Psoriasis Detection Based on Clinical and Histopathological Features

Louis Carron

May 2024

1 Introduction

Context of the study :

A medical diagnosis can be a difficult task when several diseases share the same clinical features with only small differences. This is the case of the differential diagnosis of erythemato-squamous diseases. In particular, Psoriasis is a chronic skin disease characterized by the appearance of red patches covered with scales. This disease often comes with patient that are genetically disposed. Accurate and early detection is crucial for effective treatment. Diagnostic relies on clinical examinations and histopathological analysis.

The goal of our study is to answer the following questions: Is it possible to develop a reliable classification model to detect psoriasis cases based on clinical and histopathological features ? Moreover, we want to identify the relevant factors that contribute to the prediction of psoriasis.

Dataset Presentation :

The database was first introduced in a paper made by two Computer Scientist and one Dermatologist in 1998 [1] to show the efficiency of a classification algorithm, called VFI5. It contains 366 instances for 34 variables split into two different categories : Clinical and Histopathological Attributes. Every attributes apart age and family history is given by a categorical value between $[|0, 3|]$ referring to a signification degree of matter. Zero indicates the absence of a symptom and three the largest presence as possible. Family History takes the value one or zero according to the observation of the disease in the family or not. More precisely, the observed variables are defined in the appendix to this study (last Page). We will train a binary classification model with label 1 for the psoriasis cases and 0 for the others.

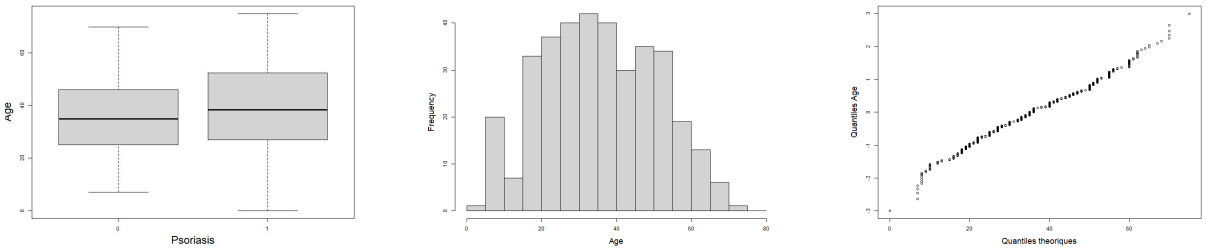
2 Exploratory Data Analysis

Before fitting a model, we first process and visualise the data. The goal is to clean raw observations, gain insights about the distributions and identify correlations between predictors.

Data Pre-Processing

Missing value

The dataset contains 8 missing values for patient's age. According to the following plot, the age distribution is symmetric, the mean is 36.3 and the median is equal to 35.



(a) Box Plot (b) Histogram (c) Normal QQ

Figure 1: Raw Age Distribution Plot

The mean is a measure that is highly sensitive to outliers. The boxplot shows us that there are no potential outliers in the distribution. The missing values proportion is 0.02, it only represents a small portion of the whole dataframe. The correlation coefficient for the qqplot is 0.993. We apply a Shapiro-Wilk normality test and the p-value is 0.002. At a signification level of 0.01, we do not have enough information to reject the Null Hypothesis that states the normality of our age distribution.

