patches on the skin.**
- **Seborrheic dermatitis- Skin condition causing red, flaky, itchy skin, often on the scalp and face.**
- **Lichen planus - Inflammatory condition causing purplish, itchy, flat-topped bumps on the skin and mucous membranes.**
- **Pityriasis rosea - Temporary rash of raised, scaly patches, often starting with a herald patch.**
- **Chronic dermatitis - Persistent eczema causing red, itchy, inflamed skin, often due to genetic and environmental factors.**
- **Pityriasis rubra pilaris - Rare disorder causing reddish-orange scaly patches and thickened skin on the palms and soles.**

Accurate diagnosis is crucial for effective treatment but is often dependent on biopsy, which can be invasive and inconvenient for patients. Additionally, the histopathological similarities among these conditions complicate the diagnostic process. Promising solutions for improving diagnostic accuracy and efficiency are offered by advancements in machine learning and artificial intelligence. This report lays the foundation for the development of a robust model by exploring key relationships through EDA.

## 2. Description of the Question

Using advanced analytical techniques to enhance classification accuracy is essential due to the limitations of traditional diagnostic methods. Therefore, the objective of this report is to perform exploratory data analysis (EDA) to address the following questions:

**1) What are the factors mainly affecting each of the six diseases?**

**2) How can we accurately classify erythemato-squamous diseases based on clinical and histopathological features?**

In pursuit of these objectives, various statistical methods will be employed to uncover the underlying patterns and relationships within the dataset. By analyzing both clinical and histopathological features, this report aims to identify key factors that differentiate each of the six diseases.

## 3. Description of the Dataset

The dataset contains 35 variables and 366 records.

Few of the variables are given below:

| Name of variable | Description of variable | Type of variable |
|---|---|---|
| erythema | Redness of the skin caused by increased blood flow | Ordinal (0,1,2,3) |

| | | |
|---|---|---|
| scaling | Flaking or shedding of the outer layer of the skin | Ordinal (0,1,2,3) |
| definite_borders | Clear and distinct edges of a lesion or affected area | Ordinal (0,1,2,3) |
| itching | Sensation causing the desire to scratch | Ordinal (0,1,2,3) |
| koebner_phenomenon | Appearance of new skin lesions on previously unaffected skin due to trauma | Ordinal (0,1,2,3) |
| polygonal_papules | Small, raised, polygonal-shaped bumps on the skin | Ordinal (0,1,2,3) |
| follicular_papules | Small, raised bumps centered around a hair follicle. | Ordinal (0,1,2,3) |
| age | Age of the patient | Continuous |
| family_history | Indicates if any of the diseases are present in the family (1 if yes, 0 if no). | Nominal(0,1) |
| class | The target variable indicates the disease class.<br>1-Psoriasis, 2-Seborrheic Dermatitis, 3-Lichen Planus, 4-Pityriasis Rosea, 5-Chronic Dermatitis, 6-Pityriasis Rubra Pilaris | Nominal |

**Table 1: Description of part of the dataset**

The rest of the variables are all ordinal each with 0-3 symptom levels and they are oral_mucosal_involvement, knee_and_elbow_involvement, scalp_involvement, melanin_incontinence, eosinophils_infiltrate, PNL_infiltrate, fibrosis_papillary_dermis, exocytosis, acanthosis, hyperkeratosis, parakeratosis, clubbing_rete_ridges, elongation_rete_ridges, thinning_suprapapillary_epidermis, spongiform_pustule, munro_microabcess, focal_hypergranulosis, disappearance_granular_layer, vacuolisation_damage_basal_layer, spongiosis, saw_tooth_appearance_retes, follicular_horn_plug, perifollicular_parakeratosis, inflammatory_mononuclear_infiltrate, band_like_infiltrate.

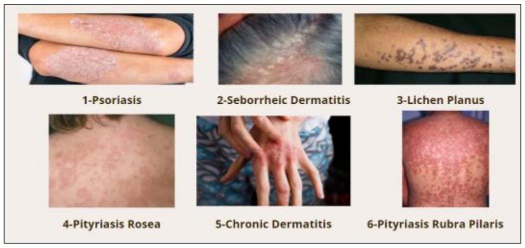

**Figure 1: Response variable Diseases**

## 4. Pre-Processing

- The datatype of 'age' was changed from object to int64.
- Datatype of all other features were changed from int64 to category for plotting purposes.
- 8 missing values(denoted as '?') were found in the 'age' variable. Mean Imputation was done since by figure 2, it is symmetrically distributed and by skewness score it was clear that impact from outliers are minimal.
- No duplicates were found.
- When Koebner phenomenon symptom by class was analyzed, an unusual observation was made when class 2(seborrheic_dermatitis) was considered. While 60 records had no Koebner phenomenon as the symptom, one observation had symptom level of severity 2 which is inaccurate according to out references which state "***Koebner phenomenon affects people with certain skin diseases, most often with psoriasis. Sometimes, it can happen to people with warts, vitiligo and lichen planus. An injury, wound or burn can cause new lesions that resemble the primary skin disease.***" **[1]**

