

Diagnosis of Erythematous-Squamous Skin Diseases with Machine Learning Algorithms

Mustafa Necati BOZOK (✉ mustafa45037@ogr.duzce.edu.tr)

Düzce University

Ali ÇALHAN

Düzce University

Research Article

Keywords: Machine learning, Erythematous-Squamous skin diseases, Outlier detection

Posted Date: May 18th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1656894/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Since erythema squamous skin diseases show very close findings in clinical examination, a biopsy is taken from the patient for definitive diagnosis and the diagnosis of the disease can be made according to the biopsy result. On literature, classification studies were carried out on these diseases using machine learning and classification methods. Researches were mostly focused on optimizing and reducing database features for better classification score. Due to importance of reflecting specifications of diseases we especially focused on dataset features named as as clinic or histopathological features findings. In this study, histopathological features of diseases were discussed first and then we developed an algorithm to remove outlier data from the dataset. This algorithm leads us to discover a threshold value to achieve better outlier removal. Logistic Regression, KNeighbors Classifier, Support Vector Classifier, Gaussian Naive Bayes, Decision Tree Classifier and Random Forest Classifier methods applied to the outlier free dataset. It was determined that the Gaussian Naive Bayes method was the most appropriate classification method with 100% score. The results we obtained as a result of the algorithm we developed, being compatible with the clinical and histopathological features of skin diseases with erythema squamous, is a positive result for this study.

1. Introduction

Chronic diseases are diseases that progress slowly, do not fully heal with treatment, and occur with multifactorial causes in which genetic factors are involved in the etiology. Erythematous-squamous skin diseases affect the individual's mental, social and quality of life, as well as cause psychological stress on the family and negatively affect social mental health. These diseases cause loss of workforce and economic negativities with high treatment costs with high cost drugs obtained from foreign countries (Akdeniz, 2019).

There are 6 different erythematous-squamous skin diseases in the data set discussed in our study (İlter & Güvenir, 1998). In order to understand the diseases in the data set and the relations of these diseases, the following information was compiled by considering the diseases first.

Psoriasis

Psoriasis is a chronic disease characterized by inflammation with periods of exacerbation and remission, with the formation of erythematous plaques or papules with pearlescent-white scales on the skin. Its prevalence in the community is around 1–2%. It is seen in all age groups, regardless of gender. Psoriasis continues with joint involvement as well as skin disease and can cause death by damaging many organs together with severe cases, accompanied by psychiatric disorders, inflammatory bowel disease, insulin resistance (Alper et al., 2016) (Aykol et al., 2011). When the crusts of the psoriasis plaque are removed, the appearance of punctate bleeding foci (Auspitz sign) is helpful in the diagnosis (Topkarcı, 2013). The histopathological findings of psoriasis are regular elongation of the rete ridges, elongation of the dermal papillae, edema of the dermal papillae, dilated blood vessels, thinning of the suprapapillary plate, intermittent parakeratosis, absence of a granular layer, perivascular and dermal infiltrates of lymphocytes (Kim et al., 2015).

Seborrheic dermatitis

Seborrheic dermatitis is a chronic inflammatory disease characterized by oily yellowish scaly and erythematous scaling patches, which can occur on the face, scalp and body folds, periodically recurs or heals (Aksoy & Aksu Arica, 2016a) (Harmanyeri et al., 1990). It is a chronic, recurrent inflammatory skin disease characterized by erythematous scaling patches and is usually located in areas rich in sebaceous glands. It is divided into infantile and adult-onset. While infantile is initially seen in the first 3 months of life, adults initially appear between the ages of 20–40. It is more common in men and peaks in the 40s. While the severity of the disease increases in the winter months, the appearance of the disease improves with the effect of ultraviolet rays in the summer months (Aksoy & Aksu Arica, 2016b). The disease increases with stress, depression and fatigue (Aksoy et al., 2012). Although sebum production is thought to be the cause of the disease because the disease involvement areas are rich in sebaceous glands, sebum in patients is at a normal level. Although the disease is not seen in areas where sebaceous glands are absent, such as the palms and soles, seborrheic dermatitis is not a sebaceous gland disease, although it occurs in sebaceous areas (Doğruk Kaçar & Özüğüz, 2016). It is seen frequently in the society and is not perceived as a disease. Winter and autumn are effective in increasing the severity of the disease (Saçar & Saçar, 2010). The histopathological findings of seborrheic dermatitis are parakeratosis, epidermal spongiosis, lymphocytic exocytosis, dermal inflammatory cell infiltration (Park et al., 2016).