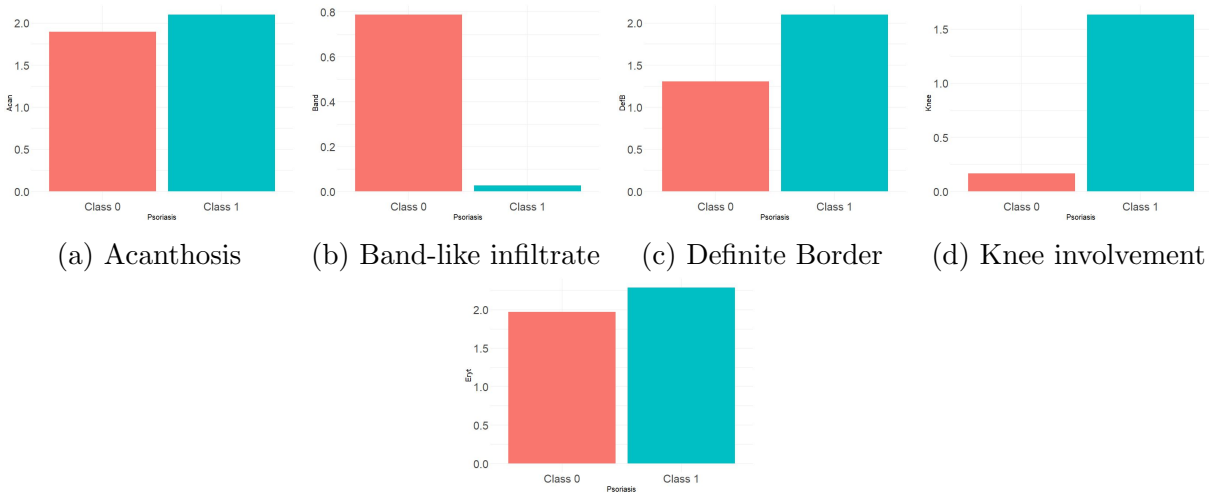
In conclusion, without loosing any generalities according to the previous statement, the missing values for the age distribution will be replaced by the mean.

Visualization

We visualize different variables mean value according to the proportion of Psoriasis cases in Figure 2. We can notice that for the Band-like infiltrate, the proportion of class 1 features are very few, less than 0.2. Same for the Knee Involvement, with less than 0.25 class 0 features. This can be the sign of outliers presence inside our dataset.

To identify any outliers in the dataset, we compute inside all the variables and for each class, the apparition frequency of the categorical levels 0,1,2,3. If the apparition frequency of the level is less than a threshold value fixed to 0.05, the observation can be considered as an outlier.

Thus, we ensure that our data are correctly distributed by excluding extreme values that can potentially influence our classification model.

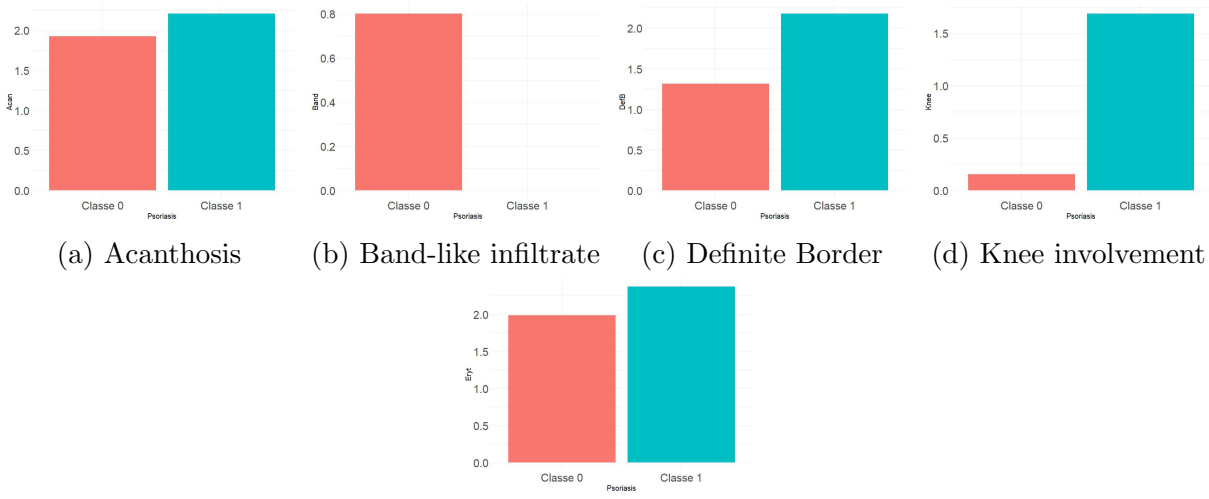


(a) Acanthosis (b) Band-like infiltrate (c) Definite Border (d) Knee involvement

(e) Erythema

Figure 2: Proportion of Psoriasis case (1, blue) according to the mean value of the categorical levels for different Variables : Acanthosis, Band-Like inflitrate, Definite Border, Knee involvement, Erythema. Before Processing outliers.