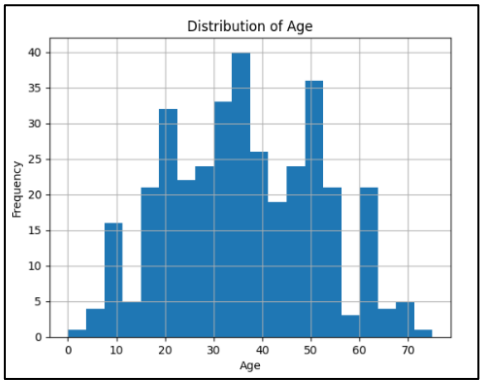

**Figure 2: Distribution of Age before imputation**

## 5. Main results of the Descriptive Analysis

### 5.1 Univariate Analysis

### 5.1.1 Response Variable

Figure 3 denotes percentage bar chart of the response and it is evident that the highest percentage of data is available for Psoriasis while Pityriasis Rubra Pilaris is indicated by the least amount of data.

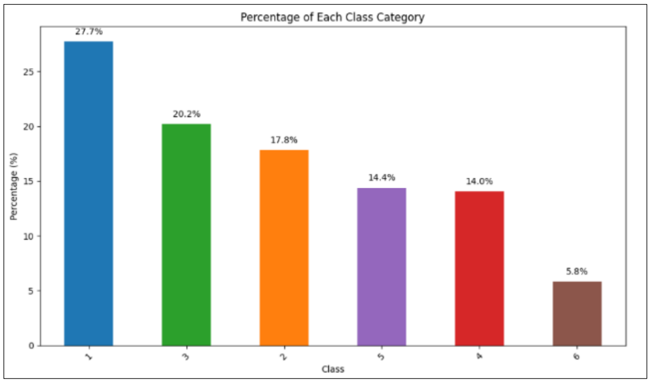

**Figure 3: Percentage Bar Chart of Class**

### 5.1.2 Age variable

Age's distribution can be seen by Figure 4.The mean age of patients is 35.7 while less patients are aged greater than 60, hence this should be kept in mind when reflecting the results of the dataset as it might not represent patients with extreme ages well.

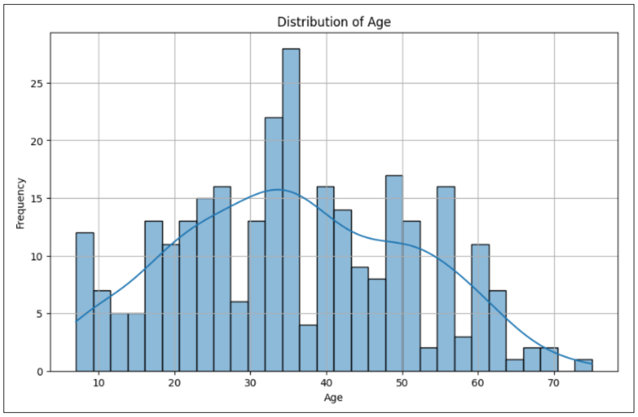

**Figure 4: Distribution of age after imputation**

### 5.1.2 Nominal variable family history

By Figure 5, it is evident that most patients are from families who don't have any of these diseases. When building and evaluating models, it's essential to keep this in mind since the results won't represent the patients from families with the diseases well.

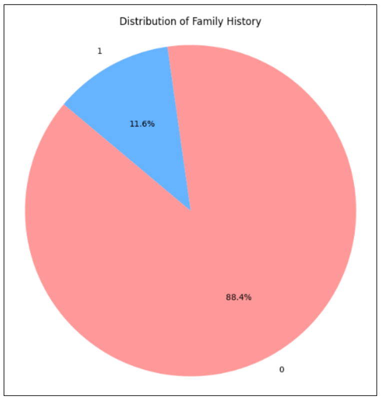

**Figure 5: Distribution of family history**

### 5.2 Bivariate Analysis

### 5.2.1 Main observations of variables by Class

According to Figure 6, most people affected by Pityriasis Rosea(class 4) seem to have no itching. This is confirmed by "***Except for mild to severe itching in up to 25% of patients, no systemic symptoms are typically present during the rash phase of pityriasis rosea.*"[1]** Majority of people affected by Lichen Planus have severe itching which is confirmed by "***Lichen planus is a non-infectious, itchy rash that can affect many areas of the body.*"[2]**

According to Figure 7, people affected by Psoriasis, Seborrheic Dermatitis, Pityriasis Rosea, Chronic Dermatitis and Pityriasis Rubra Pilaris aren't affected by the oral mucosa, but Lichen Planus commonly has oral mucosal involvement where patients face them at different severity levels according to figure 7. This is confirmed by ***"Oral lichen planus is often diagnosed by a dentist during routine dental check-ups."*****[2]** which indicates that oral lichen planus is a type of Lichen Planus, which affects the mouth.