Lichen planus

Lichen planus is a papulosquamous, non-infectious inflammatory disease involving the skin and mucous membranes (Yanık & Albayrak, n.d.). It does not show organ involvement and involves skin, nails and mucosal surfaces. It is more common in women. Oral lichen planus has an increased risk of developing oral cancer. Therefore, its clinical classification is important for follow-up. (Turkoglu Babakurban et al., n.d.). The histopathological findings of lichen planus are saw-tooth rete ridges, atrophy, acanthosis, hyperorthokeratosis, melanin incontinence (Cheng et al., 2016).

Pityriasis rosea

Pityriasis rosea is a self-limiting, common acute papulosquamous inflammatory disease affecting the trunk and extremities. It is seen in children and adults and is rarely seen under the age of 10 (Kılınç, 2019). In most of the cases, the first finding is "medallion plaque", followed by classical lesions on the trunk and proximal extremities within 5–14 days. There is no definite conclusion about which season it is specific to (Başkan et al., 2011). The histopathological findings of pityriasis rosea are spongiosis, exocytosis (Özyürek et al., 2014)

Cronic dermatitis

It is a disease known as eczema in Turkey. Dermatitis types include atopic dermatitis, seborrheic dermatitis, contact dermatitis, allergic contact dermatitis, irritant contact dermatitis, hypereosinophilic dermatitis and other types. Atopic dermatitis is a recurrent, chronic, inflammatory skin disease that usually begins in early childhood (Yıldırım & Özcanlı, 2009). The lifetime prevalence of the disease is increasing all over the world and especially in developed countries (Ertam Sağduyu et al., 2018). The histopathological findings of cronic dermatitis are elongation of the rete ridges, prominent hyperkeratosis, minimal spongiosis (Bieber, 2010).

Pityriasis rubra pilaris

Pityriasis rubra pilaris (PRP) is a rare inflammatory disease, the etiology of which is not fully known, and the definitive treatment is still being investigated. It occurs equally in men and women. Immune system diseases and infections can be a trigger for the emergence of this disease. In the PRP age distribution, the majority of patients are in their 10s, 50s and 60s. When classification is made according to age distribution, it is divided into 5 groups. More than 50% of the patients are adults and are in the first group. 80% of patients in the first group recover within 3 years. The patients in the second group are adults, but the disease progresses chronically. The patients in the third group consist of children aged 5–10 years, and the disease characteristics are the same as in the first group. The patients in the fourth group are children between the ages of 3–10. Well defined borders follicular hyperkeratosis and erythema are seen in the knees and elbows. Stress and dissipation are low. The patients in the fifth group are children between the ages of 0–4 and the disease progresses chronically. The patients in the fifth and last group have the same symptoms as the first group and these patients are HIV virus carriers. It varies with treatment (Cohen et al., n.d.). The histopathological findings of pityriasis rubra pilaris are psoriasiform epidermis with areas of parakeratosis, plugs of the follicular infundibulum (Katherine L, 2003).

2. Related Work

In the study conducted by Güvenir et al. in 1998, VF15 was used as a classification algorithm developed by Güvenir. In the classification process, the VF15 algorithm was not affected by the missing data in the database. As a result of the estimation made with the VF15 algorithm during the preparation of the data, it was revealed that the specialist doctor was wrong in two diagnoses. With cross validation (10 fold cross validation) analysis, 96.2% correct classification was made with the VF15 algorithm. In order to determine the most suitable features for classification, after the weight values of the features were determined by the genetic algorithm method, when the classification was made over the features with the appropriate weight value, the classification was 99.2% higher than the VF15 algorithm (Güvenir et al., 1998).

In the study conducted by Güvenir and Emeksiz in 2000, a comparison was made using 3 different methods in the classification of erythemato-squamous skin diseases. K-NN classification, Naive Bayesian classification and VF15 classifications were applied and their effectiveness was compared with cross-validation analysis. In addition, it is aimed to make effective use in the field by developing an interface in the C++ programming language of these classifications (Güvenir & Emeksiz, n.d.).

In the study conducted by Übeyli et al. in 2005, a classification study was carried out on the UCI database by using the adaptive network-based fuzzy inference system (ANFIS), which is a combination of artificial neural networks and fuzzy logic. The accuracy of the classification study made with the ANFIS system was measured with the confusion matrix and the classification result was obtained with an accuracy of 95.5% (Übeyli & Güler, 2005).

In the study carried out by Karabatak and Ince in 2009, a classification system based on neural networks (Neural Network) and matching rule (Assosation Rule) was used. In their study, it was stated that it was more important to determine the features of the classes rather than the classification, and the determination of the most accurate features that could provide a high rate of classification was provided by the AR2 method developed by Karabatak. As a result of classification using neural networks with 24 features determined by AR2 instead of 34 features in the database, an accuracy of 98.6% was obtained (Karabatak & Ince, 2009).