After removing the observations with extrem values, the final variables distribution are plotted in the following Figure 3 :



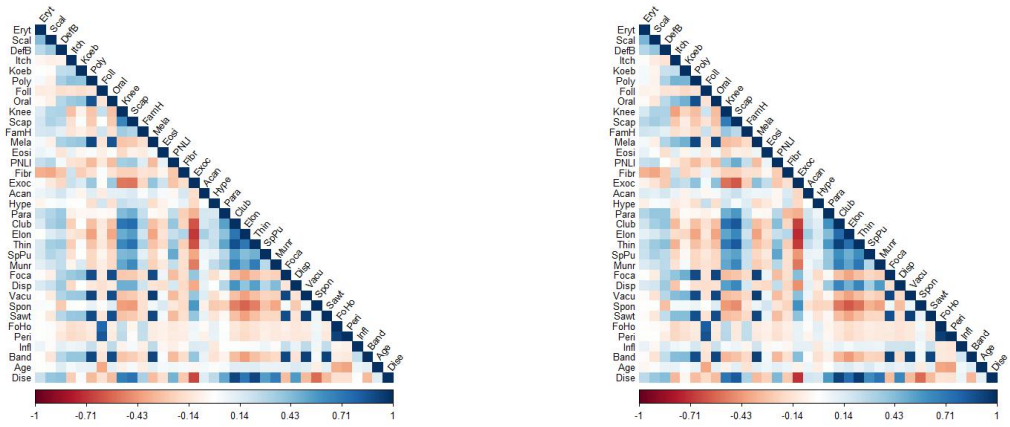
(a) Acanthosis (b) Band-like infiltrate (c) Definite Border (d) Knee involvement

(e) Erythema

Figure 3: Proportion of Psoriasis case (1, blue) according to the mean value of the categorical levels for different Variables : Acanthosis, Band-Like inflitrate, Definite Border, Knee involvement, Erythema. After Processing outliers.

Data Exploration

To identify linear relationships between pairs of variables, we plot the correlation matrix with the following heat map :



(a) With Raw Data

(b) After Data processing

Figure 4: Correlation Matrix : Red for high negative correlation, White for zero, Blue for high positive correlation

We can notice that the Age variable seems to be uncorrelated with most of the other variables in our study. The Correlation Matrix underline multiple cases of high multicolinearity. For instance Polygonal Papules (Poly) is highly positively correlated with Oral mucosal, Melanin incontinence, focal hypergranulosis, Vacuolisation, Saw-tooth appearance of retes and band-like infiltrate. These combinations of linear relationships create difficulties in determining coefficients for a classification model. They increase the variability of estimations by preventing the model from capturing the individual effect of each variable.

Model Preparation

Before fitting our model to numerical and categorical values, we encode them using the factor() function in R. We encode categorical data to convert it into a numerical format that the algorithm can process and understand. Because categories represent the degree of apparition for a symptom, we need to keep an order relationship. Using a One-hot encoding instead of factors is then less appropriate in our case of study because it guaranties an independence between each level. On the contrary, we want to keep the relation order of the categories.

3 Logistic Regression

3.1 Theoretical

Model Presentation

The logistic regression is a Generalized Linear Model ideal for classification tasks. Let $\mathbf{X} = [1, X_1, X_2, ..., X_N]^T$ be the observed variables with a term of bias and $\boldsymbol{\beta} = [\beta_0, \beta_1 ..., \beta_N]^T$ the parameters, the model is given by the following equation :

$$logit(p(\mathbf{X}) = \log(\frac{p(\mathbf{X})}{1 - p(\mathbf{X})}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_N X_N = \boldsymbol{\beta}^T \mathbf{X} \quad (1)$$

The output prediction of the model is $p \in [0, 1]$ that represents the probability of having a label 1 according to the input $p(\mathbf{X}) = \mathbb{E}[Y = 1 \mid \mathbf{X}] = \mathbb{P}[Y = 1 \mid \mathbf{X}]$ The logit function transform a probability in [0, 1] to a continuous value in $-\infty$ to $+\infty$.