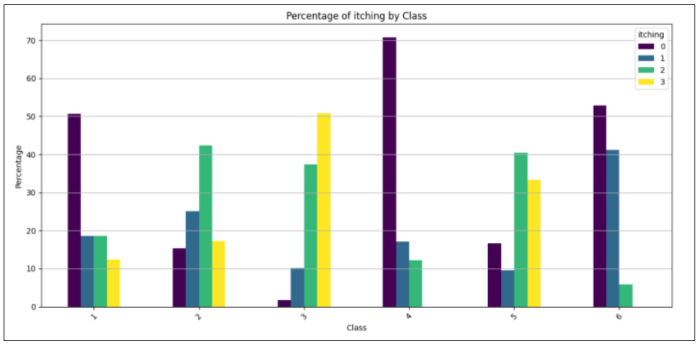

**Figure 6: Percentage bar chart of itching by class**

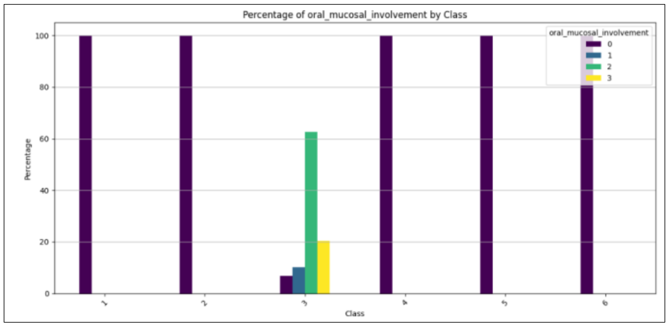

**Figure 7: Percentage bar chart of oral mucosal involvement by class**

*Based on research[3]-[11] regarding the six diseases,the below Table 2 shows the following:*
*Green: the symptoms reflected in the dataset and the references*
*Turquoise: the symptoms represented by the dataset and not the references*
*No color: the symptoms in the references but not in the dataset.*

| Psoriasis | Seborrheic Dermatitis | Lichen Planus | Pityriasis Rosea | Chronic Dermatitis | Pityriasis Rubra Pilaris |
|---|---|---|---|---|---|
| Erythema | Erythema | Erythema | Erythema | Erythema | Erythema |
| Scaling | Scaling | Definite Borders | Scaling | Scaling | Scaling |
| Definite Borders | Itching | Itching | Definite Borders | Itching | Definite Borders |
| Itching | Scalp Involvement | Polygonal Papules | Itching | Spongiosis | Follicular Papules |
| Koebner Phenomenon | Exocytosis | Oral Mucosal Involvement | Exocytosis | Scalp Involvement | Knee and Elbow Involvement |
| Knee and Elbow Involvement | Acanthosis | Melanin Incontinence | Spongiosis | Fibrosis Papillary Dermis | Scalp Involvement |
| Scalp Involvement | Parakeratosis | Exocytosis | Inflammatory Mononuclear Infiltrate | Inflammatory Mononuclear Infiltrate | Hyperkeratosis |
| PNL Infiltrate | Spongiosis | Hyperkeratosis | Koebner Phenomenon | Definite Borders | Parakeratosis |
| Acanthosis | Inflammatory Mononuclear Infiltrate | Focal Hypergranulosis | Acanthosis | Exocytosis | Inflammatory Mononuclear Infiltrate |
| Parakeratosis | PNL Infiltrate | Vacuolisation Damage Basal Layer | Parakeratosis | Acanthosis | Exocytosis |
| Clubbing Rete Ridges | Definite Borders | Saw Tooth Appearance Retes | | Elongation Rete Ridges | Acanthosis |
| Elongation Rete Ridges | | Inflammatory Mononuclear Infiltrate | | | Spongiosis |
| Thinning Suprapapillary Epidermis | | Band-like Infiltrate | | | Follicular Horn Plug |

| | | | | | |
|---|---|---|---|---|---|
| Spongiform Pustule | | Scaling | | | Perrifollicular Parakeratosis |
| Munro Microabcess | | Koebner Phenomenon | | | |
| Disappearance Granular Layer | | Acanthosis | | | |
| Inflammatory Mononuclear Infiltrate | | Parakeratosis | | | |
| | | Spongiosis | | | |

**Table 2: Comparison of symptoms for each class of diseases by research VS analysis**

### 5.2.2 Correlation between the numeric variables

By Figure 8, the correlation coefficient values between the numeric variables can be observed.
Key observations are as follows:

- Each pair in Focal hypergranulosis, Polygonal Papules, Oral Mucosal Involvement, Melanin Incontinence, Vascuolisation Damage Basal Layer, Saw Tooth Appearance Retes and Band Like Infiltrate is highly correlated with each other. These features maybe highly correlated because they are commonly found together in certain dermatological conditions, particularly lichen planus.
- Each pair in Follicular Papules, Follicular Horn Plug and Perifollicular Parakeratosis are highly correlated with each other.