In the study conducted by Abdi and Giveki in 2013, the AR-PSO-SVM system, which is a supervised learning method, was used to classify the data on erythemato-squamous diseases in the relevant database. The age attribute was excluded from the scope of the study, since there were records with missing age information in the database. By using the matching rule defined by Karabatak as the matching rule (AR), 24 features were determined by simplifying the features. The reason for using particle swarm optimization is to optimize the parameters required by the support vector machine. By determining the most appropriate input parameters with PSO, classification was made with an accuracy of 98.91% using confusion matrix measurement with SVM (Abdi & Giveki, 2013).

In the study carried out by Xie and Wang in 2011, the number of features was reduced to 21. In their study, the support vector machines have high performance in classification. But if the number of features in the used database is high, an over-learning situation will occur, a hybrid optimization process has been applied by applying the filtering with the F-score method developed on the features and the comprehension method with the forward sequential search method. Grid search method was applied to determine the necessary parameters in the support vector machines method, which will be applied after the features are determined. As a result of their study, classification was made with an accuracy rate of 98.61% (Xie & Wang, 2011).

In the literature review study conducted by Thomsen et al. in 2019, classification studies on dermatology datasets and dermatology images and the prediction rates obtained are presented in tables (Thomsen et al., 2020). In the literature review on the dermatology dataset, which is the subject of our study, in the article study titled "Medical Diagnosis by Feature Selection using Particle Swarm Optimization and Support Vector Machine" by Daliri, the features of the Daliri dermatology database were optimized by using binary particle swarm optimization (BPSO) method. The number of 34 features was reduced to 20 and classification was made with 100% accuracy (Daliri, 2012).

As a result of the comprehensive research conducted by Wang et al. in 2019, the methods of detecting outlier data to be removed from the dataset were discussed (Wang et al., 2019).

Classification study was carried out by Anifah and Haryanto in the classification study in 2021 with the Linear Vector Quantization method. As a result of the study, classification studies were carried out with an accuracy of 40% for psoriasis, 45% for seborrheic dermatitis, 30% for lichen planus, 45% for pityriasis rosea, 50% for chronic dermatitis and 90% for pityriasis rubra pilaris (Anifah & Haryanto, 2022).

In the classification study conducted by Alotaibi in 2022, a hybrid algorithm named K-nearest neighbor (KNN) and ReliefF was developed. While classification was made with the traditional KNN algorithm with an accuracy of 85.13%, the hybrid KNN model he developed was classified with an accuracy of 94.59% (Alotaibi, 2022).

In the classification study conducted by Elsayad et al. in 2022, classification was made with an accuracy of 99.07% using the Bayesian and SVM hybrid algorithm (Elsayad et al., 2022).

In the classification study conducted by Al-Kahlout et al. in 2021, a classification study was carried out with an accuracy of 98.36% with artificial neural networks and JNN (Just Neural Network) tool. (Al-Kahlout et al., 2021).

When the literature is examined in general, the minimum number of features has been obtained by optimizing the data set features by using support vector machines. Then it has been observed that classification studies have been carried out to a maximum of 100% by working classification algorithms on the optimized data set features. It was seen that the classification results were evaluated over numbers, regardless of whether the optimized data set obtained as a result of all studies reflects the clinical and histopathological features of these diseases. In order to detect outliers in our study, relatively inconsistent data were determined in each attribute data and the records of these data were removed from the data set.

In order to determine the outlier data density at the most appropriate rate, a threshold ratio was determined and classification accuracy rates were determined according to the relevant threshold value. When the threshold value is 100%, classification work is performed on all data without any elimination. Since feature loss occurred, the minimum threshold value was not lowered below 70%. Threshold values in the range of 100 - 70% were considered with 10% slices and classification studies were carried out on the data set from which outlier values were removed. By combining the classification results obtained according to the threshold values in the range of 100 - 70%, the threshold value that provides the highest number of high classification rates among the classification methods used was determined. By applying the 5-fold cross-validation method to the results obtained, the machine learning method that provides the highest classification rate was determined. Among the Logistic Regression, KNeighbors Classifier, Support Vector Classifier, Gaussian Naive Bayes, Decision Tree Classifier and Random Forest Classifier classification methods, the Gaussian Naive Bayes method gave the best result with a rate of 100%. It was determined that the data set obtained by outlier elimination was compatible with the clinical and histopathological features of these diseases.

3. Materials And Methods

3.1. Data Set

In this study, the dataset of erythematous-squamous skin diseases in the machine learning database of the University of California Irvine (UCI) at <https://archive.ics.uci.edu/ml/datasets/dermatology> was used. The data set was first prepared by Prof. Dr. H. Altay GÜVENİR, Prof. Dr. Nilsel İLTER and Dr. Gülşen DEMİRÖZ as "Diagnosis of Erythema-Scaly Diseases with VFI Algorithm" (Güvenir et al., 1998), the data set was provided by Prof. Dr. Nilsel İLTER at Gazi University Department of Dermatology (Güvenir & Emeksiz, n.d.).