The output with the inverse logit function :

$$p(\mathbf{X}) = \sigma(\boldsymbol{\beta}^T \mathbf{X}) = \frac{1}{1 + \exp(-\boldsymbol{\beta}^T \mathbf{X})}$$

Equation (1) can also be write with the odds of the probability $p(\mathbf{X})$:

$$odds(p(\mathbf{X})) = \frac{p(\mathbf{X})}{1 - p(\mathbf{X})} = \exp(\beta_0) \prod_{i=1}^{N} \exp(\beta_i X_i)$$

The odds represent the likelihood that the event is happening. They can be understood as the ratio of the number of times the event occurs to the number of times it does not occur. An increase in the event odds corresponds to an increase in the event probability. And reciprocally.

In equation (1), we can see that the odds of the event is exponentially related to the explanatory variables. For a single explanatory variable X_i and a one unit increase, we can define its odd ratio as $\exp(\beta_i)$. If the OR is equal to 1, it means that a change of the variable X, has no impact on the event odds. If $OR > 1$, it means that the variable contribute to increase the event odds. Finally $OR < 1$ results to a diminutive contribution on the event odds according to a raise of the variable.

Model Fitting

Logistic regression, as a probabilistic model, estimates its parameters $\boldsymbol{\beta}$ using the maximum likelihood method. Having a given label $Y_i = y_i$ for the (i)-th subject can be seen as a Bernoulli distribution, with a probability of success $p(\mathbf{X})$. Thus, we can write :

$$\mathbb{P}[Y_i = y_i] = p(\mathbf{X}^{(i)})^{y_i} (1 - p(\mathbf{X}^{(i)}))^{1 - y_i}$$

Thereby, the likelihood function can be written as:

$$L(\mathbf{X}^{(1)}, ..., \mathbf{X}^{(n)}, \boldsymbol{\beta}) = \prod_{i=1}^{n} p(\mathbf{X}^{(i)})^{y_i} (1 - p(\mathbf{X}^{(i)}))^{1 - y_i}$$

Maximizing $L(\boldsymbol{\beta})$ is equivalent to maximizing the log-likelihood function:

$$I(\mathbf{X}^{(1)}, ..., \mathbf{X}^{(n)}, \boldsymbol{\beta}) = \log(L(\mathbf{X}^{(1)}, ..., \mathbf{X}^{(n)}, \beta)) = \sum_{i=1}^{n} \left(y_i \log(p(\mathbf{X}^{(i)})) + (1 - y_i) \log(1 - p(\mathbf{X}^{(i)})) \right)$$

This last equation is non linear, we do not have an analytical solution for the parameters $\boldsymbol{\beta}$. To solve this problem, we use iteratively algorithms like Newton Raphson, gradient descent.

3.2 Application to our study

First Approach :

We fit a logistic regression model with all the 34 variables in our dataset. This base model shows a perfect separation between features making the Wald statistic unusable. The MLE process in logistic returns parameter estimates that tends to very high or low values. A huge standard error for estimated coefficients shows a high variability and instability of the model.

To avoid this problem of high separation, we remove some of the bias by adding a penalization term into the likelihood function. To do so, we use the Firth glm method that stabilizes our parameter estimations and ensure the convergence of the logistic regression model. After fitting the glm, we obtain the following result :

Coefficient	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	-0.50	4.74	-0.11	0.92
Eryt1	-0.65	2.69	-0.24	0.81
Eryt2	-0.76	2.80	-0.27	0.79
Eryt3	-0.70	2.91	-0.24	0.81
Scal1	-0.38	2.02	-0.19	0.85
Scal2	-0.40	2.25	-0.18	0.86
Scal3	-0.31	2.35	-0.13	0.89
DefB1	0.03	1.02	0.03	0.97
DefB2	0.10	1.08	0.09	0.93
DefB3	0.08	1.38	0.06	0.95

Table 1: Truncate summary of the full logistic model

The resulting residual deviance is 62.35. To understand the equilibrium between complexity and adjustement of our model we use the Akaike Information Criteria $AIC = 2k - 2ln(L)$ with k the number of parameters and L the maximum of the likelihood function. The value is used to detect cases of overfit with complex model and for the comparison or selection of models. Here AIC = 236.35. The overall p-value for each coefficient is too high to infer about the signification of a variable on the disease prediction. We have to simplify our model by reducing this large number of variables in order to highlight the most important predictors.