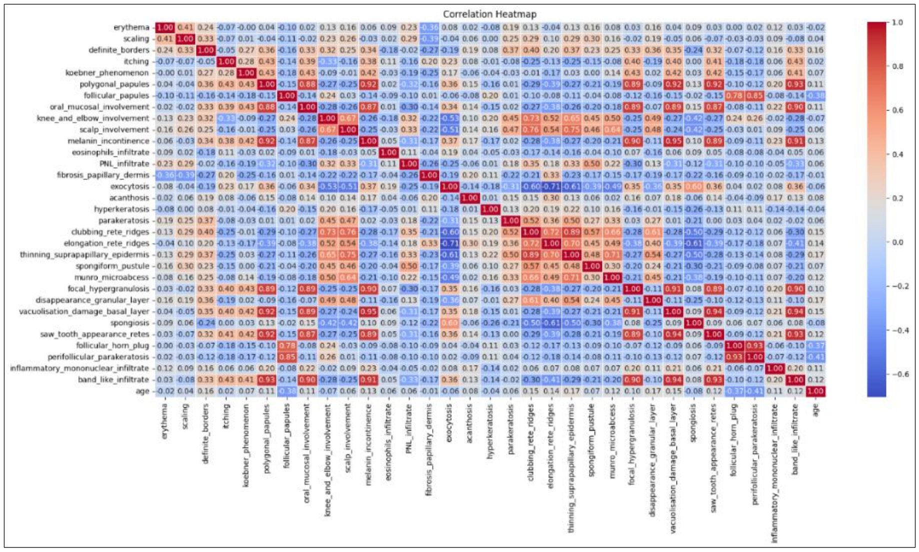

**Figure 8: Correlation Heatmap of numerical variables**

### 5.3 Multivariate Analysis

#### 5.3.1 Multicollinearity

The Variance Inflation Factor analysis was conducted to assess the multicollinearity among the features in the dermatology dataset. The results in Table 3 indicate significant multicollinearity, particularly among certain features. For instance, features like vacuolisation_damage_basal_layer (VIF = 27.79), band_like_infiltrate (VIF = 17.76), and melanin_incontinence (VIF = 16.30) exhibit very high VIF scores, suggesting a high degree of correlation with other predictors in the model. Other features such as polygonal_papules, erythema, and scaling also show substantial multicollinearity with VIF scores above 10.

| Variable Name | VIF Score | Variable Name | VIF Score |
|---|---|---|---|
| erythma | 13.840555 | hyperkeratosis | 1.921288 |
| scaling | 12.067053 | parakeratosis | 5.140062 |
| definite_borders | 8.825341 | clubbing_rete_ridges | 12.034817 |
| itching | 4.30289 | elongation_rete_ridges | 11.559817 |
| koebner_phenomenon | 2.398118 | thinning_suprapapillary_epidermis | 10.070708 |
| polygonal_papules | 13.57255 | spongiform_pustule | 2.419026 |
| follicular_papules | 4.412315 | munro_microabcess | 3.0005 |
| oral_mucosal_involvement | 9.001588 | focal_hypergranulosis | 10.176414 |
| knee_and_elbow_involvement | 4.903507 | disappearance_granular_layer | 2.468943 |
| scalp_involvement | 4.53806 | vacuolisation_damage_basal_layer | 27.794759 |
| family_history | 1.571547 | spongiosis | 4.604313 |
| melanin_incontinence | 16.303299 | saw_tooth_appearance_retes | 15.529731 |
| eosinophils_infiltrate | 1.353087 | follicular_horn_plug | 8.397643 |
| PNL_infiltrate | 2.724906 | perifollicular_parakeratosis | 12.400189 |
| fibrosis_papillary_dermis | 3.778801 | inflammatory_mononuclear_infiltrate | 9.527857 |
| exocytosis | 6.957575 | band_like_infiltrate | 17.755538 |
| acanthosis | 10.79972 | age | 7.202199 |

**Table 3: Table of VIF values**

### 5.3.2 Principal Component Analysis

By Figure 9, it is evident that to get at least 80% variance explained by the PCs, you require 11 components. But since 11 PCs cannot be visualized on a 2D plane, just for visualization Purposes 2 components can be used. However, note that only less than 45% of variance is explained by 2 PCs. This relatively low percentage of explained variance suggests that a significant portion of the dataset's information is not captured by these two components. Hence it was not considered.

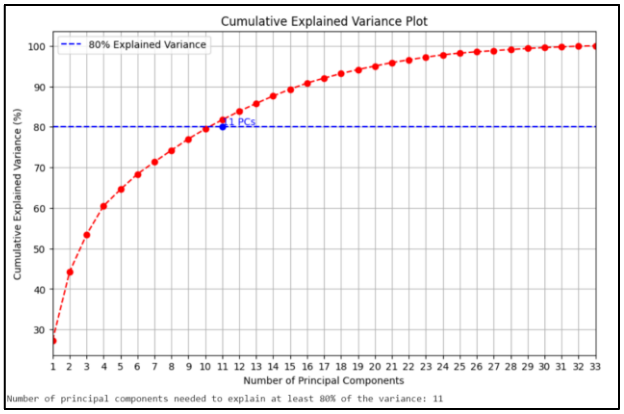

**Figure 9: Cumulative Explained Variance Plot**

### 5.3.3 Partial Least Squares Discriminant Analysis

Since PLS-DA aims to maximize the separation between classes, it was applied to the data. The components found by PLS-DA are optimized for discriminating between classes, which can lead to better separation and higher explained variance in terms of class discrimination which can be seen by the score plot.
The score plot was generated to visualize the projection of samples onto the first two PLS components. It was observed that clusters corresponding to classes were effectively separated in the plot except for classes 2 and 4 which have a little overlap. This visualization provided insight into how well the model distinguished between the other classes, with the score plot reflecting the performance of the PLS-DA in class separation.