3.2. Features

The data set in the file named "dermatology.data" consists of 366 records of 34 different features separated by a comma. When the features are examined, the features 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 34 contain the values of the clinical findings. Features between 12 and 33 contain the values of the histopathological findings. Except for the age feature number 34, all features between 1 and 33 have values between 0 and 3. Age feature has 8 unknown value and 358 values varying between 0 and 75. Number of patients are : Psoriasis 112 patients, Seborrheic dermatitis 61 patients, Lichen planus 72 patients, Pityriasis rosea 49 patients, Chronic dermatitis 52 patients, Pityriasis rubra pilaris 20 patients. Totally 366 patients available on dataset.

The clinical and histopathological features are explained as below.

Clinical features

(with values 0, 1, 2, 3)

1: Erythema (The severity of erythema in wounds)

2: Scaling (Squam, dandruff peeling off the skin, dandruff amount in the lesions)

3: Definite borders (Whether the wounds are sharply circumscribed)

4: Itching (Intensity of itching in wounds)

5: Koebner phenomenon (Limited manifestation of dermatological disease in the area of stimulation as a result of traumatic stimulation of the skin (Rifaioğlu et al., 2014))

6: Polygonal papules (Multi-edged, raised, less than 1 cm in diameter lesions on the skin)

7: Follicular papules (Swelling less than 1 cm in height, distributed at equal distances from each other)

8: Oral mucosal involvement (Lesions formation in the oral mucosa)

9: Knee and elbow involvement (Lesions formation on knees and elbows)

10: Scalp involvement (Lesions formation on the scalp)

11: Family history, (0–1) (Whether there is a family history)

34: Age (Have linear values)

Histopathological features

These are the findings obtained by biopsy taken from patients. (values are in the range of 0, 1, 2, 3)

12: Melanin incontinence (Brown granules that appear on the skin under the epidermis layer)

13: Eosinophils in the infiltrate (An increase in a type of white blood cell)

14: PNL infiltrate : Polymorphonuclear leukocyte spread. Migration and arrival of neutrophils to the disease site. Increase in the number of white blood cells of leukocytes, inflammation.

15: Fibrosis of the papillary dermis : Accumulation of new fibrotic material (collagen) due to disease in the papillary dermis layer of the skin.

16: Exocytosis : Accumulation of white blood cells towards the epidermis.

17: Acanthosis : Thickening of the epidermis layer.

18: Hyperkeratosis : Thickening of the keratin layer.

19: Parakeratosis : Nuclear cell formation in the keratin layer.

20: Clubbing of the rete ridges : Clubbing of the ridges of the rete.

21: Elongation of the rete ridges : Elongation of the ridges of the rete.

22: Thinning of the suprapapillary epidermis : Thinning of the epidermis over the papillary dermis.

- 23: Spongiform pustule : Spongy vesicles (pustules) filled with pus (neutrophils)
- 24: Munro microabcess : Small vesicles filled with neutrophils in the epidermis.
- 25: Focal hypergranulosis : Focal thickening of the granular layer of the epidermis.
- 26: Disappearance of the granular layer : Disappearance of the granular layer of the epidermis.
- 27: Vacuolisation and damage of basal layer : Formation of spongy cavities as a result of damage to the basal layer.
- 28: Spongiosis : Edema between epidermis cells.
- 29: Saw-tooth appearance of retes : Formation of rete ridges in a sawtooth appearance.
- 30: Follicular horn plug : Formation of plugs in hair follicles.
- 31: Perifollicular parakeratosis : Presence of nucleated cells around the hair follicle in the corneum layer.
- 32: Inflammatory mononuclear infiltrate : Migration of mononuclear inflammatory cells.
- 33: Band-like infiltrate : Migration of white blood cells in band appearance.

3.3. Classification of Dataset Features

In order to determine which disease class is associated with clinical and histopathological features in the data set, a ratio between 0–3 values was first made. All features are scaled except for Age and Class features. As a result of this process, the ratio of each feature belonging to the disease class was determined with a value between 0–3. The classification rates of each feature are determined by graphics and these values are shown separately in the table. While some features are seen in more than one disease, it has been determined that some of them are specific to the related disease.

All the features in the data set were charted separately and the average values of each feature were calculated according to the disease it belongs to. Values between 1 and 6 on the x-axis of the figures represent 1-Psoriasis, 2-Seboric dermatitis, 3-Lichen planus, 4-Pityriasis rosea, 5-Cronic dermatitis and 6-Pityriasis rubra pilaris diseases, respectively. The numbers in the range of 0–3 on the y-axis in the figures show the average values of these diseases in the data set of the related feature.

