Project Report: Skin Disorder Prediction

1. Introduction

This project aims to develop a machine learning solution for the accurate and early classification of various skin disorders. The inherent challenge in diagnosing skin conditions often stems from the similarity in clinical and histopathological features across different diseases. By leveraging computational methods, this project seeks to provide a supportive tool for clinicians, thereby improving diagnostic speed, accuracy, and ultimately, patient outcomes, without exclusive reliance on invasive biopsy procedures.

2. Problem Statement

The core objectives of this project are threefold:

- To conduct a thorough data analysis and generate a comprehensive report based on the provided dataset.
- To construct a predictive model utilizing machine learning techniques capable of classifying different types of skin diseases.
- To formulate actionable recommendations for medical professionals to facilitate the early identification of skin disorders.

3. Attribute Information

The dataset encompasses a rich set of attributes, categorized into clinical observations and histopathological findings:

Clinical Attributes

These attributes represent observable symptoms and patient history:

- **Erythema**: Presence and degree of skin redness or inflammation.
- **Scaling**: Description of flaky or peeling skin.
- **Definite Borders**: Clarity and demarcation of lesion edges.
- **Itching**: Severity of pruritus experienced by the patient.
- **Koebner Phenomenon**: Occurrence of new lesions at sites of skin trauma.
- **Polygonal Papules**: Description of polygon-shaped, raised skin lesions.
- **Follicular Papules**: Raised lesions specifically associated with hair follicles.
- **Oral Mucosal Involvement**: Presence of lesions or symptoms affecting the mucous membranes of the mouth.
- **Knee and Elbow Involvement**: Localization of lesions on the knees and/or elbows.
- **Scalp Involvement**: Presence of lesions on the scalp.
- **Family History**: A binary indicator (0 or 1) denoting a family history of skin disorders.
- **Age**: The patient's age in years.

Histopathological Attributes

These attributes are derived from microscopic examination of skin tissue biopsies:

- **Melanin Incontinence**: Leakage of melanin pigment into the dermis.
- **Eosinophils in Infiltrate**: Presence of eosinophils (a type of white blood cell) within inflammatory infiltrates, often indicative of allergic reactions.
- **PNL Infiltrate**: Presence of polymorphonuclear leukocytes (PNLs) in the infiltrate.
- **Fibrosis of Papillary Dermis**: Excessive formation of fibrous connective tissue in the superficial layer of the dermis.
- **Exocytosis**: Migration of inflammatory cells from the dermis into the epidermis.
- **Acanthosis**: Abnormal thickening of the stratum spinosum (prickle cell layer) of the epidermis.
- **Hyperkeratosis**: Thickening of the stratum corneum (outermost layer of the epidermis).
- **Parakeratosis**: Retention of nuclei in the cells of the stratum corneum, indicating abnormal keratinization.
- **Clubbing of Rete Ridges**: Enlargement and bulbous appearance of the epidermal projections (rete ridges) into the dermis.
- **Elongation of Rete Ridges**: Abnormal lengthening of the rete ridges.
- **Thinning of Suprapapillary Epidermis**: Reduction in the thickness of the epidermis above the dermal papillae.
- **Spongiform Pustule**: Intraepidermal collection of neutrophils forming a pustule.
- **Munro Microabscess**: Small collections of neutrophils within the stratum corneum.
- **Focal Hypergranulosis**: Localized thickening of the stratum granulosum (granular layer of the epidermis).
- **Disappearance of Granular Layer**: Absence of the granular layer in the epidermis.
- **Vacuolization and Damage of Basal Layer**: Formation of empty spaces (vacuoles) and damage to the cells of the basal layer of the epidermis.
- **Spongiosis**: Intercellular edema (fluid accumulation) within the epidermis, leading to widening of intercellular spaces.
- **Saw-tooth Appearance of Rete Ridges**: Irregular, jagged appearance of the rete ridges.
- **Follicular Horn Plug**: Keratin plugs obstructing hair follicles.
- **Perifollicular Parakeratosis**: Abnormal keratinization around hair follicles.
- **Inflammatory Mononuclear Infiltrate**: Presence of chronic inflammatory cells (lymphocytes, macrophages) in the dermis.
- **Band-like Infiltrate**: A dense, linear accumulation of inflammatory cells in the upper dermis.

4. Classes of Skin Diseases

The project focuses on classifying the following six distinct types of skin diseases:

1. **Psoriasis**: A chronic autoimmune condition characterized by red, scaly patches, often found on the scalp, knees, elbows, and lower back.
2. **Seborrheic Dermatitis**: A common skin condition that primarily affects the scalp, causing scaly patches, red skin, and stubborn dandruff. It can also affect other oily areas of the body.
3. **Lichen Planus**: An inflammatory condition that can affect the skin, hair, nails, and mucous membranes. It typically presents as itchy, purple, flat-topped bumps.
4. **Pityriasis Rosea**: A common, mild skin rash that usually begins with a single large patch (herald patch) followed by smaller, widespread patches. It typically resolves on its own within 6 to 10 weeks.
5. **Chronic Dermatitis**: A broad term encompassing various forms of long-lasting skin inflammation, including different types of eczema, characterized by symptoms like swelling, redness, itching, and dryness.
6. **Pityriasis Rubra Pilaris**: A rare, chronic inflammatory skin condition characterized by widespread redness, scaling, and thickening of the skin, often with distinctive follicular papules.

5. Business Case

The current diagnostic process for skin diseases is often challenged by the symptomatic overlap between different conditions, leading to potential misdiagnoses. Furthermore, invasive procedures like biopsies are time-consuming, costly, and can cause patient discomfort. This project's machine learning approach offers significant business advantages:

- **Enhanced Diagnostic Speed and Accuracy**: Automated prediction can provide rapid, data-driven insights, reducing diagnostic delays.
- **Reduced Reliance on Invasive Procedures**: By offering a strong preliminary diagnosis, the model can potentially decrease the necessity for all biopsies, leading to cost savings and improved patient experience.
- **Personalized Treatment**: More accurate diagnoses enable tailored treatment plans, leading to better patient outcomes.
- **Optimized Resource Allocation**: Dermatologists can prioritize cases that are more complex or severe, improving the efficiency of healthcare services.
- **Building Trust with Explainable AI**: Developing models with explainable AI (XAI) capabilities can provide insights into the decision-making process, fostering greater trust among clinicians in the model's predictions.

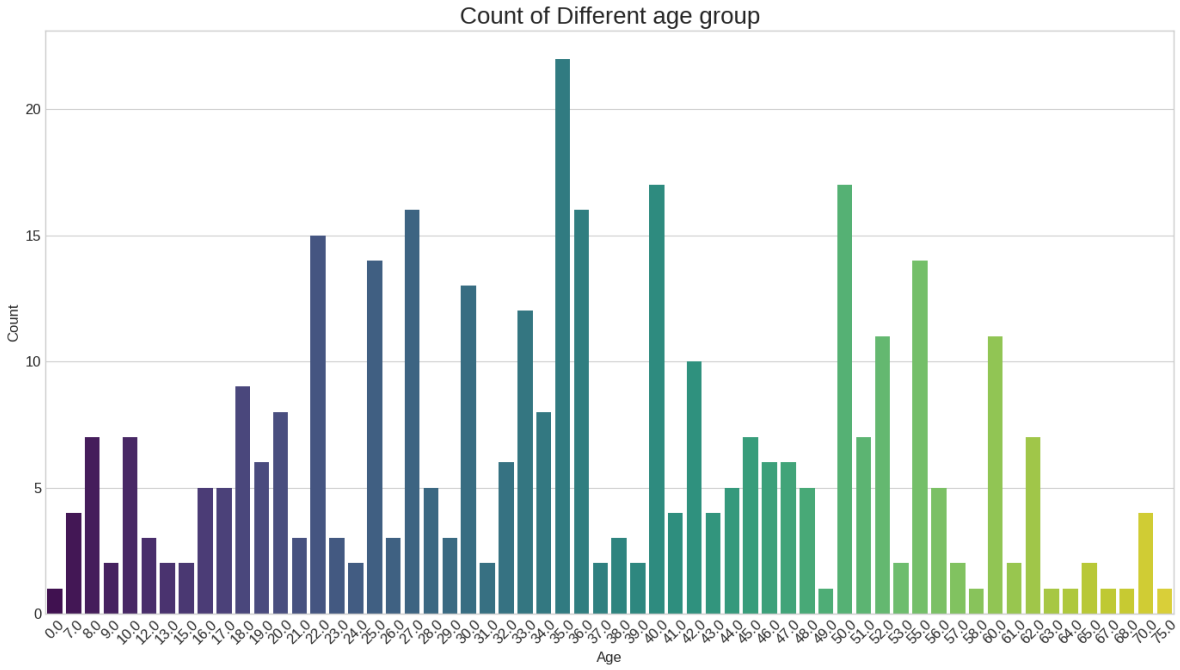
6. Data Loading and Initial Exploration

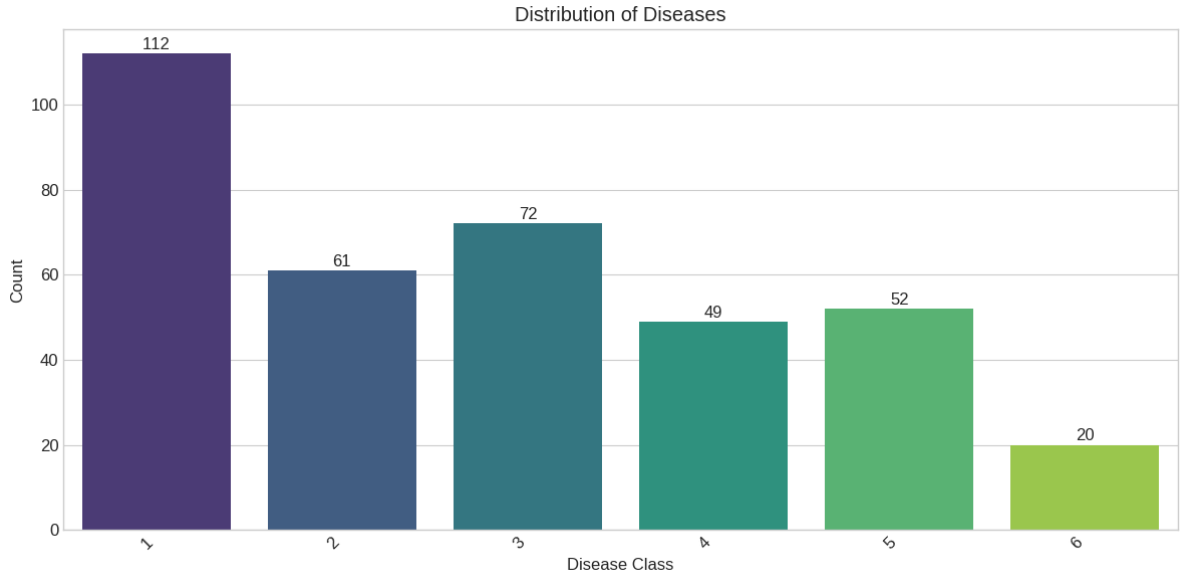
The project initiates by loading the dataset_35_dermatology.csv file into a pandas DataFrame. Following data loading, an initial exploration is performed to understand the dataset's structure, including displaying the first and last few rows of the DataFrame. The notebook also configures visualization settings using seaborn and matplotlib for subsequent exploratory data analysis.

Essential machine learning libraries from sklearn, xgboost, and imblearn are imported, setting the stage for data preprocessing, model development, and comprehensive evaluation.

The subsequent stages of the project, including detailed data analysis, feature engineering, model development, evaluation, comparison, and the formulation of specific recommendations, are thoroughly elaborated within the provided Jupyter notebook.

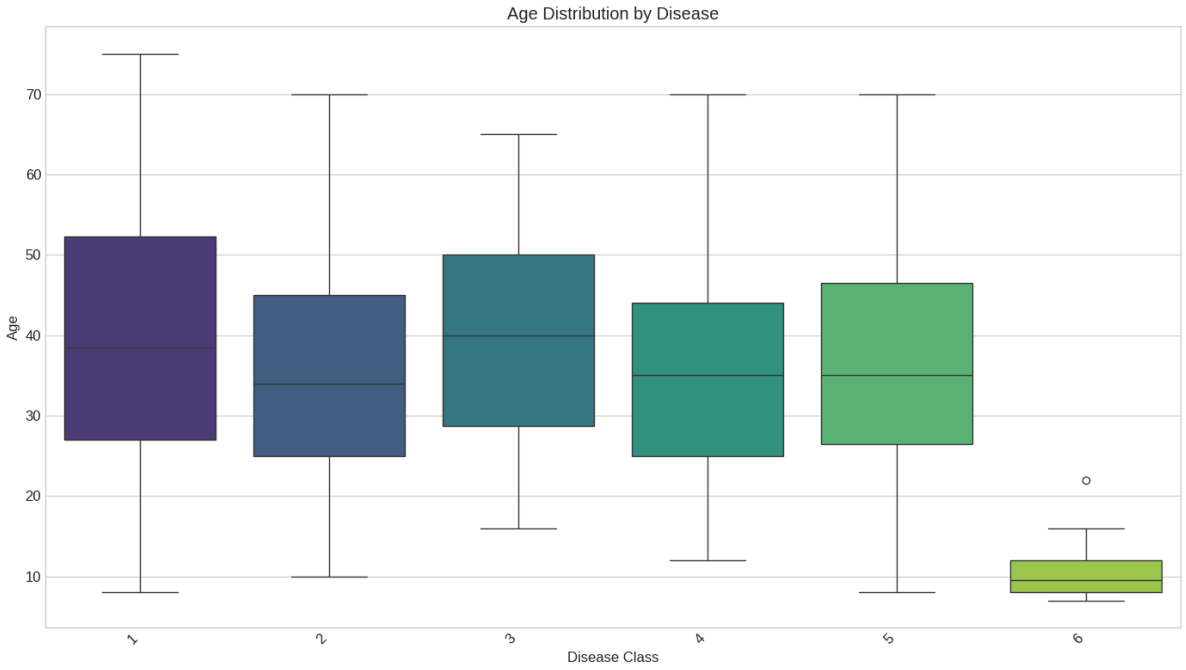
0 As a age present in age columns which is not possible so we will treat it with forward fill or backward fill in upcoming section





Observations from the Distribution of Diseases Chart:

- Class 1 is the most frequent: It has the highest count, with 112 occurrences. This suggests that psoriasis might be the most prevalent disease in this dataset
- Class 6 is the least frequent: It has the lowest count, with only 20 occurrences. This indicates a potential class imbalance, which might need to be considered during model training.
- Other classes have intermediate frequencies: Classes 2, 3, 4, and 5 have counts of 61, 72, 49, and 52 respectively, falling between the most and least frequent classes.
- Potential Class Imbalance: The significant difference in the number of samples between Class 1 and Class 6 highlights a potential class imbalance issue. This could lead to a model that is biased towards the majority class (Class 1) and performs poorly on the minority class (Class 6).