Variable Selection :

Multicolinearity

To ensure a good fit of our logistic regression model, we need to ensure the absence of multicollinearity in our variables. We saw during the exploration that variables s.a Poly can be multi-correlated to a subset of variables. The presence of such variables inside our model increase complexity and reduce the reliability of statistical significance of independent variables. It increases variability in the coefficient estimation. A little change from one can highly impact another. Thus, it reduces our model accuracy and its capability to converge well. The Variance Inflation Factor (VIF) measures this amount of multicolinearity for a variable inside a model.

The formula for VIF is given by : $VIF_i = \frac{1}{1-R_i^2}$ with R_i^2 the unadjusted coefficient of determination for regressing the ith variable on the others. Then, if a variable is highly correlated to others, the VIF tends to infinity. Otherwise it tends to 1. The vif() function in the car R package gives us the Generalized VIF (for categorical features) and an adjusted GVIF that takes into account the number of degrees of freedom for variables. We choose a threshold value for the adjusted GVIF of 2. This is equivalent to choose a maximum value for $R^2 = 0.75$.

Variable	GVIF	Df	$\text{GVIF}^{1/(2*Df)}$
Thin	932.08	3	3.13
Band	773.05	3	3.03
Club	417.69	3	2.73
Elon	282.56	3	2.56
Oral	258.69	3	2.52

Table 2: Top 5 Variables with Highest GVIF in our full logistic Model

Algorithmically, step-by-step, we remove one-by-one the variable with the highest GVIF and recompute the logistic regression model until all the GVIF adjusted are all less than 2.

We ended up by removing the following variables to our original full model : Band, Thin, Club, Oral, Poly, Foca, Peri, Vacu. As the Figure 4 shows, all these variables were highly correlated to each others, this filtering process ensures that we only the most significant variables to describe them. The model returns a residual deviance of 54.31 and an AIC of 194.31.

Variable	GVIF	Df	$\text{GVIF}^{1/(2*Df)}$
Mela	55.09	3	1.95
FoHo	38.58	3	1.84
Sawt	43.61	3	1.88
Foll	46.06	3	1.89
Knee	34.81	3	1.81

Table 3: Top 5 Variables with Highest GVIF in the filtered logistic Model

Lasso Selection

To simplify our logistic model and keep only the most significant predictors, we use a Lasso method (Least Absolute Shrinkage and Selection Operator). Lasso add a regularization term inside the log-likelihood formula equal to $\lambda \sum_i |\beta_i|$. The λ controls the intensity of the penalty on the sum of the coefficients. The choice of λ is crucial and made by cross-validation with the cv.glmnet() R function. The Lasso method provides our model from over fitting with a large number of variables by making a selection. It removes the less significant variables and handle multicolinearity.

For the Psoriasis detection, the Lasso selection returns us the following variable : Scap, Fibr, Exoc and Munr.

Variable	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	-1.78	1.26	-1.42	0.157
Scap1	2.52	1.42	1.78	0.075
Scap2	4.50	1.62	2.78	0.006 **
Scap3	2.97	2.14	1.39	0.165
Fibr1	0.72	2.30	0.31	0.755
Fibr2	-0.35	1.81	-0.19	0.846
Fibr3	-0.67	1.83	-0.37	0.713
Exoc1	-1.71	1.32	-1.29	0.196
Exoc2	-4.52	1.79	-2.53	0.011 *
Exoc3	-2.56	1.71	-1.50	0.134
Munr1	4.36	1.68	2.59	0.010 **
Munr2	3.65	1.93	1.89	0.059 .
Munr3	0.78	1.95	0.40	0.689

Table 4: Summary of the Psoriasis model obtained using the Lasso selection method with variables Scap, Fibr, Exoc, and Munr. Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

The model has a residual Deviance of 12.64 and an AIC of 38. We can notice that a degree 2 of Scalp involvement is statistically significant at a level of 0.01 which suggests a strong association of this symptom with Psoriasis cases. Same finding for degree 1 Munro microabcess. Finally Exocytosis is statistically significant at a level of 0.05.