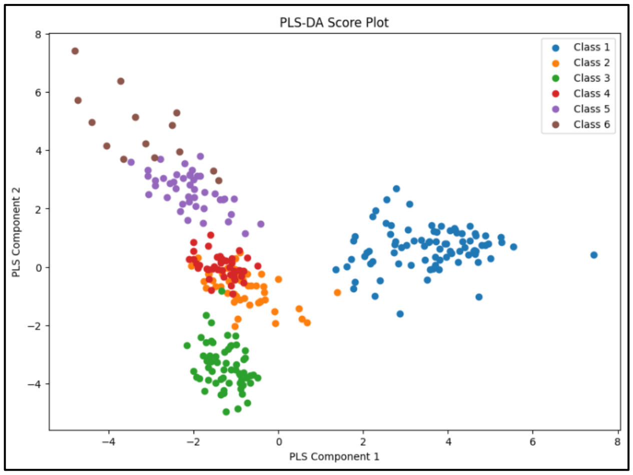

**Figure 10: PLS-DA Score plot by class**

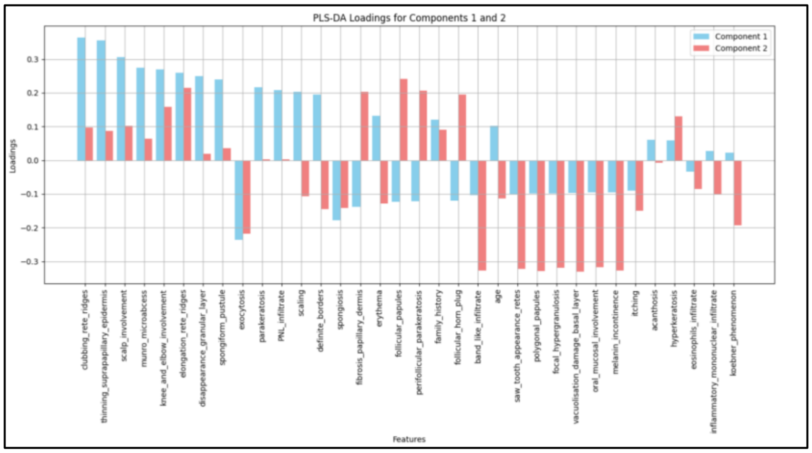

**Figure 11: PLS-DA Loadings for first two components**

Figure 11 visualizes the loadings of the first two components in the PLS-DA model. Each bar represents a feature and its position on the x-axis indicates its contribution to the respective component. The height of the bar represents the loading value, signifying the importance of the feature in explaining the variation captured by the component. For example, clubbing_rete_ridges has a high positive loading on Component 1. If Component 1 is a strong differentiator between two classes (diseases), this symptom might be significantly elevated in one disease compared to the other.

When you see all the highly positive loadings on Component 1, it is clear how symptoms clubbing_rete_ridges, thinning_suprapapillary_epidermis, scalp_involvement, munro_microabcess and others(with blue positive bars) are present in Psoriasis(class 1) from Table 2. This indicates that these symptoms allowed a good separation causing Psoriasis to be separated well(as high positive loadings pushed the observations to the right of the score plot in Figure 10).

Similarly, when you consider the negative loadings on Component 2, the symptoms band_like_infiltrate, saw_tooth_appearance_retes, polygonal_papules, focal_hypergranulous and others(with pink negative bars) are not present at all in Pityriasis Rubra Pilaris(class 6). That means that these symptoms do not affect class 6 and hence generally expect low values. Due to negative loading, this then would have positive scores. This indicates that these symptoms allowed a good separation causing PRP to be separated well(as high negative loadings pushed the observations to the top of the score plot in Figure 10).

## 5.4 Outlier Detection

In Figure 12, Isolation Forest method was applied for outlier detection after performing PLS-DA. The obtained PLS scores, were used as input for the Isolation Forest algorithm. Note that most of the observations which belong to Class 6 (as seen in Figure 10) has appeared as outliers. This may be due to how there are just 14 observations belonging to class 6 and how the uniqueness of that disease has caused these data to appear as outliers. Hence no action is taken now, as when modelling, if accuracy of models is weak, then they can be built again by removing the outliers.