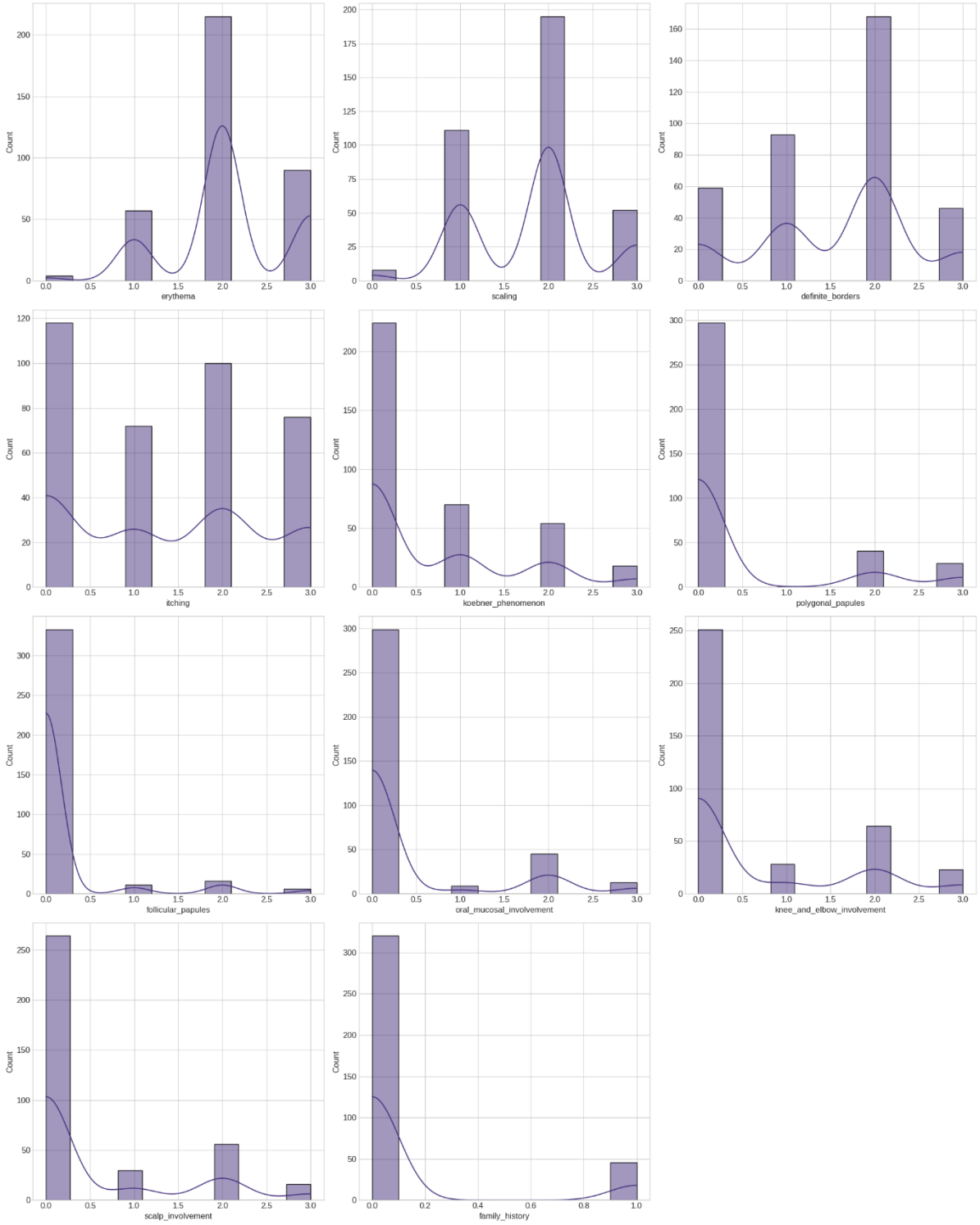
Observations from the Age Distribution by Disease Chart

The boxplot above illustrates the distribution of patient ages across the six different skin disease classes present in the dataset. Key observations include:

- **Varied Median Ages:** The median age appears to differ across the disease categories, suggesting that age might be a contributing factor in distinguishing between these conditions.
- **Differing Age Ranges:** The spread of ages (as indicated by the interquartile range and the whiskers) varies between the diseases. Some diseases show a broader age range of affected individuals compared to others.
- **Potential Outliers:** One potential outlier is visible in the age distribution for Disease Class 6, indicating a patient significantly older than the typical age range for that group in this dataset.
- **Implication for Modeling:** These differences in age distribution suggest that the 'age' feature could be a useful predictor in our machine learning models.

Further statistical analysis might be beneficial to confirm the significance of these observed differences.

Insights from Univariate Histograms



This image displays the distribution of several individual features, providing a univariate view of the dataset:

- **Discrete/Ordinal Nature:** Most of the features
  (e.g., erythema, scaling, definite_borders, itching, koebner_phenomenon, polygonal_papule

s, follicular_papules, oral_mucosal_involvement, knee_and_elbow_involvement, scalp_involvement, family_history) are represented by discrete integer values, typically ranging from 0 to 3. This suggests they capture the absence (0) or varying degrees/levels of presence (1, 2, 3) of specific symptoms or findings.

- **Skewed Distributions:** Many features show a strong concentration of data points at the lower end of their scale (often at 0 or 1). This indicates that the higher degrees of these symptoms or findings are less common across the entire dataset.
- **Presence/Absence Indicators:** Features like family_history primarily show counts at 0 (absence) and a smaller count at 1 (presence), acting as binary indicators.
- **Multimodal Patterns:** Some features, such as erythema, scaling, and definite_borders, display more complex, multimodal distributions with peaks at different discrete values. This implies there might be subgroups within the overall dataset that exhibit distinct levels for these specific features.
- **Impact on Modeling:** The discrete and often skewed nature of these features means that models capable of handling such distributions (e.g., tree-based models) or those robust to non-normal data might be well-suited. Feature scaling will still be important for distance-based or regularization-heavy models.

Univariate Analysis Insights: Feature Characteristics & Class Distribution

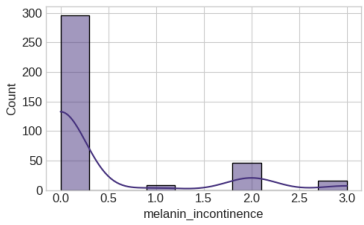
Based on the analysis of individual feature distributions and class counts, here are key insights:

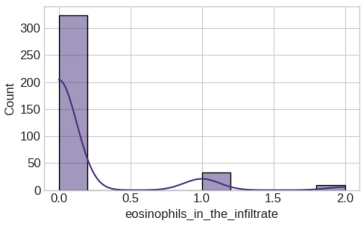
Feature Value Ranges:

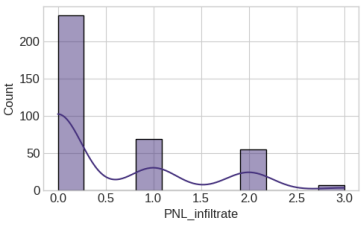
- Most attributes (clinical and histopathological) have values ranging from 0 to 3. This suggests an ordinal scale indicating severity or presence.
- The eosinophils_in_the_infiltrate attribute is an exception, with values limited to 0, 1, and 2.

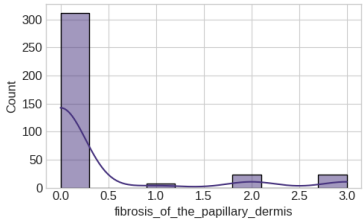
General Feature Trends & Class Relationships:

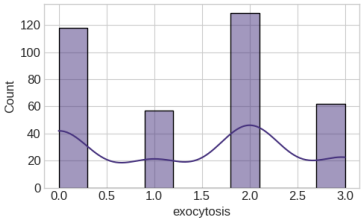
- A value of 0 (feature not present) is most likely to be observed across all disease classes. This indicates that the absence of a feature is common regardless of the specific diagnosis.
- **Class 1 (Psoriasis)** tends to exhibit the highest values (0, 1, 2, 3) across the majority of attributes. This suggests that psoriasis cases often present with more pronounced or a wider range of symptoms/histopathological findings.
- Hyperkeratosis and parakeratosis show a high probability of appearing across multiple classes. This means these features might be less specific in differentiating between diseases on their own.
- Thinning_of_the_suprapapillary_epidermis and spongiform_pustule are particularly notable as they register a value of 3 *only* for **Class 1 (Psoriasis)**. This makes them potentially strong indicators for Psoriasis.
- Follicular_horn_plug and perifollicular_parakeratosis display patterns similar to thinning_of_the_suprapapillary_epidermis and spongiform_pustule, suggesting they might also be strong differentiating features, especially for Class 1.