Figure 1 shows the mean values of the diseases of the Erythema feature in the data set. In Psoriasis as number 1 and Seboric dermatitis as number 2, Erythema has the highest incidence rate with a value of 2.3 according to the ratio values in the 0–3 range. With a value of 1.5, the Erythema feature has the lowest incidence in chronic dermatitis.

In Fig. 2, the Definitive borders feature in the data set is seen at the highest rate in 1-Psoriasis and 3-Lichen Planus diseases, with the lowest rate of 0.8 in 5-Cronic dermatitis.

In Fig. 3, the Itching feature in the data set is seen at the highest rate in 3-Lichen planus disease, with the lowest rate of 0.5 in 4-Pityriasis rosea and Pityriasis rubra pilaris diseases.

In Fig. 4, the Koebner phenomenon feature in the dataset is included in 1-Psoriasis, 3-Lichen planus and 4-Pityriasis rosea diseases, while it is not seen in other diseases such as 2- Seborrheic dermatitis, 5- Cronic dermatitis and 6- Pityriasis rubra pilaris and does not have any value.

The Band-like infiltrate feature in Fig. 5 and dataset is only seen in 3-Lichen planus disease and has a rate of 2.7 in this disease.

The Clubbing of the rete ridges feature in Fig. 6 and dataset is seen only in 1-Psoriasis disease and has a rate of 2.1 in this disease.

Perifollicular parakeratosis feature in the data set in Fig. 7 is seen only in 6-Pityriasis rubra pilaris disease and has a rate of 2.0 in this disease.

The Fibrosis of the papillary dermis feature in the data set in Fig. 8 is only seen in 5-Cronic dermatitis disease and has a rate of 2.3 in this disease.

Figure 9 shows the age distribution of the Age feature in the data set according to the diseases. While patients between the ages of 0–40 are observed in 1-Psoriasis, 2-Seborrheic dermatitis, 3-Lichen planus, 4-Pityriasis rosea and 5-Cronic dermatitis diseases, the age range of patients in 6-Pityriasis rubra pilaris disease has values of 0–10.

In Fig. 10, a statistical graph was drawn instead of the mean values of the age ranges of the patients in the data set. Thanks to this graph, the value of a 22-year-old patient who was found to be outlier in 6-Pityriasis rubra pilaris disease outside the 7–16 age range is shown as a dot in the 6th column.

$$meanC_iF_j = \frac{\sum_1^n F_j}{n} \quad (2)$$

The values found on C_iF_j are indexed. The number of each index value in the corresponding attribute is counted. As seen in Eq. 3 letter v indicates indexed values of active feature of active class. And letter q_{1-33} indicates that the count of the indexed v values for each feature separately.

$$count(\forall C_iF_j \forall v_n) = \begin{cases} v_1, v_2, \dots, v_n \rightarrow q_1 \\ \dots \\ v_1, v_2, \dots, v_n \rightarrow q_n \end{cases} \quad (3)$$

x letter represent the maximum value of q gives us the most repeated values of active feature. Eq. 4 will be as follows:

$$x = \max(count(q_1, \dots, q_n)) \quad (4)$$

y letter represent the minimum value of q gives us the least repeated values of active feature. Eq. 5 will be as follows:

$$y = \min(count(q_1, \dots, q_n)) \quad (5)$$

As used before in Eq. 2, we used n as the number of elements of the active class. n reflects us the count of active classes active features elements.

The ratio of the number of the most repeated values in the studied feature to the number of values of the related feature is calculated by Eq. 6.

$$ratio = \frac{x}{n} \quad (6)$$

We need a treshold value to be able to evaluate the ratio value we found.

In order to determine the outlier data density at the most appropriate rate, a threshold ratio was determined and classification accuracy rates were determined according to the relevant threshold value. When the threshold value is 100%, classification work is performed on all data without any elimination. Since feature loss occurred, the minimum treshold value was not lowered below 70%. Threshold values in the range of 100 - 70% were considered with 10% slices and classification studies were carried out on the data set from which outlier values were removed.

If ratio value resulted in 1, its reflects the situation that all values of active feature has same values without not having any outlier value.

Otherwise if ratio value bigger than treshold and lower than 1, its shows that active feature has outlier values to be delete from database. Algorithm will delete records which indicates y from Eq. 5.

This is illustrated in Eq. 6 below.

$$ratio = \begin{cases} 1 \rightarrow \text{All values of active feature are same.} \\ (ratio > \text{treshold}) \wedge (ratio < 1) \rightarrow \text{delete record which value is 'sindicate'} \end{cases} \quad (6)$$

With Eq. 7, the Ratio evaluation is applied separately for all the features in the entire dataset.

$$ratio \rightarrow (\forall C_i \forall F_j) \quad (7)$$

With this method, all records that will be considered as outlier data are removed from the data set. Then we can apply machine learning methods on the remaining error-free data set.

