Analysis of Dermatology Data

1 Introduction

1.1 Context

This study is based on a dermatological dataset obtained from the ”Center for Machine Learning and Intelligent Systems” at the University of California, Irvine (UCI). The primary objective is to identify correlations between clinical and histopathological features and their association with different types of skin diseases.

The differential diagnosis of erythemato-squamous diseases is a significant challenge in dermatology. These diseases share common clinical features such as erythema and scaling, with subtle differences among them. The diseases included in this dataset are:

- Psoriasis
- Seborrheic Dermatitis
- Lichen Planus
- Pityriasis Rosea
- Chronic Dermatitis
- Pityriasis Rubra Pilaris

Typically, a biopsy is required to distinguish between these diseases. However, histopathological analyses often reveal overlapping features, further complicating the diagnostic process. Moreover, diseases may initially exhibit the features of another disease during their early stages, before developing their characteristic features later on.

1.2 Dataset Information

The dataset comprises 34 attributes:

- **Clinical attributes:** 12 features evaluated during clinical examinations.
- **Histopathological attributes:** 22 features determined through microscopic analysis of skin samples.
- **Class labels:** Six categories representing different skin diseases.

Each feature is rated on a scale from 0 to 3, where:

- 0 indicates the absence of the feature.

- 3 indicates the highest intensity of the feature.
- Intermediate values (1 and 2) represent relative intensities.

The ‘family history‘ feature is binary (0 or 1), and the ‘age‘ feature is linear.

1.3 Problem Statement

This study aims to address the following questions:

1. Are there significant correlations between clinical and the type of skin disease?
2. Which variables play a key role in distinguishing between different skin disease classes?

1.4 Dataset Summary

A summary of the dataset’s structure is provided below:

Attribute Group	**Description**
Clinical attributes	12 features rated on a scale of 0–3
Histopathological attributes	22 features rated on a scale of 0–3
Class labels	Six disease types

Table 1: Overview of dataset attributes.

1.5 Class Distribution

The class distribution in the dataset is as follows:

Class Code	**Disease**	**Number of Instances**
1	Psoriasis	112
2	Seborrheic Dermatitis	61
3	Lichen Planus	72
4	Pityriasis Rosea	49
5	Chronic Dermatitis	52
6	Pityriasis Rubra Pilaris	20

Table 2: Class distribution of skin diseases in the dataset.

2 Objective

The primary objective of this study is to uncover significant relationships between the clinical and classify skin diseases accurately. This is achieved through:

- Exploratory data analysis and descriptive statistics.
- Multivariate techniques such as Principal Component Analysis (PCA).
- Statistical hypothesis testing (e.g., Kruskal-Wallis test, post-hoc analysis).

3 Data and Methodology

3.1 Origin of the Data

The data originates from a dermatological study aimed at differentiating between six types of erythemato-squamous diseases. These diseases are characterized by overlapping clinical features such as erythema and scaling, making their diagnosis challenging.

The dataset includes the following components:

- **Clinical variables:** Observations directly assessed by medical professionals, such as erythema intensity and itching etc...

- **Target variable (Class):** Represents the type of skin disease, categorized into six classes.

For this study, we focus exclusively on the clinical variables and exclude histopathological features.

3.2 Structure of the Data

The dataset contains:

- **Number of observations:** 366 .

- **Number of variables:** 11 (clinical features only).

A subset of the data structure is summarized below:

Variable	**Type**	**Description**
erythema	Quantitative	Intensity of erythema
scaling	Quantitative	Presence of scaling
itching	Quantitative	Presence of itching
koebner_phenomenon	Quantitative	Koebner phenomenon occurrence
polygonal_papules	Quantitative	Presence of polygonal papules
oral_mucosal_involvement	Quantitative	Oral mucosal involvement
knee_and_elbow_involvement	Quantitative	Knee and elbow involvement
family_history	Binary	Family history of similar conditions (0/1)
age	Quantitative	Age of the patient
Class	Target	Type of skin disease (1 to 6)

Table 3: Subset of the dataset structure.