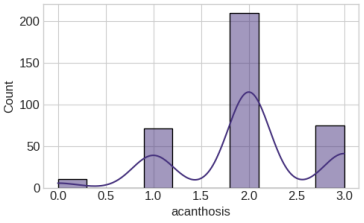

Count
melanin_incontinence

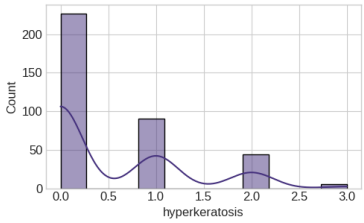

Count
eosinophils_in_the_infiltrate

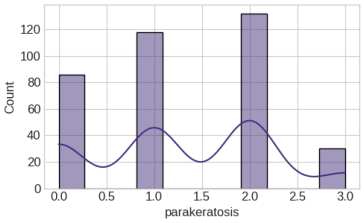

Count
PNL_infiltrate

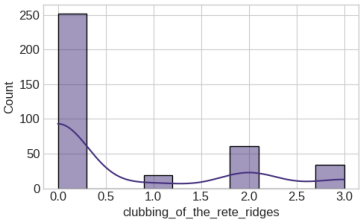

Count
fibrosis_of_the_papillary_dermis

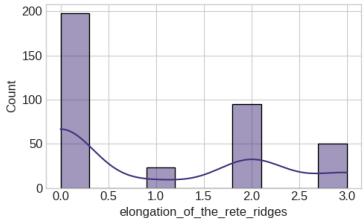

Count
exocytosis

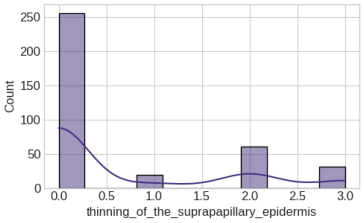

Count
acanthosis

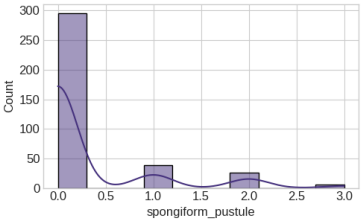

Count
hyperkeratosis

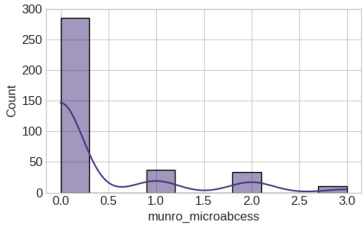

Count
parakeratosis

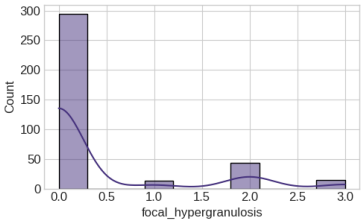

Count
clubbing_of_the_rete_ridges

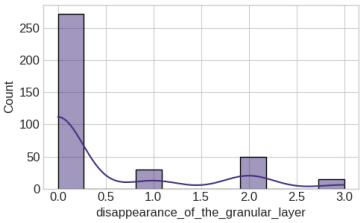

Count
elongation_of_the_rete_ridges

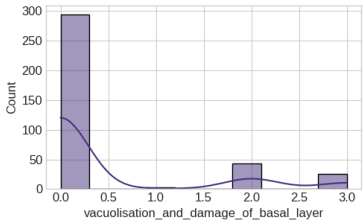

Count
thinning_of_the_suprapapillary_epidermis

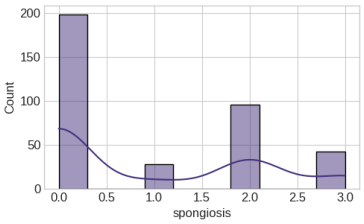

Count
spongiform_pustule

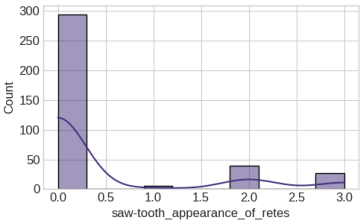

Count
munro_microabcess

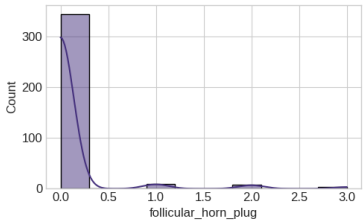

Count
focal_hypergranulosis

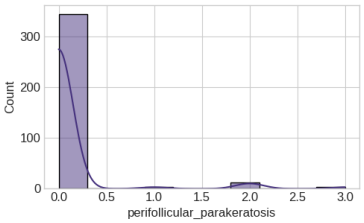

Count
disappearance_of_the_granular_layer

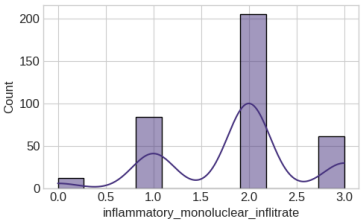

Count
vacuolisation_and_damage_of_basal_layer

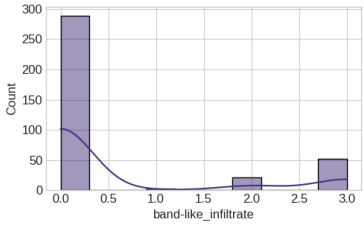

Count
spongiosis


Count
saw-tooth_appearance_of_retes


Count
follicular_horn_plug


Count
perifollicular_parakeratosis


Count
inflammatory_monoluclear_infiltrate

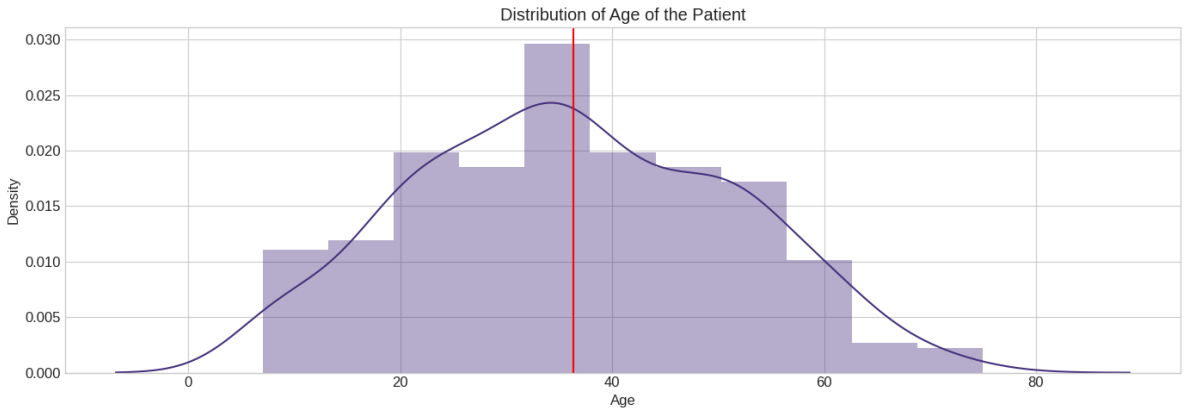

Count
band-like_infiltrate

Disease Class Distribution:

- The dataset includes six disease classes: Psoriasis, Lichen_Planus, Seboreic_Dermatitis, Chronic_Dermatitis, Pityriasis_Rosea, and Pityriasis_rubra_pilaris.
- Based on count plots, **Psoriasis** is the most common disease in the dataset.
- **Pityriasis_rubra_pilaris** is the least common disease, confirming the class imbalance you addressed with SMOTE.

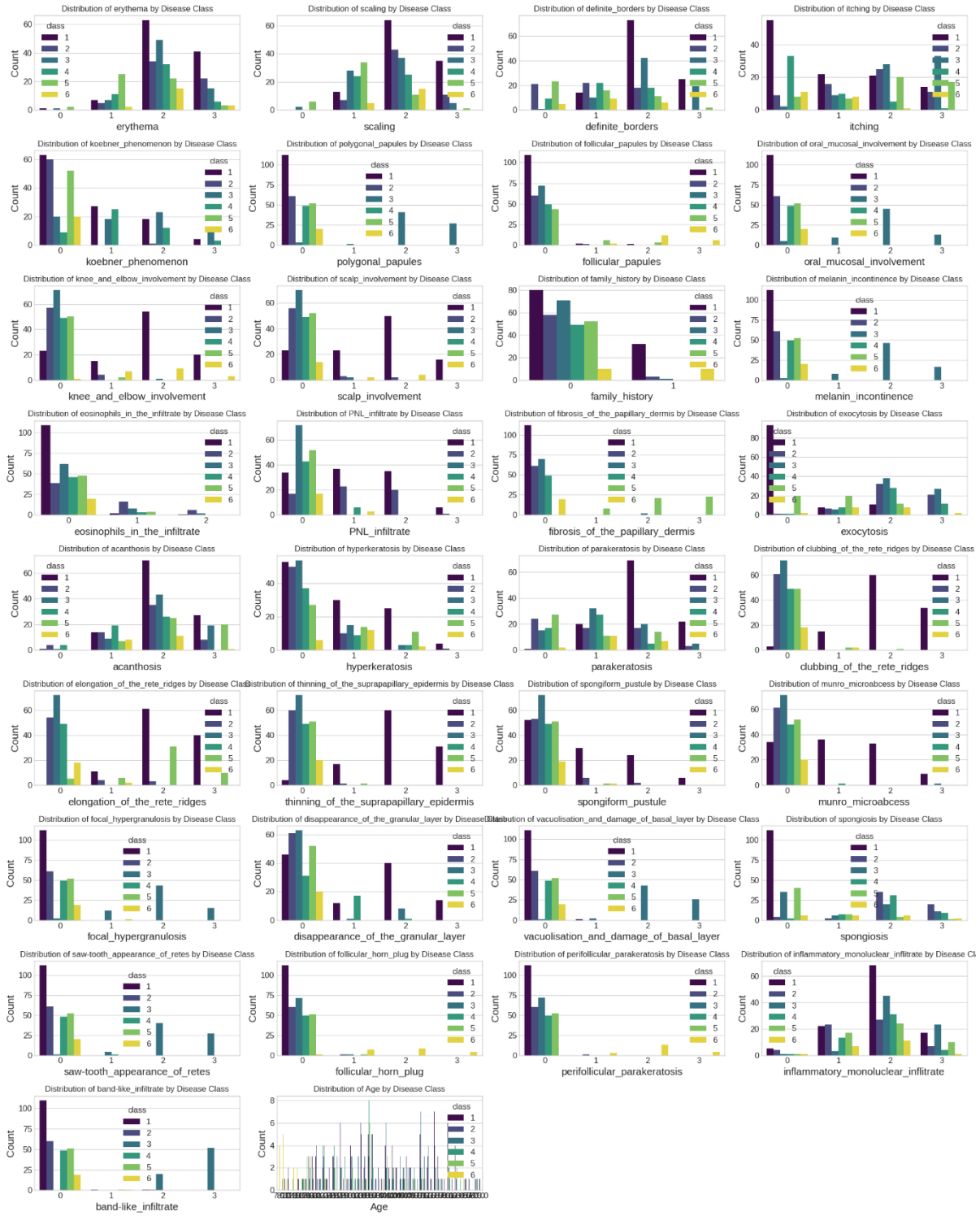
Implications for Modeling:

- Features with values of 3 exclusively for certain classes (e.g., thinning_of_the_suprapapillary_epidermis, spongiform_pustule, follicular_horn_plug, perifollicular_parakeratosis for Class 1) are highly discriminative and will likely be very important for classification models.
- The prevalence of '0' values for many features suggests that the *presence* of a feature (value > 0) is often more informative than its absence.
- The distinct characteristics of Class 1 (Psoriasis) across many features are a positive sign for its accurate classification.
- The overall class imbalance noted (Psoriasis most common, Pityriasis_rubra_pilaris least common) underscores the necessity of techniques like SMOTE for robust model training.



Insights

- Skin diseases can occur even if there is no family history.
- One type of skin problem shows signs like strange bumps, mouth involvement, pigment changes, and harm to a specific layer.
- Another type, class 1 disease, affects the scalp, causing changes like ridge clubbing, tiny abscesses, skin thining, and pimplr-like blisters.
- Good skin health is indicated by the absence of redness, scaling, thickness of the skin, and inflammation.



Key Observations:

Feature Types:

- Most features (like erythema, scaling, acanthosis) are discrete (0, 1, 2, 3).
- Age is continuous.

Diagonal Plots (Individual Feature Distributions by Class):

- Age: Shows clear differences in age distribution across classes. Some classes are concentrated in younger/older groups.
- Clinical/Histological Features: For erythema, scaling, acanthosis, etc., we see different classes peaking at different score levels. This indicates these features help separate classes.

Off-Diagonal Plots (Feature Relationships by Class):

Features vs. Age (Bottom Row / Right Column):

- These plots (e.g., scaling vs. Age, acanthosis vs. Age) are crucial for understanding features_vs_age.
- Different classes often cluster in distinct areas, showing how certain features combine with age to define a class.

Feature vs. Feature:

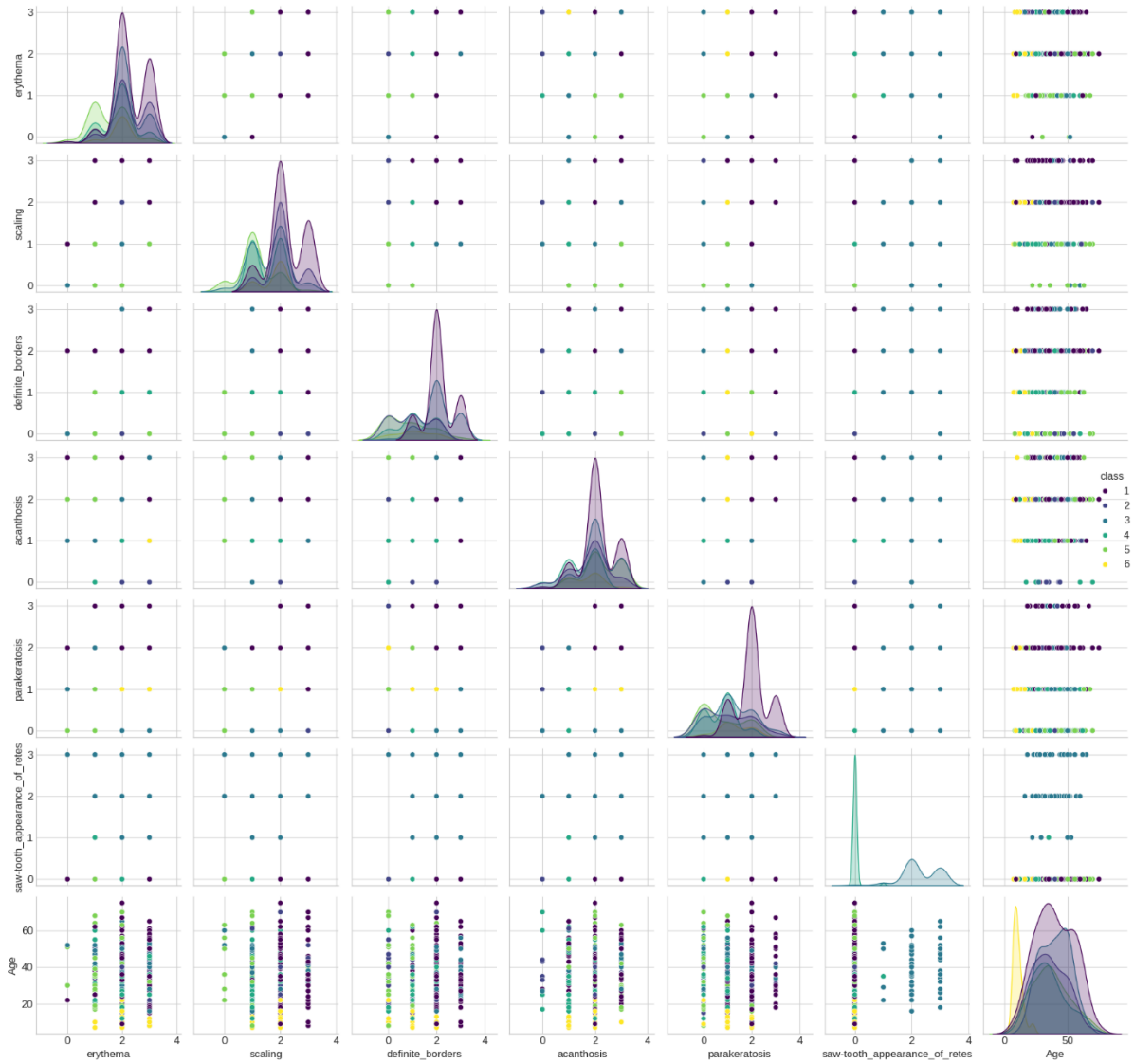
- For discrete features, you see grids of points. The color density within a grid cell shows which classes frequently have those combined feature values. If different colors (classes) separate or occupy distinct regions in any plot, it suggests strong differentiation.

The selected features, both individually and in combination (especially with Age), show clear patterns that help distinguish between the different disease classes. This visual analysis confirms their potential for classification.

✅ Strong Differentiating Features (Highly Class-Specific):

- **polygonal_papules**: Shows a clear presence predominantly in **Class 3**, making it a strong indicator for this class.
- **follicular_papules**: Almost exclusively observed in **Class 6**.
- **family_history**: Uniquely prevalent in **Class 5**, distinguishing it from others.
- **koebner_phenomenon**: More commonly seen in **Class 2 and 3**.
- **oral_mucosal_involvement**: Primarily present in **Class 6**.
- **knee_and_elbow_involvement**: More common in **Class 6** and somewhat in **Class 1**.