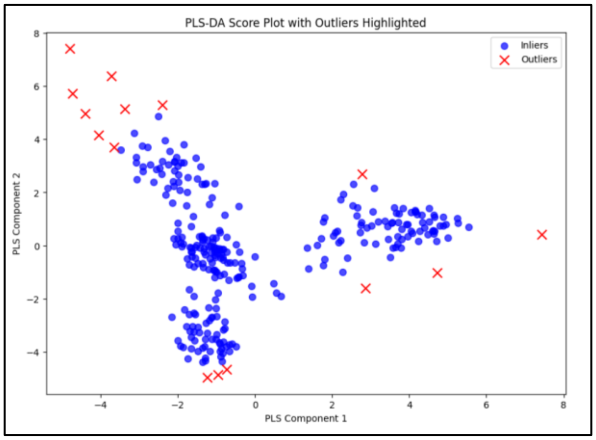

**Figure 12: PLS-DA Score plot with outliers**

## 6. Important Results of the Advanced analysis

Since the dataset is imbalanced, SMOTE-NC (Synthetic Minority Over-sampling Technique for Nominal and Continuous) which is an extension of SMOTE for datasets with both numerical and categorical features.

It generated synthetic samples to balance class distributions. Before building a model, basic descriptive analysis was done on the sampled dataset and there were no anomalies, and generally same patterns were found. Hence the data was ready for advanced analysis.Since there was some overlap between some diseases and symptoms which was observed in the descriptive analysis, it was assumed that non-linear models can fit better to the data. However, both linear and non-linear models were fit. Models that can handle multi-collinearity were given priority considering the evidence given by Table 3. Models were also built before and after removing outliers, but models without removing proved to give better results.

The F1 score, Recall, and Precision were calculated and averaged using the "Macro" method to provide a balanced evaluation across all classes. However, it is important to note that in medical contexts, the F1 score is particularly valuable. This is because it provides a balanced measure of a model's performance by combining both precision and recall into a single metric, which is crucial for evaluating models in scenarios where the costs of false positives and false negatives are significant. Hence F1 score will be given priority in this case.

The results are as follows:

| | **Training Set** | | | | **Test Set** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Precision** | **Recall** | **F1** | **Accuracy** | **Precision** | **Recall** | **F1** | **Accuracy** |
| **Logistic Ridge** | 0.99 | 0.99 | 0.99 | 0.9882 | 0.99 | 0.98 | 0.98 | 0.9863 |
| **Logistic Lasso** | 0.99 | 0.99 | 0.99 | 0.9863 | 0.99 | 0.98 | 0.98 | 0.9863 |
| **KNN** | 0.99 | 0.99 | 0.99 | 0.9863 | 0.96 | 0.97 | 0.96 | 0.9726 |
| **Multinomial Naive Bayes** | 0.98 | 0.98 | 0.98 | 0.9824 | 0.96 | 0.96 | 0.96 | 0.9726 |
| **SVM** | 0.99 | 0.99 | 0.99 | 0.9882 | 0.99 | 0.98 | 0.98 | 0.9863 |
| **Gradient Boost** | 1.00 | 1.00 | 1.00 | 1.00 | 0.98 | 0.99 | 0.98 | 0.9863 |
| **Random Forest** | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.98 | 0.98 | 0.9863 |

**Table 4: Table of results of models after applying SMOTENC**

Since overall, all models performed well, a voting classifier, which combines the predictions of above models to improve generalization, was applied. The voting classifier showed a slight improvement in performance, benefiting from the diverse decision-making of all models. However, when considering feature importance and reducing variables, it was observed that using the Random Forest classifier yielded better results. Since top 16 features from the Random Forest Model were similar for most of the above models, those were chosen as the final set of features for the model and can be viewed in Figure 13.

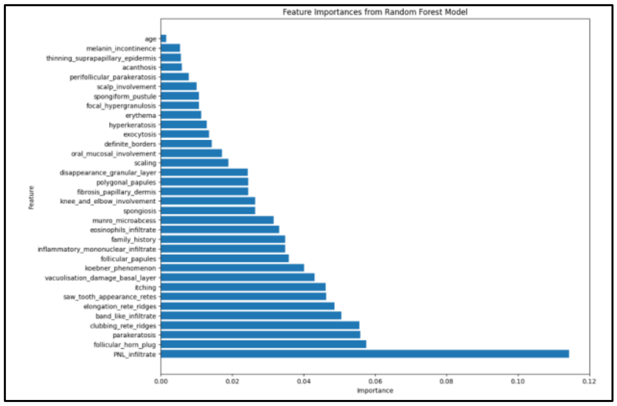

**Figure 13: Feature importance values of Random Forest Model**

The table below shows the results of the fitted models after choosing the top 16 features of Random Forest.