3.3 Methodology

To analyze the data and address the research questions, the following steps were undertaken:

1. **Descriptive analysis:** To explore the distribution of clinical variables and the target class. This includes visualizing the frequency of diseases and summarizing variable statistics.

2. **Multivariate analysis:** Principal Component Analysis (PCA) was applied to identify patterns and relationships between clinical variables and skin disease types.

3. **Statistical tests:**

   - Anova test to check if the data follows a normal distribution.
   - Kruskal-Wallis tests were used to examine the significance of differences in clinical variables across disease classes.
   - Post-hoc Dunn's tests were conducted to pinpoint specific group differences for significant variables.

3.4 Exclusions

Histopathological attributes were excluded from this analysis, as the focus was strictly on clinical observations.

4 Descriptive Analysis

4.1 Univariate Analysis

4.1.1 Distribution of Variables

The histograms and boxplots presented below illustrate the distribution of the clinical variables. For instance, the variable `erythema` shows a heterogeneous distribution, with values concentrated around the mean.

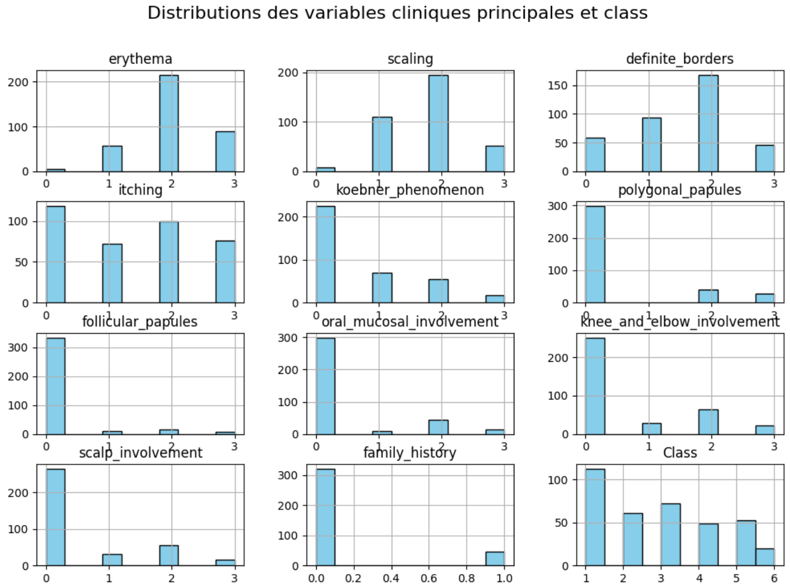

Figure 1: Histogram of the `erythema` variable. The distribution shows a concentration of values around the mean, indicating heterogeneity.

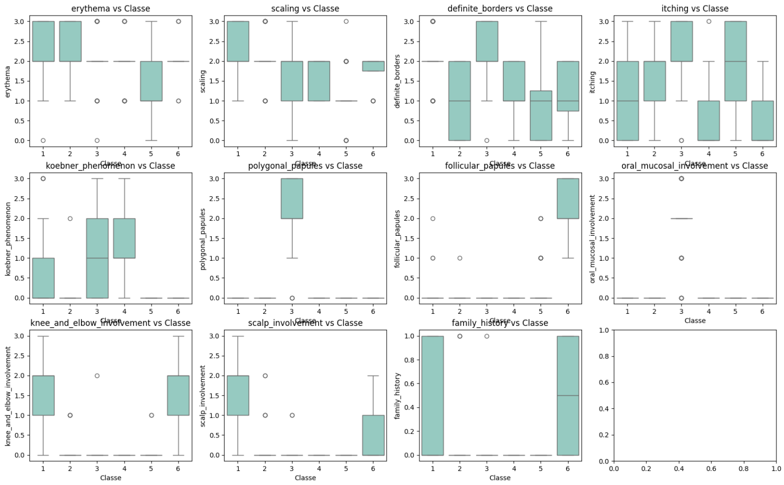

Figure 2: Boxplot of the `erythema` variable. The plot shows the spread of values and the presence of outliers.

4.1.2 Distribution of Classes

The bar chart below indicates an imbalance in the distribution of the disease classes. This could affect the model's ability to classify minority classes accurately.