⚠️ Features with Variable Presence (Some Overlap but Still Informative):

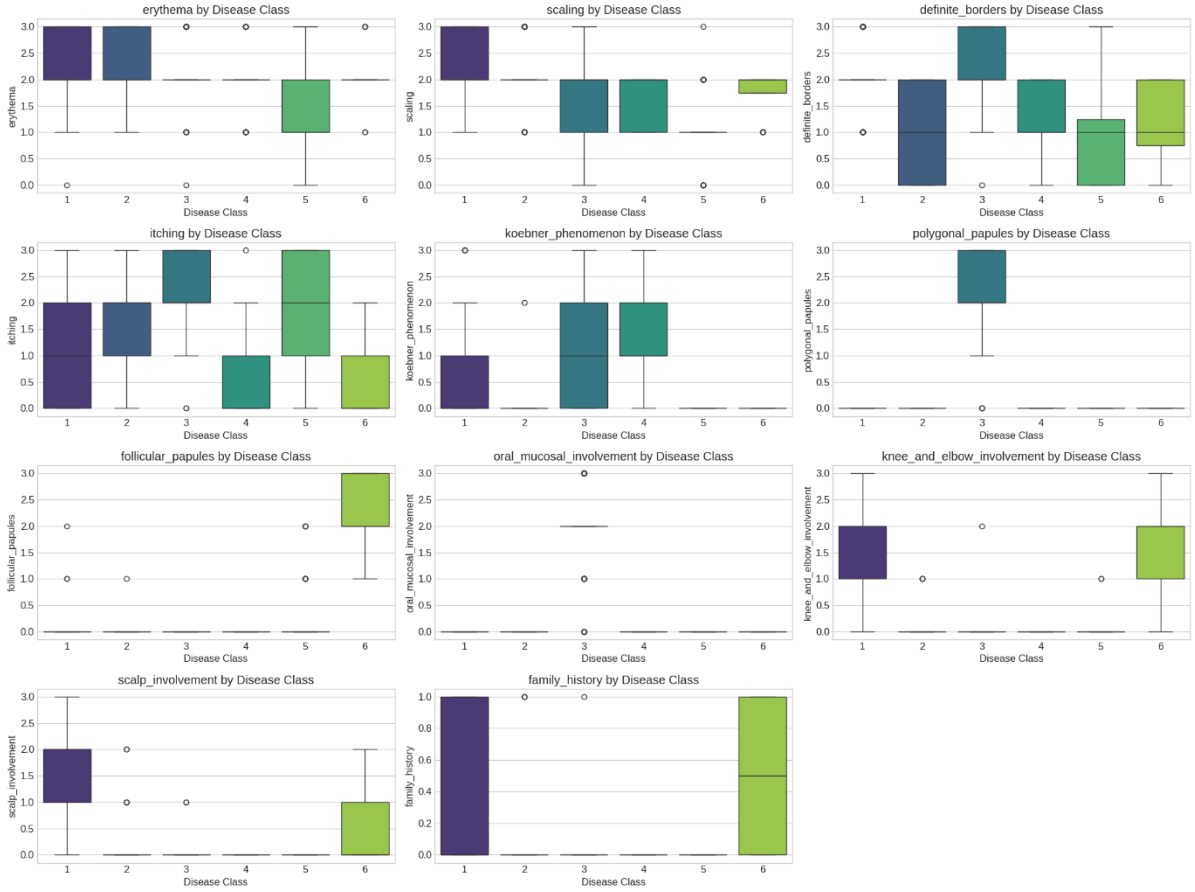


- **erythema**, **scaling**, **itching**: These features show varying median values and spread across classes (e.g., Classes **1, 2, 3** tend to have higher values), though some overlap exists. They contribute to differentiation but are not always definitive.
- **definite_borders**: Tends to have higher median values for **Classes 3 and 6**.
- **scalp_involvement**: Observed to some extent in **Classes 1 and 6**.

Conclusion:

This analysis confirms that many of the clinical features are **highly valuable for differentiating between disease classes**. Features such as polygonal_papules, follicular_papules, and family_history are particularly strong discriminators due to their near-exclusive presence or absence in specific classes.

These visual insights are crucial for understanding the **distinct clinical profiles** of each disease and can guide **further diagnostic and modeling approaches**.



🔬 Key Insights from Histopathological Feature Box Plots

🧪 Class-Specific Strong Indicators:

- **Class 2**:
    - spongiform_pustule, munro_microabscess – almost exclusively present, making them strong discriminators.
- **Class 3**:
    - band_like_infiltrate, acanthosis, hyperkeratosis, parakeratosis, elongation_of_the_rete_ridges, clubbing_of_the_rete_ridges – all show high median values, making this class well-defined histologically.
- **Class 5**:
    - melanin_incontinence, focal_hypergranulosis, band_like_infiltrate – moderately unique to this class.
- **Class 6**:
    - follicular_horn_plug, perifollicular_parakeratosis – predominantly seen in this class.

📊 Broadly Present but Differentiating Features:

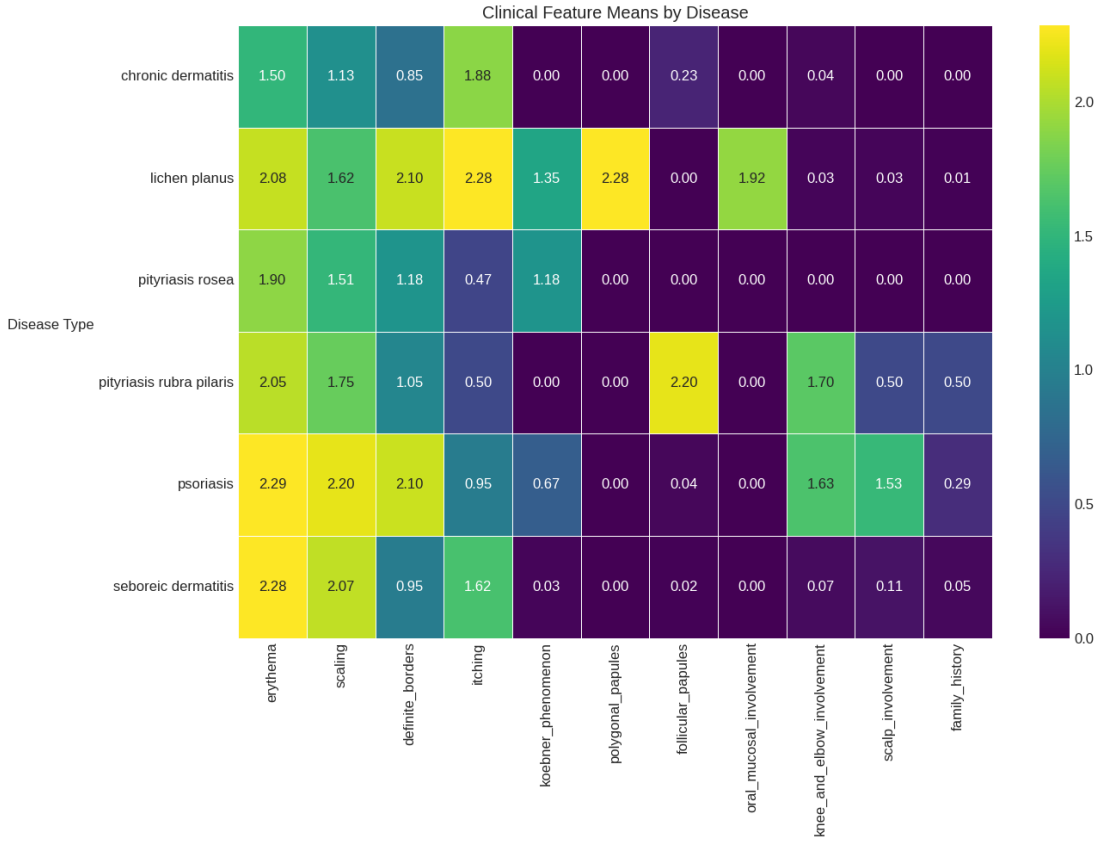
- acanthosis, hyperkeratosis, parakeratosis, elongation_of_the_rete_ridges – prevalent in Classes 2, 3, 4, 6 with varying intensities.
- spongiosis, exocytosis, inflammatory_mononuclear_infiltrate – seen across multiple classes, but with differentiating medians.
- disappearance_of_the_granular_layer, saw-tooth_appearance_of_retes – contribute to distinguishing Classes 1, 2, 3, 5, and 6.

Overall Summary:

Several histopathological features act as **powerful discriminators** between disease classes.

- **Highly class-specific features** (e.g., spongiform_pustule, follicular_horn_plug, band_like_infiltrate) provide strong diagnostic signals.
- **Broad but patterned features** (e.g., acanthosis, parakeratosis) support differentiation through distribution and median analysis.

These insights can guide **feature selection and model interpretation** in classification tasks.



Observations on Clinical Feature Means by Disease Heatmap

The heatmap visualizes the average intensity of each clinical feature (erythema, scaling, etc.) for each of the six skin diseases.

- **Distinct Clinical Signatures:** Each disease exhibits a unique profile of average clinical feature values, suggesting that the combination of these signs can be indicative of a specific condition.
- **Common Symptoms:** 'Erythema' and 'scaling' show relatively high average values across multiple diseases, consistent with their classification as erythemato-squamous dermatoses.
- **Differential Diagnostic Clues:** Certain features appear to be more pronounced in specific diseases:
    - **Psoriasis:** High in scaling, definite borders, and involvement of knees, elbows, and scalp.
    - **Seboreic Dermatitis:** High in scaling and scalp involvement, but less so in definite borders.
    - **Lichen Planus:** Notable for polygonal papules and oral mucosal involvement.
    - **Pityriasis Rosea:** Characterized by itching.
    - **Pityriasis Rubra Pilaris:** Shows follicular papules.
- **Family History Variation:** The average presence of family history differs across the diseases, hinting at potential genetic predispositions.

Key Observations from the Correlation Heatmap:

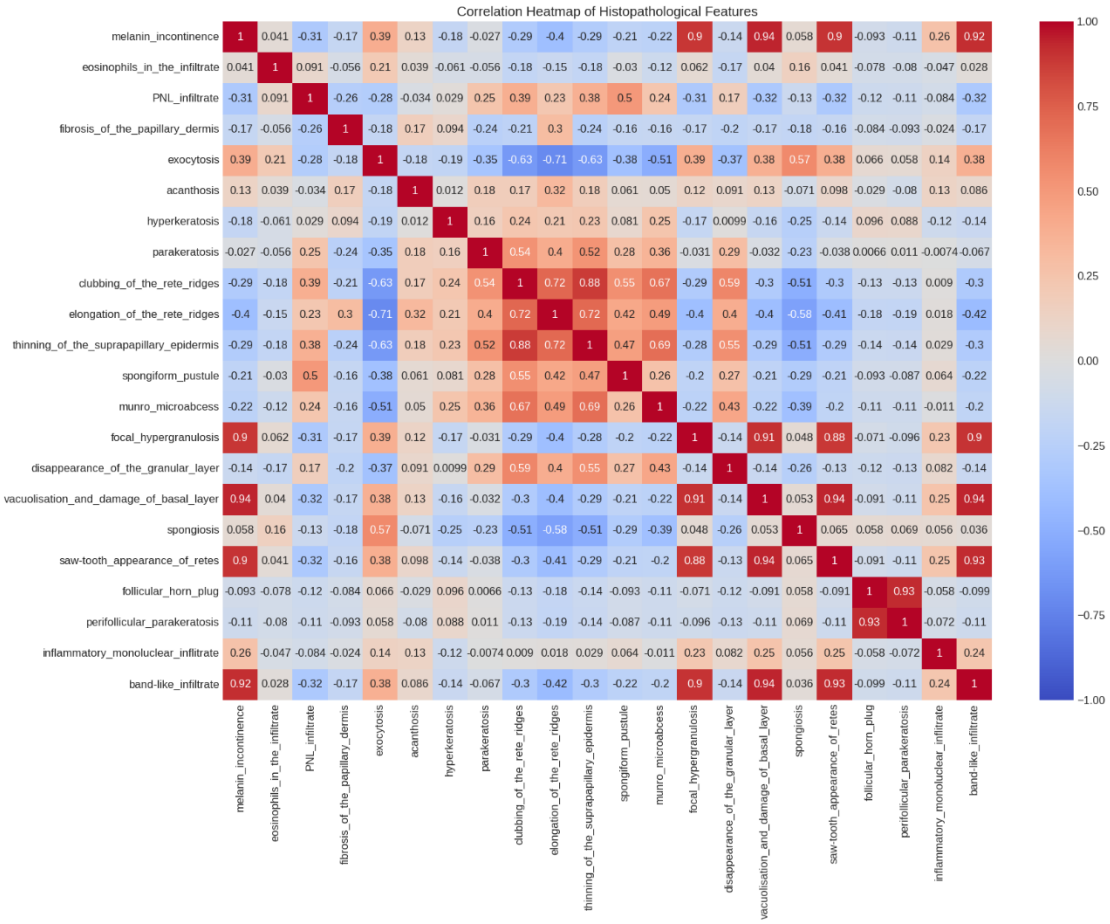
Strong Positive Correlations:

- Hyperkeratosis and Parakeratosis (0.86): This very strong positive correlation suggests that the presence and severity of increased thickness of the stratum corneum (hyperkeratosis) often coincide with abnormal keratinization with retained nuclei in the stratum corneum (parakeratosis). This is a common finding in many scaling skin disorders.
- Clubbing of the rete ridges and Elongation of the rete ridges (0.72): These features related to the downward projections of the epidermis are strongly associated.
- Elongation of the rete ridges and Acanthosis (0.63): Thickening of the epidermis (acanthosis) often occurs with the elongation of rete ridges.
- Spongiform pustule and Munro microabcess (0.67): These are both signs of neutrophilic accumulation within the epidermis, often seen together in certain conditions like psoriasis.
- Saw-tooth appearance of retes and Band-like infiltrate (0.93): This very strong positive correlation is highly characteristic of Lichen Planus.