| Training Set | | | | | Test Set | | | |
|---|---|---|---|---|---|---|---|---|
| | **Precision** | **Recall** | **F1** | **Accuracy** | **Precision** | **Recall** | **F1** | **Accuracy** |
| **Logistic Ridge** | 0.96 | 0.96 | 0.96 | 0.9588 | 0.86 | 0.89 | 0.87 | 0.8767 |
| **Logistic Lasso** | 0.97 | 0.97 | 0.97 | 0.9686 | 0.92 | 0.94 | 0.93 | 0.9315 |
| **KNN** | 0.98 | 0.98 | 0.98 | 0.9824 | 0.88 | 0.87 | 0.86 | 0.8767 |
| **Multinomial Naive Bayes** | 0.97 | 0.97 | 0.97 | 0.9725 | 0.92 | 0.94 | 0.93 | 0.9315 |
| **SVM** | 0.98 | 0.98 | 0.98 | 0.9843 | 0.95 | 0.95 | 0.94 | 0.9315 |
| **Gradient Boost** | 1.00 | 1.00 | 1.00 | 0.9980 | 0.93 | 0.95 | 0.94 | 0.9315 |
| **Random Forest** | 1.00 | 1.00 | 1.00 | 0.9980 | 0.94 | 0.96 | 0.95 | 0.9452 |
| **Voting Classifier** | 0.94 | 0.94 | 0.94 | 0.9431 | 0.86 | 0.88 | 0.87 | 0.8767 |

**Table 5: Table of results of models with top 16 features**

Random Forest model emerged as the best model with the highest F1 score and highest accuracy on unseen data.

## 7. The Best Model

The best model is the Random Forest Model with its top 16 features:
PNL_infiltrate, follicular_horn_plug, parakeratosis, clubbing_rete_ridges, band_like_infiltrate, elongation_rete_ridges, saw_tooth_appearance_retes, itching, vacuolisation_damage_basal_layer, koebner_phenomenon, follicular_papules, inflammatory_mononuclear_infiltrate, family_history, eosinophils_infiltrate, munro_microabcess, spongiosis.

Hyper-parameter tuning was performed and the best parameters were as follows:
**{'max_depth': None, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}**

| Training Set | | | | | Test Set | | | |
|---|---|---|---|---|---|---|---|---|
| | **Precision** | **Recall** | **F1** | **Accuracy** | **Precision** | **Recall** | **F1** | **Accuracy** |
| **Random Forest** | 1.00 | 1.00 | 1.00 | 0.9980 | 0.94 | 0.96 | 0.95 | 0.9452 |

**Table 6: Table of final tuned best model**

Now for the best model, the partial dependency plots were analyzed for each class.

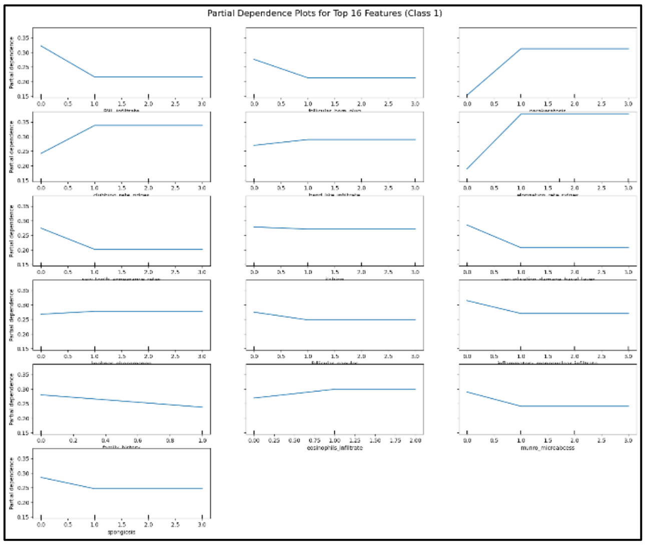

Figure 14: PDP plot of class 1

By Figure 14, it was evident that when most features increase in value, the average predicted probability of the specific class decreases which means that these may not affect class 1(Psoriasis). However, Parakeratosis, Clubbing Rate Ridges and Elongation Rete Ridges seem to affect this disease as when its severity increases, the average predicted probability of this disease increases, indicating that these features may be the main symptoms of this disease(leading to a good accuracy of the model). These symptoms are indeed accurate by the research done in Table 2.

Similarly, the other plots for each class were also analyzed and in summary, the following are considered as the main indicators that contributed to the detection of the 6 diseases:

**Class 1 (Psoriasis) - Parakeratosis, Clubbing Rate Ridges, Elongation Rete Ridges**
**Class 2 (Seborrheic dermatitis) - Vacuolisation_damage_basal_layer, Eosinophils_infiltrate**
**Class 3 (Lichen planus) – family_history, Spongiosis, Munro_microabcess**
**Class 4 (Pityriasis rosea) - Koebner Phenomenon , Vacuolisation_damage_basal_layer**
**Class 5 (Chronic dermatitis) - PNL Infiltrate**
**Class 6 (Pityriasis rubra pilaris) - Saw Tooth Appearance Retes, Follicular Horn Plug**

## 7. Issues Encountered and Proposed Solutions

1) Principal Component Analysis (PCA) with 2 components explained only a small percentage of the total variance in the dataset. This limited the effectiveness of PCA for dimensionality reduction and subsequent analysis, as it did not capture enough of the variability in the data. Given the limitations of PCA, Partial Least Squares Discriminant Analysis (PLS-DA) was employed which enabled better dimensionality reduction by using both the features and the response variable to find latent structures in the data.