Diagram 1 : General flow chart of classification work

The general flow chart of the classification work to be done is shown in Diagram 1.

Diagram 2 : Detailed flow chart of detecting the best classification method.

With Scheme 2, the flow chart of the machine learning methods applied on the dataset obtained by evaluating the features according to the threshold value and deleting the related records is given.

3.4. Performance Analysis of Classification Methods

In our study, classification methods of the Sklearn library in the Python programming language were used. The data set is divided into 33% test and 67% train sections, and classification studies will be carried out on 121 records, which corresponds to 33% of the 366 records in the data set. By using training and test data, LogisticRegression, KNeighborsClassifier, Support Vector Classification (SVC), GaussianNB, DecisionTreeClassifier and RandomForestClassifier classification methods were applied, respectively.

The classification results were evaluated by k-fold cross validation values. The results obtained without deleting the outlier records were first evaluated with the single fold cross validation method.

Since feature loss occurred, the minimum threshold value was not lowered below 70%. Threshold values in the range of 100 – 70% were considered with 10% slices and classification studies were carried out on the data set from which outlier values were removed. By combining the classification results obtained according to the threshold values in the range of 100 – 70%, the threshold value that provides the highest number of high classification rates among the classification methods used was determined. By applying the 5-fold cross-validation method to the results obtained, the machine learning method that provides the highest classification rate was determined.

Logistic regression, KNeighbors Classifier, Support Vector Classification, Gaussian Naive Bayes, Decision Tree Classifier and Random Forest Classifier methods were used as classification methods.

3.4.1. Logistic Regression

With logistic regression, a discrimination model is created according to the number of groups in the structure of the data. With this model, the new data taken into the dataset is classified. The purpose of using logistic regression is to create a model that will establish the relationship between the least variable and the most suitable dependent and independent variables. (Bircan, 2004).

3.4.2. KNeighbors Classifier

In the KNeighbors Classifier classification (k-nearest neighbors), a clustering is created according to the distance values of the classes depending on the k parameter value in the existing data set, and the method of classifying the new data according to the similarity to these clusters is applied. (Keskenler et al., 2021)

3.4.3. Support Vector Classification (SVC)

Support vector classification is the process of predicting what will be the outputs of new data based on existing data. Support vector classification performs classification by finding the separator plane with the widest range between classes (Güner et al., n.d.).

3.4.4. Gaussian Naive Bayes

Gaussian Naive Bayes classification applies a classification algorithm based on the probability of the Gaussian distribution (Karatay & Algahani, 2021).

3.4.5. Decision Tree Classifier

Decision tree classifier method consists of 3 components consisting of node, branch and leaf. Questions are asked to create a tree structure using the attributes in the training data, and these processes continue until node or leaves without branches are reached (Çölkesen & Kavzoğlu, 2010).

3.4.6. Random Forest Classifier

The purpose of the random forest classifier classification method is to bring together the decisions made by many trees trained in different training sets instead of a single decision tree. (Daş et al., n.d.).

Table 2
Single fold cross validation scores obtained between 100% - 70% treshold values

No	Threshold	Classification Type	Single fold cross validation Score	Cronic dermatitis	Lichen planus	Pityriasis rosea	Pityriasis rubra pilaris	Psoriasis	Seboreic dermatitis	Total patients
1	1.0	Logistic Regression	0.992	52.0	72.0	49.0	20.0	112.0	61.0	366.0
2	1.0	KNeighborsClassifier	0.934	52.0	72.0	49.0	20.0	112.0	61.0	366.0
3	1.0	SVC	0.975	52.0	72.0	49.0	20.0	112.0	61.0	366.0
4	1.0	GaussianNB	0.818	52.0	72.0	49.0	20.0	112.0	61.0	366.0
5	1.0	DecisionTreeClassifier	0.95	52.0	72.0	49.0	20.0	112.0	61.0	366.0
6	1.0	RandomForestClassifier	0.959	52.0	72.0	49.0	20.0	112.0	61.0	366.0
7	0.90	Logistic Regression	0.99	42.0	65.0	45.0	15.0	103.0	46.0	316.0
8	0.90	KNeighborsClassifier	0.981	42.0	65.0	45.0	15.0	103.0	46.0	316.0
9	0.90	SVC	0.981	42.0	65.0	45.0	15.0	103.0	46.0	316.0
10	0.90	GaussianNB	0.933	42.0	65.0	45.0	15.0	103.0	46.0	316.0
11	0.90	DecisionTreeClassifier	0.99	42.0	65.0	45.0	15.0	103.0	46.0	316.0
12	0.90	RandomForestClassifier	1.0	42.0	65.0	45.0	15.0	103.0	46.0	316.0
13	0.80	Logistic Regression	1.0	29.0	41.0	40.0	15.0	87.0	31.0	243.0
14	0.80	KNeighborsClassifier	0.988	29.0	41.0	40.0	15.0	87.0	31.0	243.0
15	0.80	SVC	1.0	29.0	41.0	40.0	15.0	87.0	31.0	243.0
16	0.80	GaussianNB	1.0	29.0	41.0	40.0	15.0	87.0	31.0	243.0
17	0.80	DecisionTreeClassifier	1.0	29.0	41.0	40.0	15.0	87.0	31.0	243.0
18	0.80	RandomForestClassifier	1.0	29.0	41.0	40.0	15.0	87.0	31.0	243.0
19	0.70	Logistic Regression	1.0	29.0	30.0	31.0	6.0	63.0	12.0	171.0
20	0.70	KNeighborsClassifier	0.965	29.0	30.0	31.0	6.0	63.0	12.0	171.0
21	0.70	SVC	1.0	29.0	30.0	31.0	6.0	63.0	12.0	171.0
22	0.70	GaussianNB	1.0	29.0	30.0	31.0	6.0	63.0	12.0	171.0
23	0.70	DecisionTreeClassifier	0.947	29.0	30.0	31.0	6.0	63.0	12.0	171.0
24	0.70	RandomForestClassifier	0.982	29.0	30.0	31.0	6.0	63.0	12.0	171.0