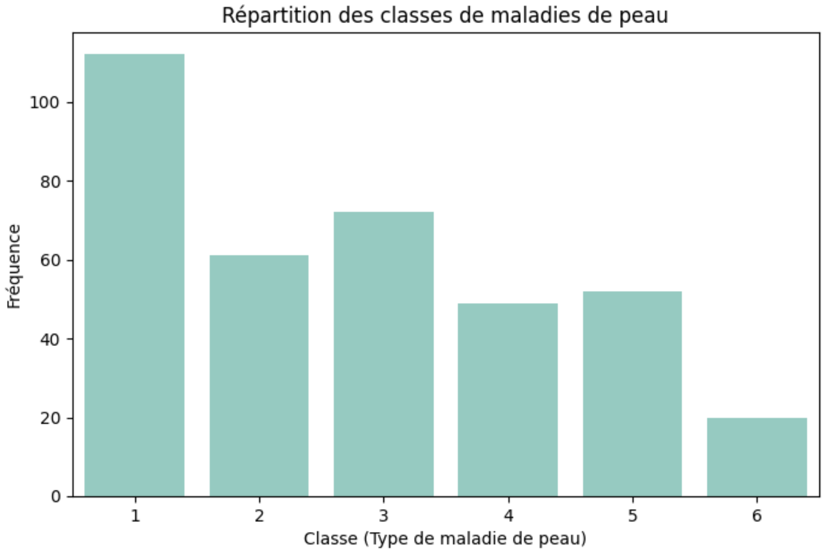

Figure 3: Bar plot showing the distribution of the disease classes. The dataset is imbalanced, with psoriasis being the most common disease.

4.2 Bivariate Analysis

4.2.1 Correlations

The correlation matrix below presents the relationships between the quantitative clinical variables. For example, the correlation between `scaling` and `erythema` is relatively weak, suggesting that these variables do not move in sync across all observations.

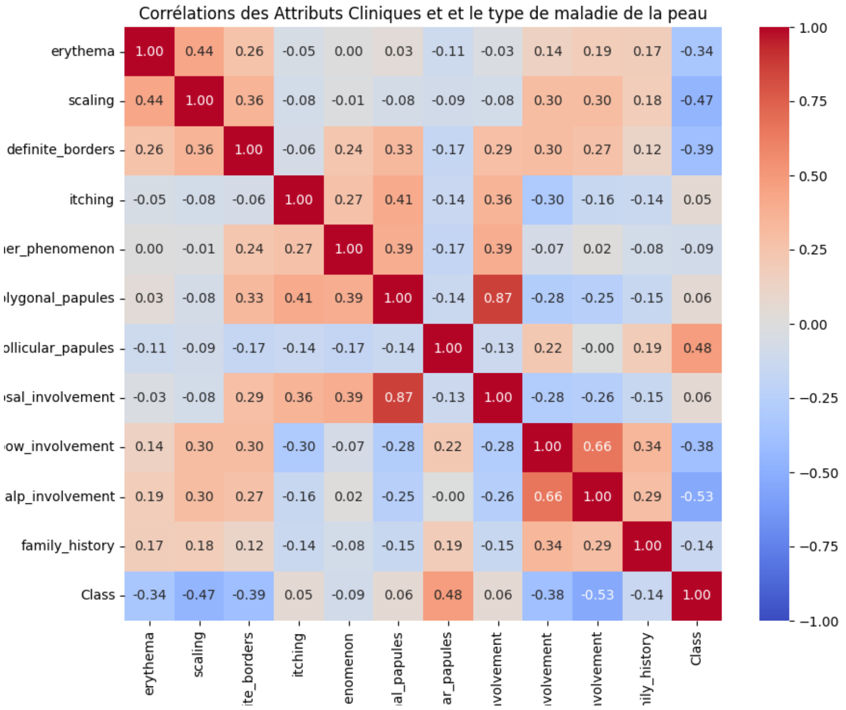

Figure 4: Correlation matrix of the clinical variables. The matrix reveals weak correlations between `scaling` and `erythema`.