2) The dataset had 34 features in total, hence having all of them in the final model and the data product can prove to be complex. Descriptive analysis gave an overview of some variables that were important but removing them didn't seem like a smart choice. Hence models were built keeping all the variables and then feature importance values were calculated to choose the top features.

3) Multicollinearity was dealt with by predominantly fitting models that handled it.

4) Very small number of outliers were present and when modelling, those were removed but it proved to reduce the F1 score, hence they were kept in the data.

## 7. Discussion and Conclusion

1) The dataset contained 32 ordinal variables, which were initially treated as categorical. In this analysis, ordinal variables were treated as numerical features, assuming that their order could be captured in numerical form. This allowed for more straightforward application of techniques such as Principal Component Analysis (PCA) and Partial Least Squares Discriminant Analysis (PLS-DA).

2) Fisher's Linear Discriminant Analysis could not be applied effectively due to the violation of its underlying assumptions. Specifically, it assumes that the features follow a normal distribution and that each class has the same covariance matrix, which were not satisfied in the dataset. Given the assumptions were not met, PLS-DA was analyzed in depth to understand that maybe non-linear models fitted better. However, both linear and non-linear models were fit and best model which was Random Forest was chosen.

3) The Random Forest model, with its inherent capability to evaluate feature importance and handle a wide range of feature interactions, proved more effective in capturing the essential patterns in the data after feature reduction and this approach ensured a more streamlined model that maintained high performance while focusing on the most relevant features. Hence, Random Forest with the top 16 features proved to be the best model.

## 7. References

[1] Cleveland Clinic, "Koebner Phenomenon," Cleveland Clinic, Apr. 27, 2022. [Online]. Available: https://my.clevelandclinic.org/health/diseases/22860-koebner-phenomenon. [Accessed: Aug. 19, 2024].

[2] DermNet. "Pityriasis Rosea," DermNet NZ. [Online]. Available: https://dermnetnz.org/topics/pityriasis-rosea. [Accessed: 20-Jul-2024].

[3] NHS Inform. "Lichen Planus," NHS Inform. [Online]. Available: https://www.nhsinform.scot/illnesses-and-conditions/skin-hair-and-nails/lichen-planus/. [Accessed: 22-Jul-2024].

[4] Mayo Clinic. "Lichen Planus - Symptoms and Causes," [Online]. Available: https://www.mayoclinic.org/diseases-conditions/lichen-planus/symptoms-causes/syc-20351378. [Accessed: 23-Jul-2024].

[5] National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS). "Psoriasis - Symptoms and Causes," [Online]. Available: https://www.niams.nih.gov/health-topics/psoriasis#:~:text=Symptoms%20of%20psoriasis%20vary%20from,Thick%2C%20ridged%2C%20pitted%20nails. [Accessed: 22-Jul-2024].

[6] NHS. "Psoriasis - Symptoms," [Online]. Available: https://www.nhs.uk/conditions/psoriasis/symptoms/. [Accessed: 23-Jul-2024].

[7] National Eczema Association. "Types of Eczema: Seborrheic Dermatitis," [Online]. Available: https://nationaleczema.org/eczema/types-of-eczema/seborrheic-dermatitis/. [Accessed: 23-Jul-2024].

[8] Mayo Clinic. "Seborrheic Dermatitis - Symptoms and Causes," [Online]. Available: https://www.mayoclinic.org/diseases-conditions/seborrheic-dermatitis/symptoms-causes/syc-20352710. [Accessed: 24-Jul-2024].

[9] NHS. "Pityriasis Rosea - Symptoms," [Online]. Available: https://www.nhs.uk/conditions/pityriasis-rosea/#:~:text=A%20widespread%20rash%20of%20small,body%20and%20may%20be%20itchy. [Accessed: 25-Jul-2024].

[10] Mayo Clinic. "Dermatitis - Symptoms and Causes," [Online]. Available: https://www.mayoclinic.org/diseases-conditions/dermatitis-eczema/symptoms-causes/syc-20352380#:~:text=Dermatitis%20is%20a%20common%20condition,%2C%20ooze%2C%20crust%20or%20flake. [Accessed: 23-Jul-2024].

[11] Mount Sinai. "Pityriasis Rubra Pilaris - Overview," [Online]. Available: https://www.mountsinai.org/health-library/diseases-conditions/pityriasis-rubra-pilaris#:~:text=Pityriasis%20rubra%20pilaris%20is%20an,a%20characteristic%20orange%2Dred%20color. [Accessed: 24-Jul-2024].

[12] National Organization for Rare Disorders (NORD). "Pityriasis Rubra Pilaris," [Online]. Available: https://rarediseases.org/rare-diseases/pityriasis-rubra-pilaris/. [Accessed: 22-Jul-2024].

## 8. Appendix

Code: https://github.com/Naethree/Dermatology_ML