Table 2 shows the single-fold cross-validation score values of the machine classification methods applied to the database according to each threshold value. While the Threshold value was 100%, all records were listed, while at 70%, the number of patient records decreased to 171.

When the whole table was evaluated, it was determined that the classification algorithms with 80% threshold value had the highest accuracy rate. Since the rates here are determined by the single fold validation method, the 5-fold cross validation evaluation method was applied to the results obtained for the 80% threshold value found in order to obtain more accurate results.

Table 3
5-fold cross validation result table

Classification Method	Normal Classification		Classification by %80 Threshold value	
	Single Fold	5-Fold Cross Validation	Single fold	5-Fold Cross Validation
Logistic Regression	0.992	0.954	1.0	0.987
KNeighbors Classifier	0.934	0.922	0.988	0.975
Support Vector Classification (SVC)	0.975	0.962	1.0	0.975
Gaussian Naive Bayes	0.818	0.881	1.0	1.0
Decision Tree Classifier	0.95	0.959	1.0	0.968
Random Forest Classifier	0.983	0.967	1.0	0.981

When Table 3 is examined, it is seen that there is a difference between single fold validation and 5-fold cross validation values. Since the 5-fold cross validation is applied on the entire dataset due to its structure, it gives more accurate results about the performance of the classification method.

When the normal classification results and the classification results of the Threshold value are compared, it is seen that the success rate of the threshold value classification is higher. Again, when the 5-fold cross-validation method was applied, the Gauss Naive Bayes method achieved the highest classification success with 100%.

As can be seen from the Fig. 11, it has been determined that classification success rates are in the range of 100–96% when the threshold value is 1, and when it is at the 0.8 point, all methods show success at values close to 100%. From this graph, it is understood that by decreasing the threshold value from 1 to 0.8, 20% outlier record was detected and deleted from the table.

As a result of the normal classification study obtained in Table 2, the correlation graph of the dataset features was drawn with Graph 2. As a result of the normal classification, it was observed that the features of the diseases did not appear clearly.

The correlation graph of the outlier-free dataset obtained after the threshold value determined in Table 2 is shown in Fig. 13. The white areas in the graph show that there is a high correlation between the related feature and the disease.

Clinical and histopathological findings such as knee and elbow involvement, scalp involvement, clubbing or the rete ridges, elongation of the rete ridges, thinning of suprapapillary dermis are compatible with psoriasis. It was observed that the correlation graph of the psoriasis disease in the database, which was cleared of outlier data, was exactly compatible with the findings of this disease.

4. Conclusion And Findings

Many different machine learning methods are applied in the diagnosis of erythema squamous skin diseases. Each method classifies disease with reasonable accuracy.

Clinical and histopathological data obtained from the patient are used in the diagnosis of the disease. The specialist doctor uses these data to make the most appropriate diagnosis decision for the patient. With the experience of the medical profession, the specialist physician can decide whether the erroneous data is compatible with the relevant disease and can eliminate the erroneous values.

Outlier values in the data set cause incorrect rates to be obtained in the results of the applied classification method.

Outlier data were removed from the dataset in order to obtain a classification result that fully reflects the clinical and histopathological features of the diseases.

As a result, machine learning classification rates have been successfully achieved.