Variable	GVIF	Df	GVIF$^{1/(2*Df)}$
Scap	3.00	3	1.20
Fibr	1.91	3	1.11
Exoc	2.22	3	1.14
Munr	2.23	3	1.14

Table 5: Multicolinearity analysis for the Psoriasis detection model using variables Scap, Fibr, Exoc and Munr

3.3 Model Assessment

The logistic regression model states several hypothesis : Independent observations (for MLE), linear relation between log-odds and predictors (1), absence of perfect colinearity between variables, homoscedasticity and no outliers.

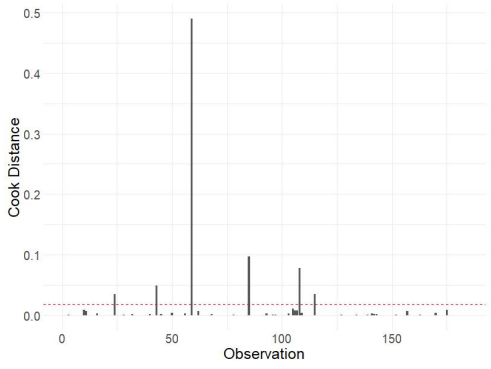
With all the adjusted GVIF under 1.2, our model does not present any sign of multicolinearity Table 6.

To identify outliers in our model prediction, we use the Cook Distance. This measure allows us to identify the influence that one observation can have on the coefficient estimation of our logistic model.

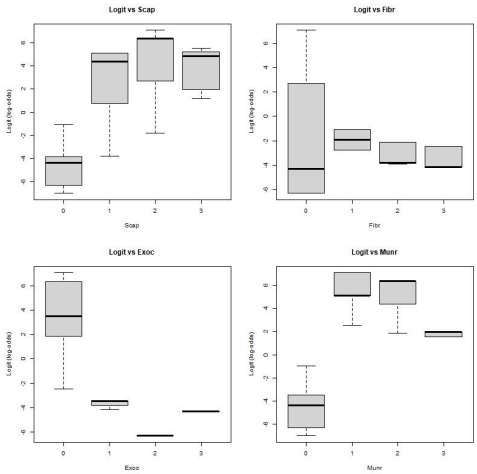
We define it s.a : $D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{kMSE}$

where $\hat{Y}_j$ is the full model prediction for the j-th observation. And $\hat{Y}_{j(i)}$ is the prediction for the j-th observation with a logistic regression model without i-th observation.

The Figure 5. (a) shows a majority of the observation under the threshold value and 6 outside. The highest one is under 0.5 that stays a reasonable value to reject the presence of outliers in our logistic model.



(a) Cook distance as a function of observations



(b) Logit as function of predictors

Figure 5: Model assessment plot, Psoriasis - Scap + Fibr + Exoc + Munr

The linearity between the logit and the categorical values of the predictors can be attested by the plot Figure 5 (b) for the Scap predictor and Fibr.

Final Model

$$\hat{\text{logit}}(p) = -1.78 + \text{Scap}\beta_1 + \text{Fibr}\beta_2 + \text{Exoc}\beta_3 + \text{Munr}\beta_4 \quad (1)$$

with $Var = [Var1, Var2, Var3]$ and $\beta_1 = [2.52, 4.50, 2.97]^T$, $\beta_2 = [0.72, -0.35, -0.67]^T$, $\beta_3 = [-1.71, -4.52, -2.56]^T$, $\beta_4 = [4.36, 3.65, 0.78]^T$

Classification Results

We split our data in a train (0.8) and test set (0.2). We make a 10-fold cross validation and obtain an accuracy rate of 95% on the test set. The adjusted accuracy kappa that underlines the robustness of the model gives us 88%.

4 Conclusion

Our logistic regression analysis has proven effective for detecting cases of Psoriasis within our dataset. The primary challenge of this study was in the difficult distinctions between symptoms of Erythemato-Squamous Skin Diseases. To solve this, we explored the relationships among these variables by constructing correlation plots and calculating the Variance Inflation Factor (VIF). This analysis revealed patterns of multicollinearity, which allowed us to gain insights about the selection of variables inside our model.

