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 histopathological features 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 histopathological features 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).