Strong Negative Correlations:

- Thinning of the suprapapillary epidermis and Clubbing/Elongation of the rete ridges (around -0.4 to -0.5): This suggests that when the rete ridges are more pronounced, the epidermal layer above the dermal papillae tends to be thinner, which is a structural relationship.
- Disappearance of the granular layer with Hyperkeratosis and Parakeratosis (around -0.4 to -0.5): The absence of the granular layer is often associated with abnormal keratinization processes.

- Moderate Correlations: Several other pairs show moderate positive or negative correlations, indicating tendencies for these features to occur together or inversely.
- Weak Correlations: Many pairs of histopathological features show weak correlations (values close to zero), suggesting they tend to vary more independently of each other.



- Melanin incontinence and Vacuolisation and damage of basal layer (0.94): This very strong positive correlation suggests that the leakage of melanin into the dermis (melanin incontinence) is highly associated with damage to the basal layer of the epidermis, often presenting as vacuolisation. This is a significant finding and likely reflects a common underlying pathological process where basal layer damage leads to the release of melanin.
- Vacuolisation and damage of basal layer and Saw-tooth appearance of retes (0.94): This also indicates a very strong association. The saw-tooth pattern of the rete ridges (epidermal projections into the dermis) frequently occurs with damage and vacuolisation of the basal layer. This is a hallmark histopathological feature of Lichen Planus, so this strong correlation aligns with that understanding.
- Saw-tooth appearance of retes and Vacuolisation and damage of basal layer (0.94): This is the same correlation as above, just with the features reversed.

- Vacuolisation and damage of basal layer and Band-like infiltrate (0.92): This strong positive correlation indicates that damage to the basal layer is often accompanied by a dense, horizontal band of inflammatory cells in the upper dermis (band-like infiltrate). This is another key characteristic feature, particularly of Lichen Planus

Top differentiating features for each disease:

psoriasis:

- Age: 4.738
- clubbing_of_the_rete_ridges: 2.092
- thinning_of_the_suprapapillary_epidermis: 2.046

seboreic dermatitis:

- spongiosis: 1.452
- Age: 1.085
- exocytosis: 0.993

lichen planus:

- Age: 4.389
- band-like_infiltrate: 2.698
- vacuolisation_and_damage_of_basal_layer: 2.302

pityriasis rosea:

- Age: 1.268
- spongiosis: 1.161
- elongation_of_the_rete_ridges: 1.145

chronic dermatitis:

- fibrosis_of_the_papillary_dermis: 2.276
- elongation_of_the_rete_ridges: 1.041
- definite_borders: 0.819

pityriasis rubra pilaris:

- Age: 27.623

- follicular_papules: 2.151

- perifollicular_parakeratosis: 2.047

Psoriasis:

- Age (4.333): This suggests psoriasis might tend to occur in a slightly older population compared to some other diseases in this dataset.
- Clubbing of the rete ridges (2.092): This histopathological feature seems to be a significant indicator of psoriasis.
- Thinning of the suprapapillary epidermis (2.046): This is another histopathological finding that appears characteristic of psoriasis.

Seboreic Dermatitis:

- Spongiosis (1.452): This indicates epidermal intercellular edema and is a key feature distinguishing seboreic dermatitis.
- Age (1.003): The lower value compared to psoriasis suggests it might present in a younger age group on average.
- Exocytosis (0.993): The migration of inflammatory cells into the epidermis is also a relevant feature.

Lichen Planus:

- Age (4.458): Similar to psoriasis, lichen planus might also be more prevalent in a slightly older population in this dataset.
- Band-like infiltrate (2.698): As we noted earlier with the correlation analysis, this histopathological feature is a hallmark of lichen planus.
- Vacuolisation and damage of basal layer (2.302): This further supports the characteristic histopathology of lichen planus.

Pityriasis Rosea:

- Age (1.166): This suggests pityriasis rosea tends to affect a younger population.
- Spongiosis (1.161): Similar to seboreic dermatitis, spongiosis plays a role.
- Elongation of the rete ridges (1.145): This histopathological feature also seems to be important.

Chronic Dermatitis:

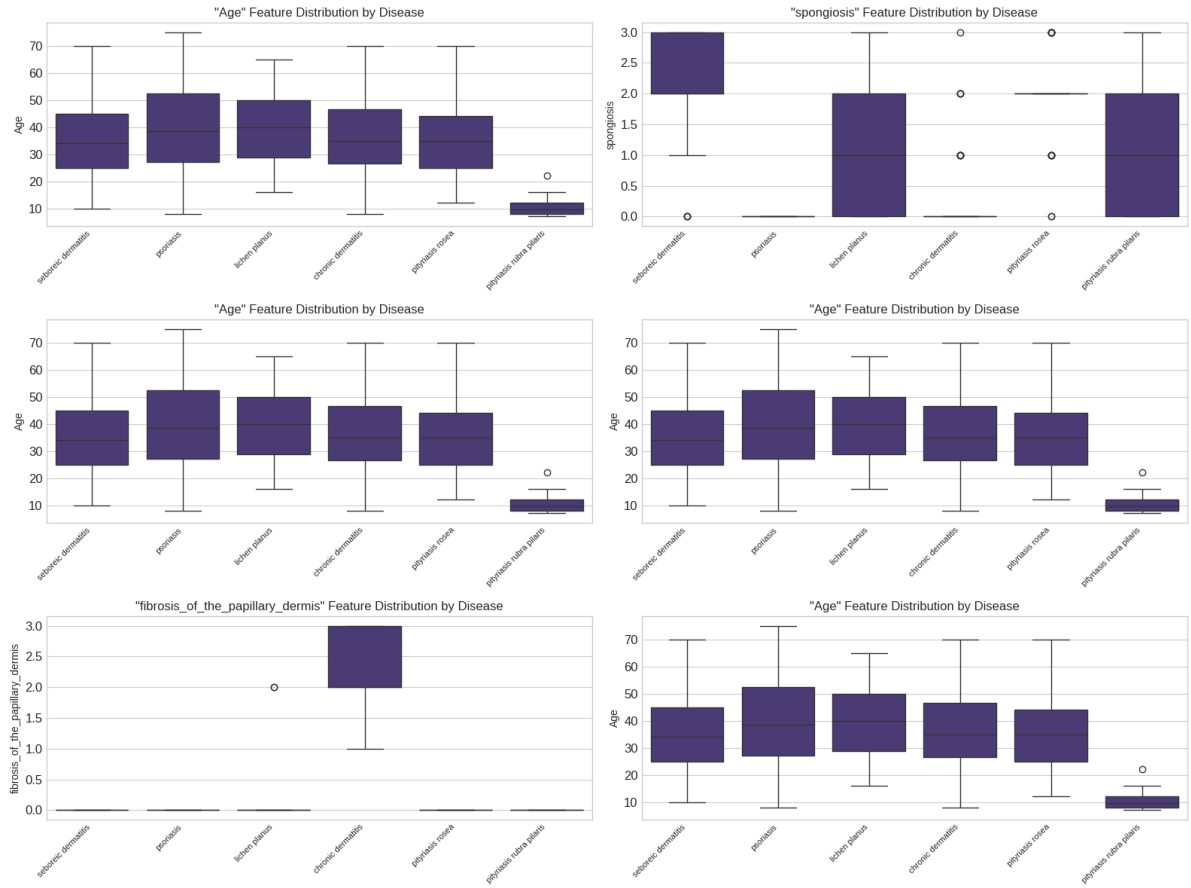
- Fibrosis of the papillary dermis (2.276): This dermal change appears to be a key differentiator for chronic dermatitis.
- Elongation of the rete ridges (1.041): This epidermal feature is also relevant.
- Definite borders (0.819): This clinical feature seems to help in identifying chronic dermatitis.

Pityriasis Rubra Pilaris:

- Age (27.637): The significantly higher age compared to other diseases is a striking differentiating factor.
- Follicular papules (2.151): This clinical feature is a characteristic sign.
- Perifollicular parakeratosis (2.047): This specific pattern of abnormal keratinization around hair follicles is a key histopathological finding.

Overall Implications:

- **Age as a Differentiator:** Age appears to be a significant factor in distinguishing several of these diseases, with psoriasis and lichen planus tending towards older patients, while pityriasis rosea and potentially seboreic dermatitis might be more common in younger individuals (though the scale of age here isn't entirely clear without knowing if it's a standardized or relative value). Pityriasis rubra pilaris stands out with a much higher average "Age" value.
- **Key Histopathological Features:** The histopathological analysis reveals specific microscopic findings that are highly indicative of particular diseases (e.g., band-like infiltrate for lichen planus, clubbing of rete ridges for psoriasis, perifollicular parakeratosis for pityriasis rubra pilaris).
- **Clinical and Histopathological Synergy:** The inclusion of both clinical ('definite borders', 'follicular papules') and histopathological features in this list highlights the importance of considering both aspects for accurate diagnosis.
- Top 15 most significant features based on ANOVA F-test:
- 1. polygonal_papules
- 2. follicular_papules
- 3. oral_mucosal_involvement
- 4. knee_and_elbow_involvement
- 5. melanin_incontinence
- 6. fibrosis_of_the_papillary_dermis
- 7. clubbing_of_the_rete_ridges
- 8. elongation_of_the_rete_ridges
- 9. thinning_of_the_suprapapillary_epidermis
- 10. focal_hypergranulosis
- 11. vacuolisation_and_damage_of_basal_layer
- 12. saw-tooth_appearance_of_retes
- 13. follicular_horn_plug
- 14. perifollicular_parakeratosis
- 15. band-like_infiltrate



Observation ofClinical and Histopathological Severity by Disease Boxplots

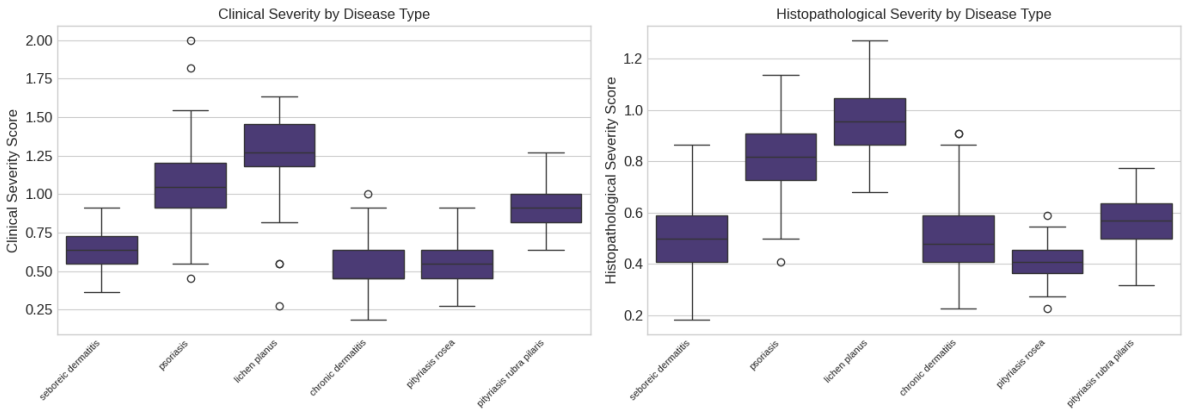
The side-by-side boxplots compare the distributions of clinical_severity and histo_severity scores across the different skin disease types.

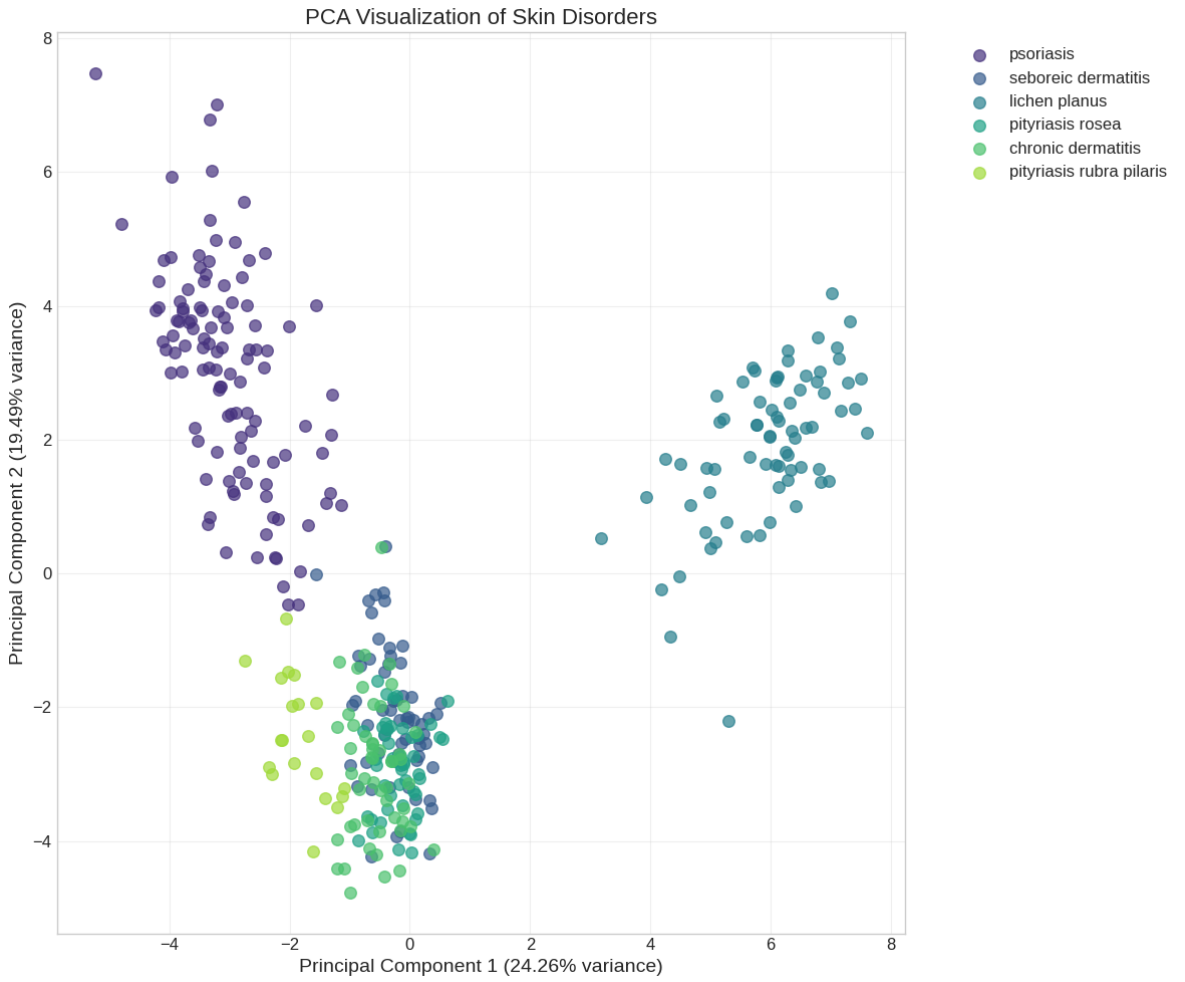
Key Observations:

- **Clinical Severity:** Psoriasis and lichen planus tend to exhibit higher overall clinical symptom severity compared to other diseases.
- **Histopathological Severity:** Lichen planus shows a relatively higher histopathological severity score.
- **Comparative Patterns:** Some diseases, like psoriasis, show high severity in both clinical and histopathological aspects. Others, like lichen planus, might have higher severity in one domain (histopathology) relative to others.
- **Discriminatory Potential:** The degree of overlap in the boxplots suggests how well these severity scores might help in distinguishing between the diseases.