To simplify and upgrade our model, we made a step-by-step approach for variable selection and used techniques such as Lasso penalization. This process gave us the final most significant variables in the prediction of Psoriasis cases which are : Scalp involvement, Exocytosis, Fibrosis of the papillary dermis, and Munro microabcess.

Dealing with categorical variables was the potential for perfect separation, which reduce the efficiency of the model. In complex models with numerous variables, identifying and addressing multicollinearity was crucial for simplifying our analysis. To handle issues of variable separation and enhance the reliability of our conclusions, we applied the Firth method via the blglm2 package, which adjusted a penalization in our generalized linear model. This approach helped us to get reliable statistical values that allowed us to infer about the relevant variables for Psoriasis detection.

This study was not only a biological challenge of identification and selection of relevant symptom but it also highlights the statistical challenge of modelling medical categorical observations.

References

[1] P. I. M. Schmitz. *Logistic Regression in Medical Decision Making and Epidemiology*. Proefschrift, Erasmus University Rotterdam, April 1986. Available at: `https://repub.eur.nl/pub/38819/860423_SCHMITZ,%20Paulus%20Ignatius%20Maria.pdf`

[2] H. Altay Güvenir, G. Demiröz, N. Ilter. *Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals*. Artificial Intelligence in Medicine, 1 July 1998.

[3] UC Irvine - Machine Learning Repository. *Dermatology Dataset*. Available at: `https://archive.ics.uci.edu/dataset/33/dermatology`

5 Appendix

Table 6: Summary of Clinical and Histopathological Features

Feature Number	**Description**	**Value Range**
1	Erythema (The severity of erythema in wounds)	0, 1, 2, 3
2	Scaling (Squam, dandruff peeling off the skin, dandruff amount in the lesions)	0, 1, 2, 3
3	Definite borders (Whether the wounds are sharply circumscribed)	0, 1, 2, 3
4	Itching (Intensity of itching in wounds)	0, 1, 2, 3
5	Koebner phenomenon (Limited manifestation of dermatological disease in the area of stimulation as a result of traumatic stimulation of the skin)	0, 1, 2, 3
6	Polygonal papules (Multi-edged, raised, less than 1 cm in diameter lesions on the skin)	0, 1, 2, 3
7	Follicular papules (Swellings less than 1 cm in height, distributed at equal distances from each other)	0, 1, 2, 3
8	Oral mucosal involvement (Lesions formation in the oral mucosa)	0, 1, 2, 3
9	Knee and elbow involvement (Lesions formation on knees and elbows)	0, 1, 2, 3
10	Scalp involvement (Lesions formation on the scalp)	0, 1, 2, 3
11	Family history (Whether there is a family history)	0, 1
34	Age	Continuous
12	Melanin incontinence (Brown granules that appear on the skin under the epidermis layer)	0, 1, 2, 3
13	Eosinophils in the infiltrate (An increase in a type of white blood cell)	0, 1, 2, 3
14	PNL infiltrate: Polymorphonuclear leukocyte spread	0, 1, 2, 3
15	Fibrosis of the papillary dermis	0, 1, 2, 3
16	Exocytosis	0, 1, 2, 3
17	Acanthosis	0, 1, 2, 3
18	Hyperkeratosis	0, 1, 2, 3
19	Parakeratosis	0, 1, 2, 3
20	Clubbing of the rete ridges	0, 1, 2, 3
21	Elongation of the rete ridges	0, 1, 2, 3
22	Thinning of the suprapapillary epidermis	0, 1, 2, 3
23	Spongiform pustule	0, 1, 2, 3
24	Munro microabcess	0, 1, 2, 3
25	Focal hypergranulosis	0, 1, 2, 3
26	Disappearance of the granular layer	0, 1, 2, 3
27	Vacuolisation and damage of basal layer	0, 1, 2, 3
28	Spongiosis	0, 1, 2, 3
29	Saw-tooth appearance of retes	0, 1, 2, 3

Continued on next page

Table 6 – continued from previous page

Feature Number	Description	Value Range
30	Follicular horn plug	0, 1, 2, 3
31	Perifollicular parakeratosis	0, 1, 2, 3
32	Inflammatory mononuclear infiltrate	0, 1, 2, 3
33	Band-like infiltrate	0, 1, 2, 3