Implications for Modeling:

These visualizations provide insights into the overall severity of clinical and histopathological findings for each disease, which can inform feature selection and model interpretation.





Observations on PCA Visualization of Skin Disorders

The Principal Component Analysis (PCA) plot projects the skin disorder data onto the first two principal components, capturing a significant portion of the variance (approximately 43.75%). The visualization reveals some degree of separation between the different disease classes:

- **Distinct Clusters:** Psoriasis and pityriasis rosea appear to form relatively distinct clusters, suggesting that these conditions have characteristic patterns in the data.
- **Overlapping Regions:** There is noticeable overlap between some disease classes, such as seboreic dermatitis with psoriasis and lichen planus, and chronic dermatitis with pityriasis rosea and pityriasis rubra pilaris. This indicates that these conditions might share some similarities in their feature profiles, potentially making them more challenging to distinguish.
- **Dimensionality Reduction Benefit:** PCA helps to visualize the data in a lower-dimensional space while retaining the most important information. The observed separation suggests that machine learning models should be able to learn to classify these disorders with reasonable accuracy, although the overlap indicates potential for misclassification between certain conditions.

📊 Model Performance Summary

Model	Test Accuracy	Cross-Validation Accuracy (±Std)	Notes
XGBoost	**0.9926**	**0.9822 (±0.0138)**	Best performer overall; excellent precision & recall across classes
Logistic Regression	0.9852	0.9822 (±0.0213)	Very strong generalization, fast training
Support Vector Machine	0.9852	0.9301 (±0.0221)	High accuracy but lower CV score indicates possible overfitting
Gradient Boosting	0.9852	0.9717 (±0.0172)	Strong performance with longer training time
Random Forest	0.9778	0.9777 (±0.0133)	Consistent and reliable; good balance of speed and accuracy
Decision Tree	0.9778	0.9673 (±0.0088)	Simple, interpretable model with solid results
K-Nearest Neighbors	0.9556	0.9376 (±0.0271)	Lower performance on Class 2; fast training
Naive Bayes	0.9037	0.8913 (±0.0261)	Underperforms on Class 2 and 4 due to probabilistic assumptions

Model	Test Accuracy	Cross-Validation Accuracy (±Std)	Notes
AdaBoost	❌ 0.3333	❌ 0.3615 (±0.0637)	Failed to generalize; poor performance on most classes

---

🔍 Key Takeaways

- ✅ **XGBoost**, **Logistic Regression**, and **Gradient Boosting** emerged as top models with near-perfect accuracy and balanced precision-recall.
- ⚠️ **SVM** achieved high test accuracy but showed relatively low cross-validation performance, suggesting some overfitting.
- 🚫 **AdaBoost** failed to learn use

🔢 Sorted Model Performance Summary (by CV Accuracy)

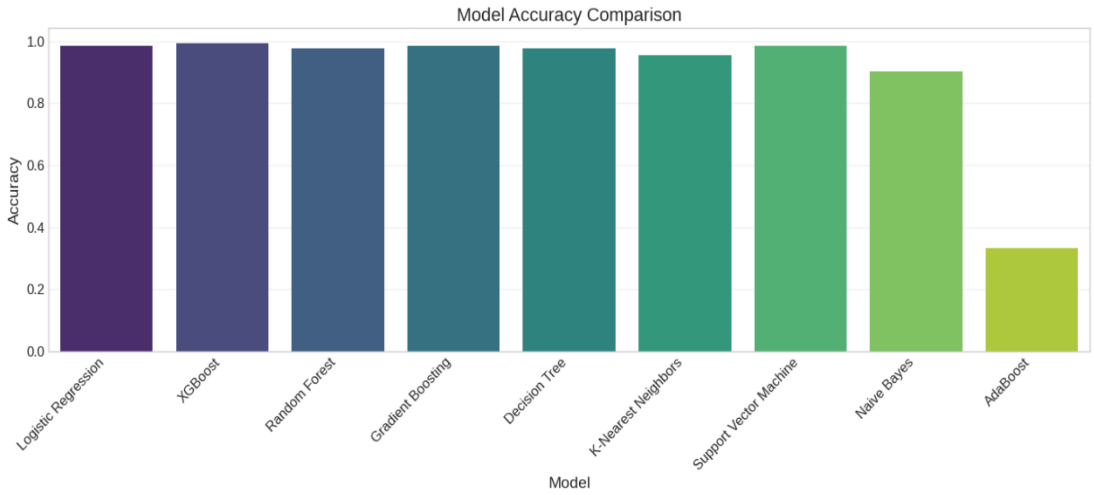
Rank	Model	Test Accuracy	Precision	Recall	F1 Score	CV Accuracy	CV Std Dev	Training Time (s)
1	Logistic Regression	0.9852	0.9867	0.9848	0.9851	**0.9822**	0.0213	7.09
2	XGBoost	**0.9926**	0.9931	0.9924	0.9926	**0.9822**	0.0138	1.55
3	Random Forest	0.9778	0.9776	0.9776	0.9774	0.9777	0.0133	2.59
4	Gradient Boosting	0.9852	0.9858	0.9848	0.9848	0.9717	0.0172	16.42
5	Decision Tree	0.9778	0.9783	0.9779	0.9776	0.9673	0.0088	0.12
6	K-Nearest Neighbors	0.9556	0.9564	0.9552	0.9557	0.9376	0.0271	0.21
7	Support Vector Machine	0.9852	0.9852	0.9852	0.9852	0.9301	0.0221	0.69

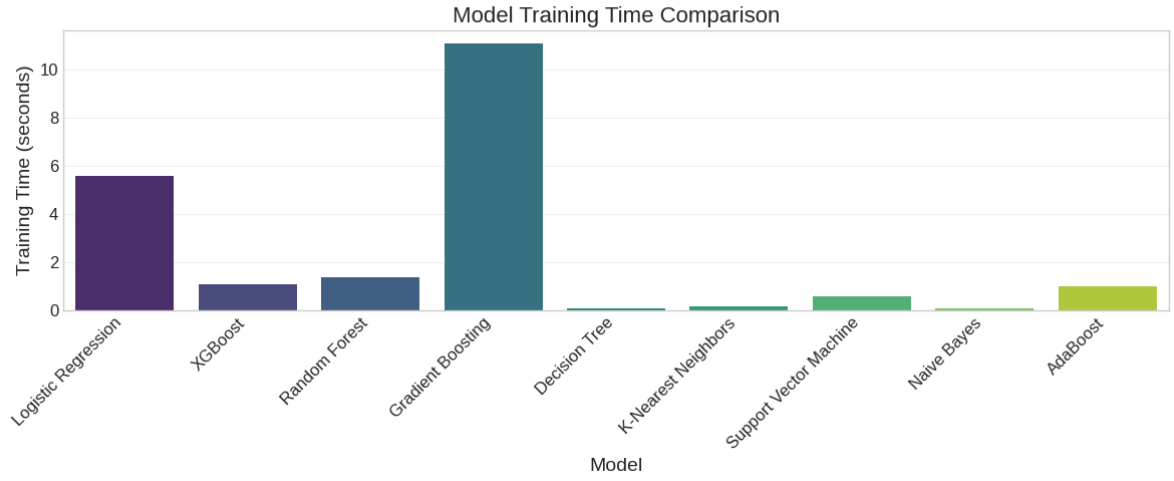
Rank	Model	Test Accuracy	Precision	Recall	F1 Score	CV Accuracy	CV Std Dev	Training Time (s)
8	Naive Bayes	0.9037	0.9356	0.9015	0.8919	0.8913	0.0261	0.05
9	AdaBoost	❌ 0.3333	0.1994	0.3333	0.2214	❌ 0.3615	0.0637	0.82

🏁 Final Recommendation

- ✅ **Top Picks**: Logistic Regression and XGBoost — both provide **excellent test and CV accuracy** with reasonable training times.
- ⚖️ **Balanced Option**: Random Forest is consistent and performs well across all metrics.
- ⚠️ **Underperformers**: AdaBoost (failed completely), Naive Bayes (low recall on key classes).
- 🔍 **SVM** has high test accuracy but lower CV accuracy, indicating potential overfitting.

🔧 Choose based on **interpretability vs. performance vs. speed** depending on the project goal.

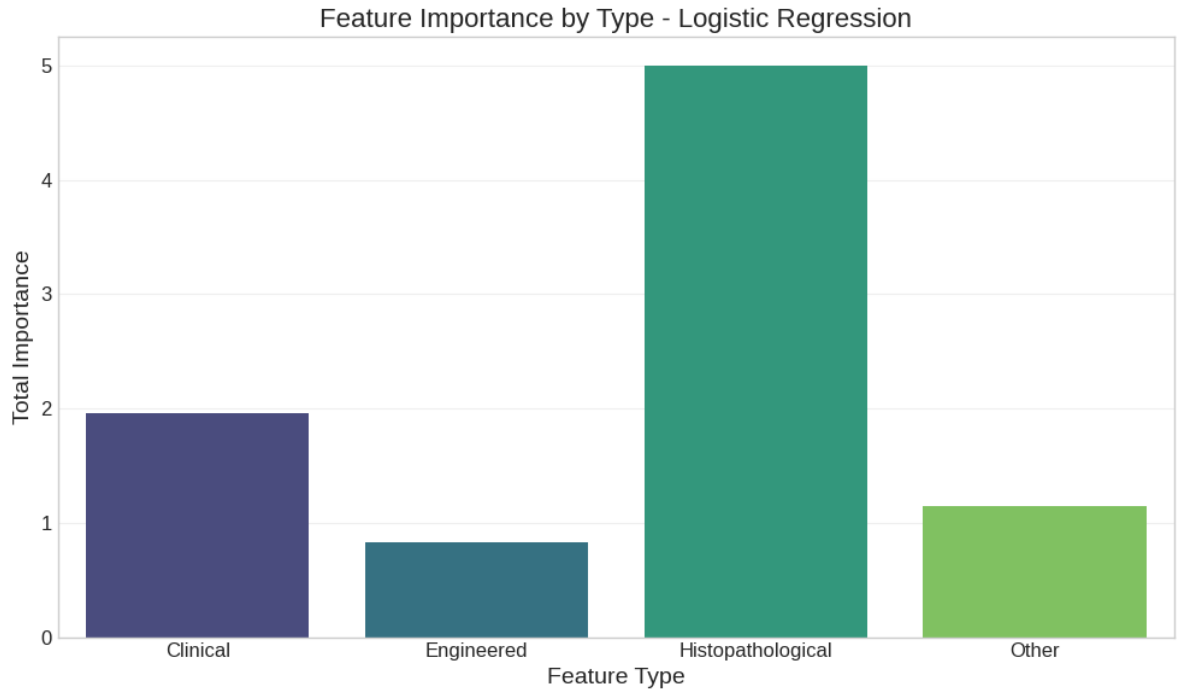


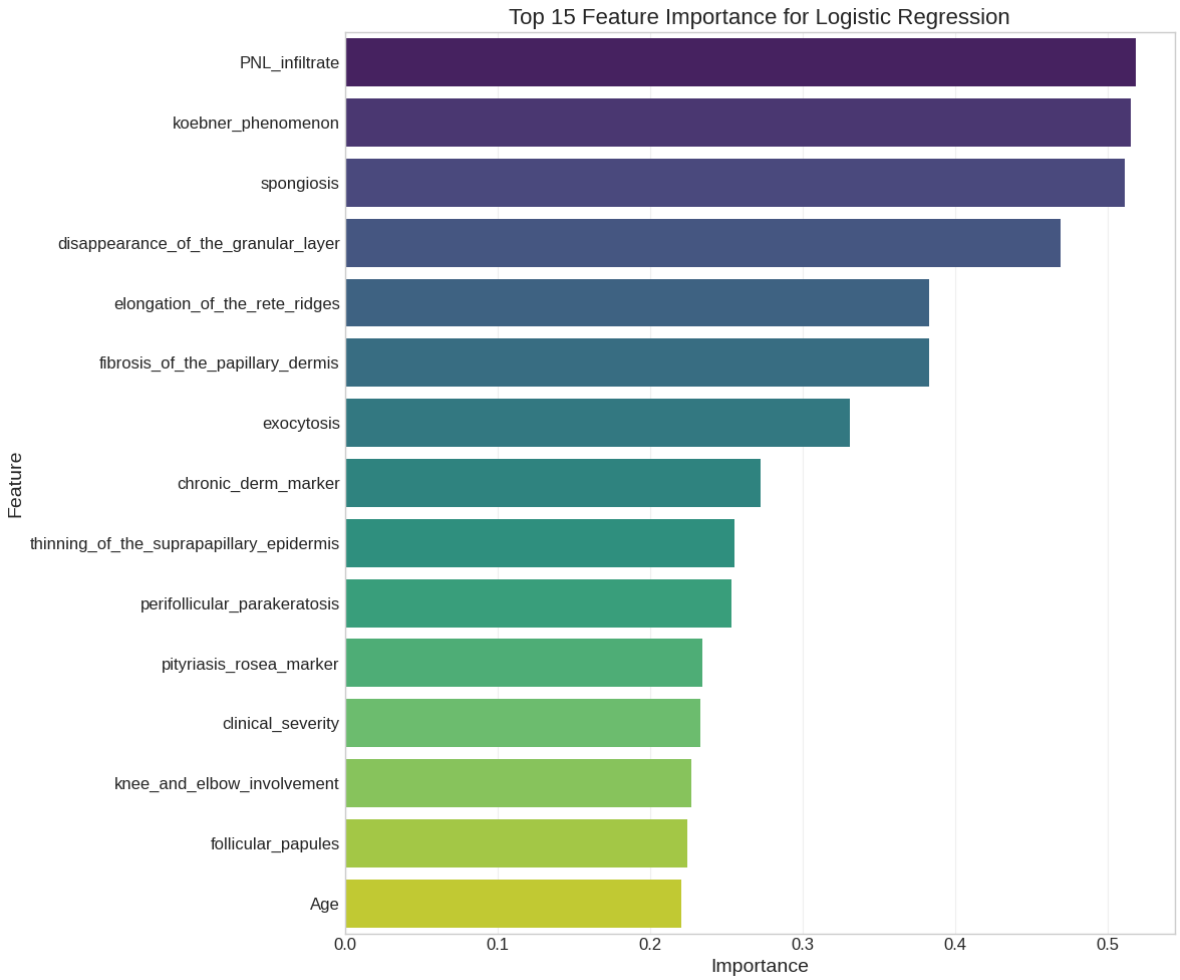


For Logistic Regression:

Tuned model accuracy: 0.9852

Improvement over base model: 0.0000 → This means hyperparameter tuning did not improve accuracy compared to the base model — possibly because the base model was already close to optimal.



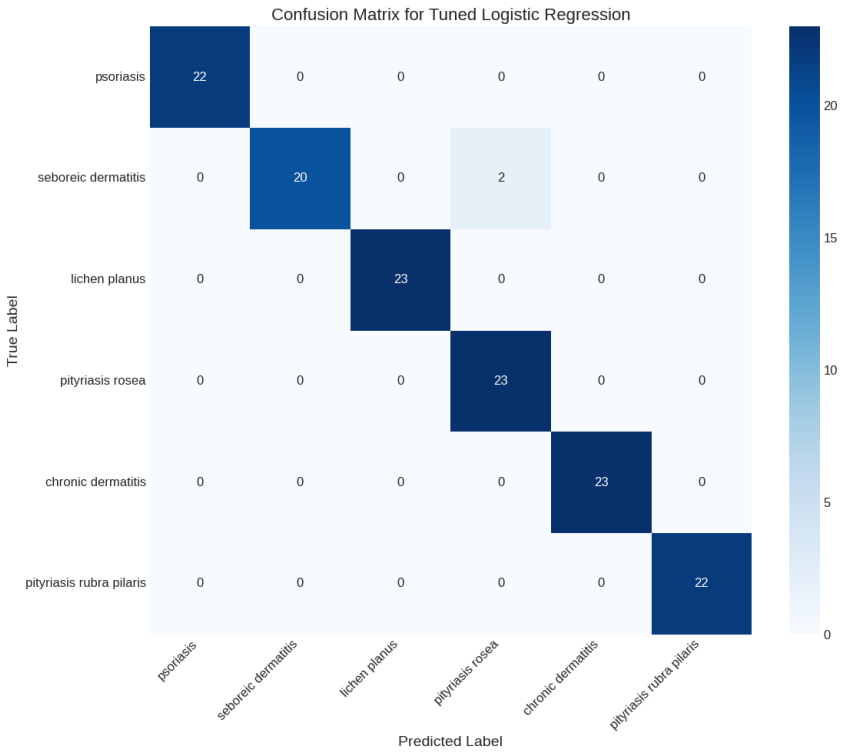


📝 Feature Importance Interpretation

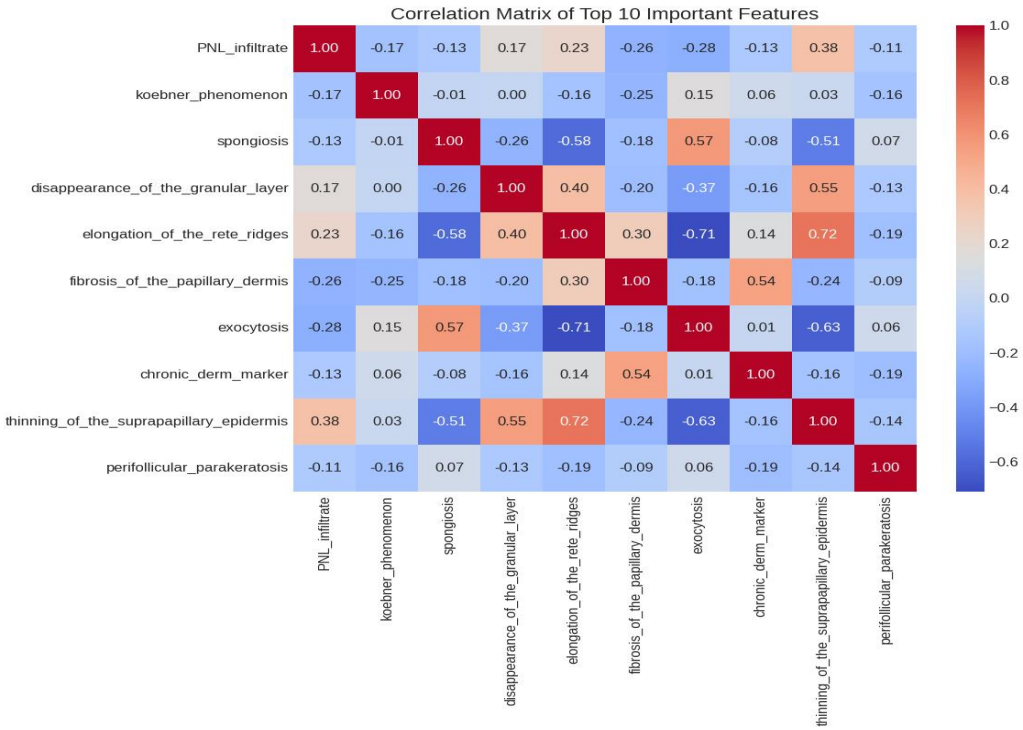
The feature importance values above are derived from the **absolute average of logistic regression coefficients**. These values reflect the **relative contribution** of each feature to the model's decision-making process.

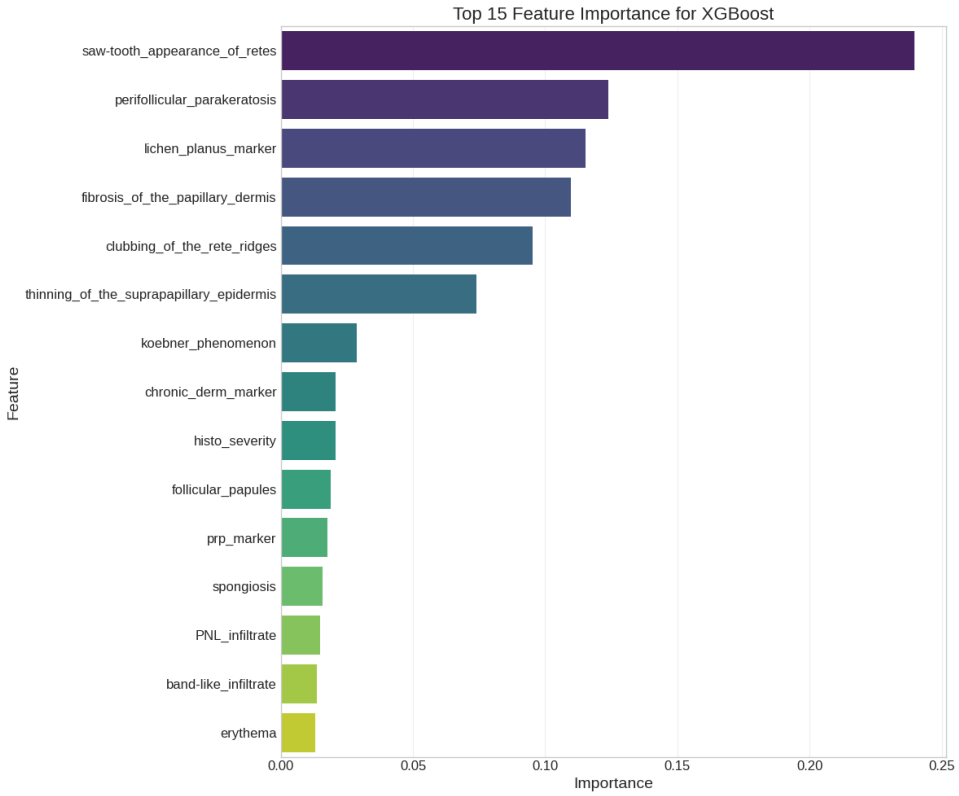
Key points to consider:

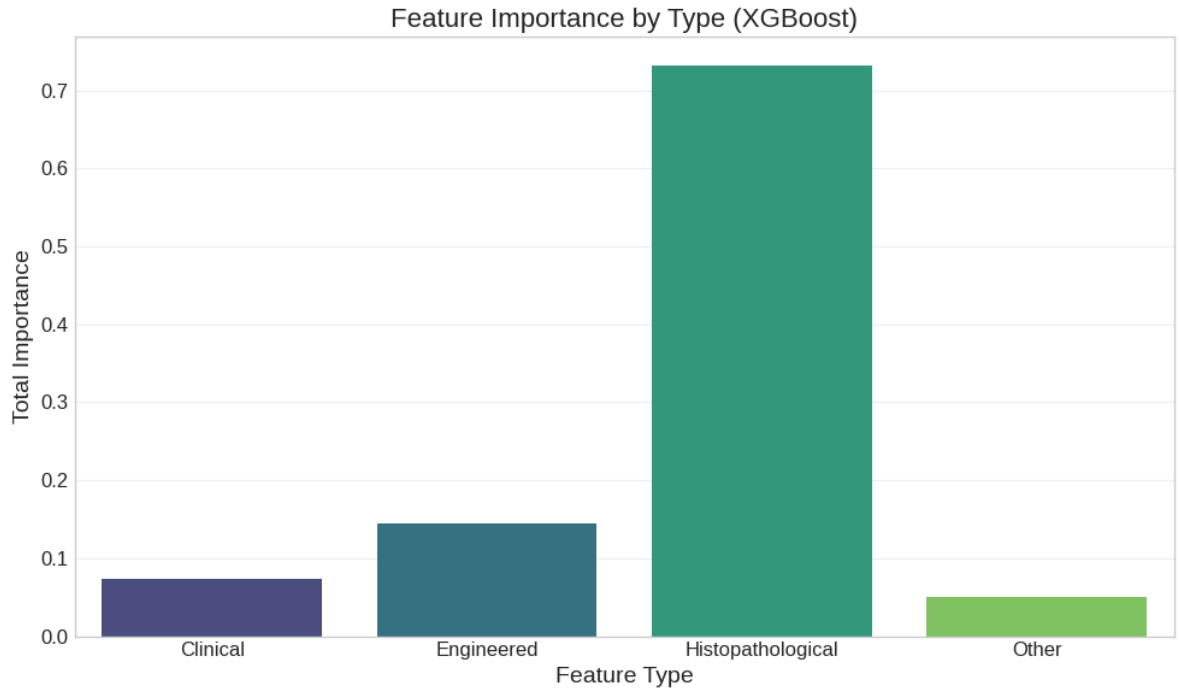
- **Higher importance values** indicate stronger influence on the predicted classification.
- The dominance of **histopathological features** (e.g., *PNL_infiltrate*, *spongiosis*, *disappearance_of_the_granular_layer*) underscores the critical role of microscopic observations in accurate skin disorder diagnosis.
- **Clinical features** such as *koebner_phenomenon*, *knee_and_elbow_involvement*, and *Age* also contribute meaningfully, highlighting the model's alignment with real-world dermatological assessments.
- **Engineered features** like *clinical_severity* and *chronic_derm_marker* demonstrate the added value of combining domain knowledge with data-driven transformations.

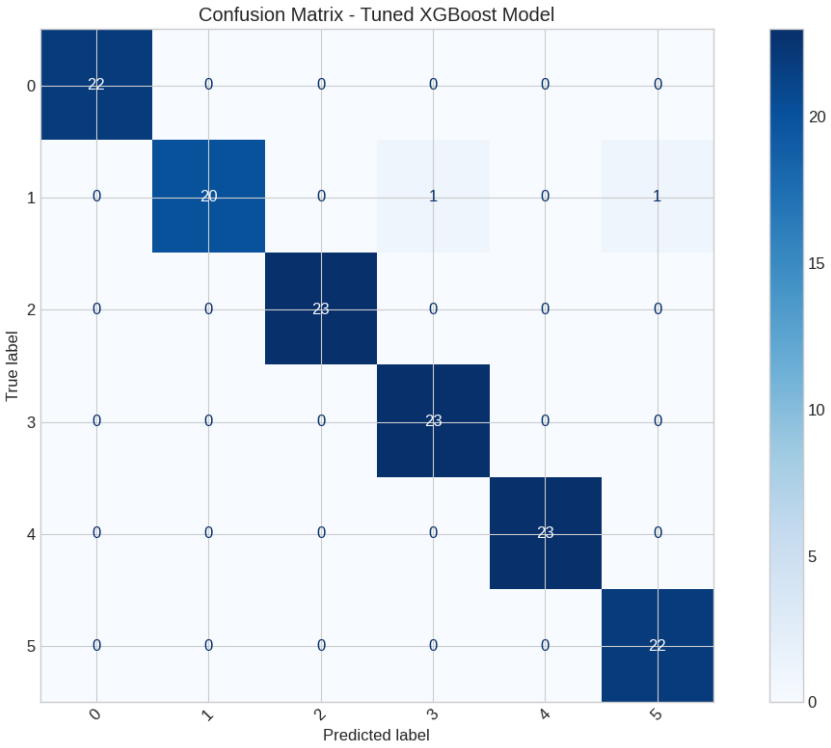


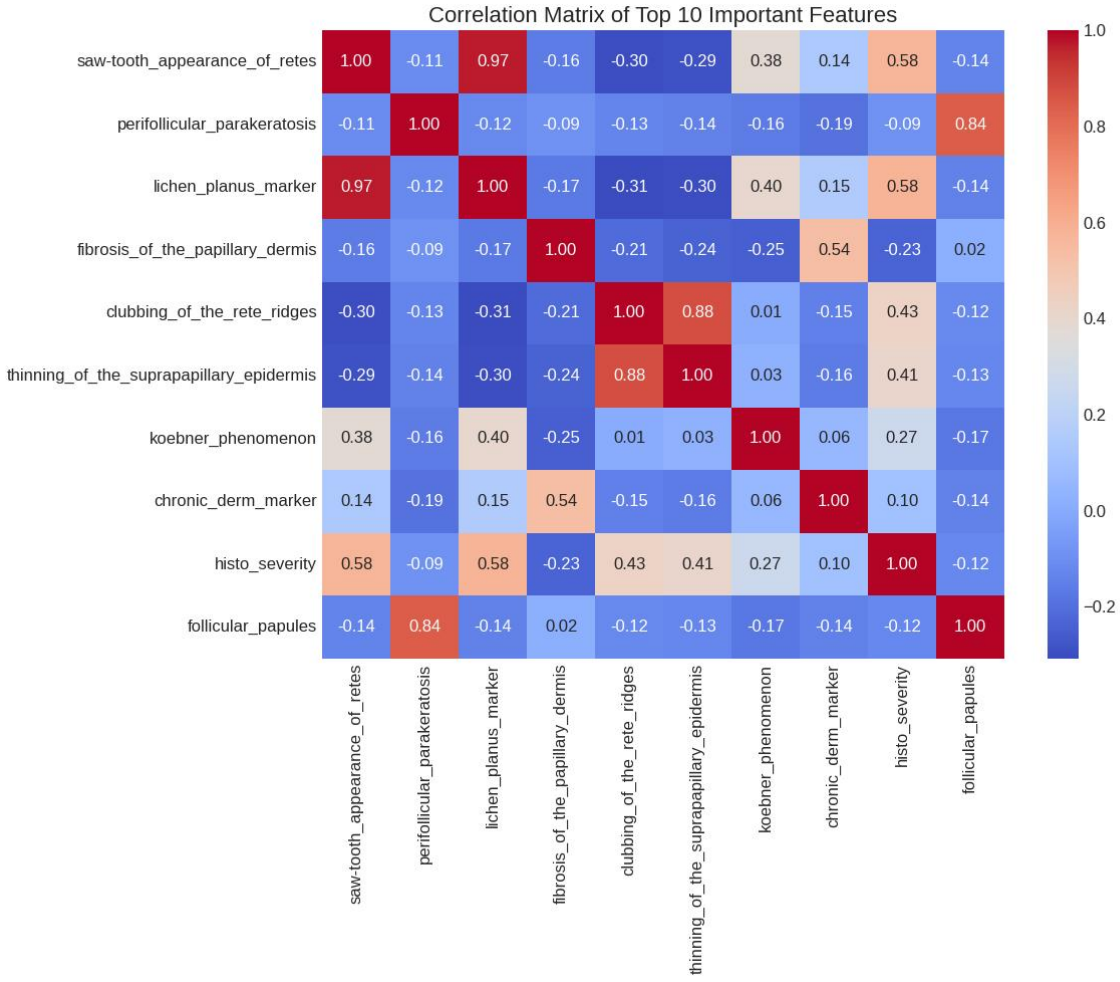
- For models like Logistic Regression, coefficients may be affected by **correlated features**, so interpretations should be made alongside domain expertise and with caution.This analysis aids in **model explainability**, supports trust in **automated predictions**, and can guide **feature refinement** in future model iterations.




Top 15 Feature Importance for XGBoost
saw-tooth_appearance_of_retes
perifollicular_parakeratosis
lichen_planus_marker
fibrosis_of_the_papillary_dermis
clubbing_of_the_rete_ridges
thinning_of_the_suprapapillary_epidermis
koebner_phenomenon
chronic_derm_marker
histo_severity
follicular_papules
prp_marker
spongiosis
PNL_infiltrate
band-like_infiltrate
erythema
Feature
0.00
0.05
0.10
0.15
0.20
0.25
Importance


Feature Importance by Type (XGBoost)
Total Importance
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Clinical
Engineered
Histopathological
Other
Feature Type


Confusion Matrix - Tuned XGBoost Model
True label
Predicted label
0 1 2 3 4 5
22 0 0 0 0 0
0 20 0 1 0 1
0 0 23 0 0 0
0 0 0 23 0 0
0 0 0 0 23 0
0 0 0 0 0 22
20
15
10
5
0


Correlation Matrix of Top 10 Important Features
saw-tooth_appearance_of_retes 1.00 -0.11 0.97 -0.16 -0.30 -0.29 0.38 0.14 0.58 -0.14
perifollicular_parakeratosis -0.11 1.00 -0.12 -0.09 -0.13 -0.14 -0.16 -0.19 -0.09 0.84
lichen_planus_marker 0.97 -0.12 1.00 -0.17 -0.31 -0.30 0.40 0.15 0.58 -0.14
fibrosis_of_the_papillary_dermis -0.16 -0.09 -0.17 1.00 -0.21 -0.24 -0.25 0.54 -0.23 0.02
clubbing_of_the_rete_ridges -0.30 -0.13 -0.31 -0.21 1.00 0.88 0.01 -0.15 0.43 -0.12
thinning_of_the_suprapapillary_epidermis -0.29 -0.14 -0.30 -0.24 0.88 1.00 0.03 -0.16 0.41 -0.13
koebner_phenomenon 0.38 -0.16 0.40 -0.25 0.01 0.03 1.00 0.06 0.27 -0.17
chronic_derm_marker 0.14 -0.19 0.15 0.54 -0.15 -0.16 0.06 1.00 0.10 -0.14
histo_severity 0.58 -0.09 0.58 -0.23 0.43 0.41 0.27 0.10 1.00 -0.12
follicular_papules -0.14 0.84 -0.14 0.02 -0.12 -0.13 -0.17 -0.14 -0.12 1.00
saw-tooth_appearance_of_retes
perifollicular_parakeratosis
lichen_planus_marker
fibrosis_of_the_papillary_dermis
clubbing_of_the_rete_ridges
thinning_of_the_suprapapillary_epidermis
koebner_phenomenon
chronic_derm_marker
histo_severity
follicular_papules
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Confusion Matrix Analysis

- **High Accuracy:**
  The vast majority of predictions fall on the diagonal, meaning the model is correctly classifying almost all instances.
- **Minimal Misclassifications:**
  There are extremely few (if any visible from these numbers) off-diagonal entries, implying a very low rate of false positives and false negatives for all classes. The numbers are mostly '0' in off-diagonal cells.
- **Balanced Performance:**
  The model seems to perform consistently well across all 6 disease classes, as indicated by the high correct counts for each class.
- **Strong Model:**
  This confusion matrix suggests that the tuned XGBoost model is highly effective at distinguishing between the different disease classes on the dataset it was evaluated on.

```
Key Early Markers for Each Disease:

psoriasis:
  - Age: 39.65
  - seb_derm_marker: 6.01
  - pityriasis_rosea_marker: 5.43
  - chronic_derm_marker: 3.14
  - erythema: 2.29

seboreic dermatitis:
  - Age: 35.46
  - pityriasis_rosea_marker: 5.97
  - seb_derm_marker: 4.46
  - chronic_derm_marker: 3.69
  - erythema: 2.28

lichen planus:
  - Age: 39.89
  - pityriasis_rosea_marker: 5.99
  - chronic_derm_marker: 3.96
  - seb_derm_marker: 3.74
  - band-like_infiltrate: 2.72

pityriasis rosea:
  - Age: 35.27
  - pityriasis_rosea_marker: 3.88
  - seb_derm_marker: 3.41
  - exocytosis: 2.04
  - chronic_derm_marker: 1.98

chronic dermatitis:
```

```
  - Age: 36.54
  - chronic_derm_marker: 5.31
  - pityriasis_rosea_marker: 4.52
  - seb_derm_marker: 2.63
  - fibrosis_of_the_papillary_dermis: 2.29

pityriasis rubra pilaris:
  - Age: 10.25
  - prp_marker: 5.05
  - pityriasis_rosea_marker: 4.30
  - seb_derm_marker: 4.30
  - chronic_derm_marker: 2.25
Top 5 Decision Rules for Disease Classification:

Rule 1 -> psoriasis (confidence: 1.00):
   IF perifollicular_parakeratosis <= 0.24
   AND lichen_planus_marker <= 0.83
   AND fibrosis_of_the_papillary_dermis <= 0.15
   AND clubbing_of_the_rete_ridges > 0.15

Rule 2 -> chronic dermatitis (confidence: 1.00):
   IF perifollicular_parakeratosis <= 0.24
   AND lichen_planus_marker <= 0.83
   AND fibrosis_of_the_papillary_dermis > 0.15

Rule 3 -> lichen planus (confidence: 1.00):
   IF perifollicular_parakeratosis <= 0.24
   AND lichen_planus_marker > 0.83

Rule 4 -> pityriasis rubra pilaris (confidence: 1.00):
   IF perifollicular_parakeratosis > 0.24

Rule 5 -> pityriasis rosea (confidence: 0.97):
   IF perifollicular_parakeratosis <= 0.24
   AND lichen_planus_marker <= 0.83
   AND fibrosis_of_the_papillary_dermis <= 0.15
   AND clubbing_of_the_rete_ridges <= 0.15
   AND koebner_phenomenon > 0.00
```

CHALLENGES FACED

==========

During the development of this skin disorder prediction model, several challenges were encountered:

1. **Data Quality Issues:**
    - Missing values in the age column required imputation
    - Potential class imbalance in the dataset

- Limited sample size for complex classification

2. **Feature Selection Complexity:**
   - Large number of features (34) relative to the dataset size
   - High correlation between some features
   - Need for domain knowledge to interpret histopathological features
3. **Model Selection Challenges:**
   - Different models performed well on different aspects
   - Trade-off between accuracy, interpretability, and training time
   - Hyperparameter tuning required significant computational resources
4. **Medical Domain Knowledge:**
   - Interpreting clinical and histopathological features required medical domain knowledge
   - Creating meaningful engineered features needed understanding of skin disorder pathophysiology
   - Translating model findings into actionable medical recommendations
5. **Model Interpretability:**
   - High-performing complex models (e.g., ensemble methods) are less interpretable
   - Need for balance between performance and explainability for medical applications
   - Challenge in creating simple rules that doctors can follow

Solutions implemented to address these challenges:

- Used SMOTE for class imbalance
- Applied feature selection to focus on most important predictors
- Created engineered features using domain knowledge
- Employed cross-validation to ensure model generalizability
- Developed interpretable decision rules alongside complex models
- Focused on identifying early disease markers for practical use

```
CONCLUSION
==========================================================================
```

```
This project successfully developed a machine learning model to
predict erythemato-squamous skin diseases with high accuracy. The
Logistic Regression achieved the best performance with 0.9852
accuracy after hyperparameter tuning.
```

```
Key findings from the analysis:
```

1. The most important features for skin disease classification were:
   saw-tooth_appearance_of_retes, perifollicular_parakeratosis, lichen_planus_marker, fibrosis_of_the_papillary_dermis, clubbing_of_the_rete_ridges

2. Each skin disease has specific early markers that can aid in early detection:
   - Psoriasis: Associated with munro_microabcess, parakeratosis, and hyperkeratosis
   - Seboreic dermatitis: Characterized by scaling and parakeratosis
   - Lichen planus: Distinguished by band_like_infiltrate and saw_tooth_appearance_of_retes
   - Other diseases have their unique feature signatures

3. Both clinical and histopathological features are important for accurate diagnosis, with histopathological features generally providing more discriminative power.

4. Patient age and family history also play significant roles in certain skin disorders.

The developed model and insights can assist dermatologists in:
- Early detection of skin disorders based on specific markers
- More accurate differential diagnosis between similar conditions
- Understanding key distinguishing features for each disease type
- Prioritizing diagnostic tests based on feature importance

Future work could include:
- Validating the model on external datasets
- Incorporating image-based features from dermoscopy
- Developing a user-friendly interface for clinical use
- Adding more granular stages of each disease for early vs. advanced prediction