When the findings obtained from all these studies are evaluated:

- Although machine learning methods give effective results, a specialist doctor examination is always required.
- Outlier datas should be corrected as much as possible and these datas should be added to the data set again for an advanced working method and evaluated.
- For individuals who are suitable for the ordinary course of life, the data in this dataset is suitable for machine learning methods. However, in cases where pregnancy or other chronic diseases are accompanied, these methods will be insufficient.
- When outlier records are not deleted, classification methods consider these outlier records as part of the disease and reveal the results of misclassification.
- A feedback should be obtained as to whether the results obtained are compatible with the intended study.

References

1. Abdi, M. J., & Giveki, D. (2013). Automatic detection of erythematous diseases using PSO-SVM based on association rules. *Engineering Applications of Artificial Intelligence*, 26(1), 603–608. <https://doi.org/10.1016/j.engappai.2012.01.017>
2. Al-Kahlout, B. I., Naeem, M. M., & Shepherd, M. J. (2021). ANN for the Classification of Erythematous Disease. *International Journal of Academic and Applied Research (IJAAAR)*, 47–55. <http://dstore.alazhar.edu.ps/xmlui/handle/123456789/2707>
3. Alotaibi, A. S. (2022). Hybrid Model Based on ReliefF Algorithm and K-Nearest Neighbor for Erythematous Diseases Forecasting. *Arabian Journal for Science and Engineering*, 47(3), 1299–1307. <https://doi.org/10.1007/s13369-021-05921-z>
4. Anifah, L., & Haryanto. (2022). *Decision Support System Erythematous Diseases Classification Diagnosis Using Linear Vector Quantization Based Clinical Attributes*. 254–258. <https://doi.org/10.1109/ICRACOS53680.2021.9701952>
5. Bieber, T. (2010). Atopic dermatitis. *Annals of Dermatology*, 22(2), 125–137. <https://doi.org/10.5021/ad.2010.22.2.125>
6. Bircan, H. (2004). *Lojistik Regresyon Analizi* (Issue 2).
7. Cheng, Y. S. L., Gould, A., Kurago, Z., Fantasia, J., & Muller, S. (2016). Diagnosis of oral lichen planus: a position paper of the American Academy of Oral and Maxillofacial Pathology. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, 122(3), 332–354. <https://doi.org/10.1016/j.oooo.2016.05.004>
8. Çölkesen, İ., & Kavzoğlu, T. (2010). Classification of Satellite Images Using Decision Trees: Kocaeli Case. *Electronic Journal of Map Technologies*, 2(1), 36–45. www.tuik.gov.tr
9. Daliri, M. R. (2012). Feature selection using binary particle swarm optimization and support vector machines for medical diagnosis. *Biomedizinische Technik*, 57(5), 395–402. <https://doi.org/10.1515/bmt-2012-0009>
10. Daş, B., Türkoğlu, İ., Üniversitesi, F., Fakültesi, T., Mühendisliği, Y., Türkiye, E. / , Üniversitesi, B., Ve, M., Fakültesi, M., & Bölümü, B. M. (n.d.). *DNA Dizilimlerinin Sınıflandırılmasında Karar Ağacı Algoritmalarının Karşılaştırılması*. <http://www.ncbi.nlm.nih.gov/Genbank/genomes/bacteria>
11. Elsayad, A. M., Nassef, A. M., & Al-Dhaifallah, M. (2022). Bayesian optimization of multiclass SVM for efficient diagnosis of erythematous diseases. *Biomedical Signal Processing and Control*, 71, 103223. <https://doi.org/10.1016/J.BSPC.2021.103223>
12. Güner, N., Çomak, E., Üniversitesi, P., Fakültesi, M., & Bölümü, B. M. (n.d.). *Mühendislik Öğrencilerinin Matematik I Derslerindeki Başarısının Destek Vektör Makineleri Kullanılarak Tahmin Edilmesi Predicting Performance of First Year Engineering Students in Calculus by Using Support Vector Machines*.
13. Güvenir, H. A., Demiröz, G., & İtler, N. (1998). Learning differential diagnosis of erythematous diseases using voting feature intervals. In *Artificial Intelligence in Medicine* (Vol. 13).
14. Güvenir, H. A., & Emeksiz, N. (n.d.). *An expert system for the differential diagnosis of erythematous diseases*. www.elsevier.com/locate/eswa
15. Karabatak, M., & Ince, M. C. (2009). A new feature selection method based on association rules for diagnosis of erythematous diseases. *Expert Systems with Applications*, 36(10), 12500–12505. <https://doi.org/10.1016/j.eswa.2009.04.073>
16. Karatay, S., & Algahani, M. (2021). 1999 Marmara Depremi ve Güneş Tutulmasının Naive Bayes Sınıflayıcısı ile İstatistiksel Analizi. *European Journal of Science and Technology*. <https://doi.org/10.31590/ejosat.876223>
17. Katherine L, M. W. (2003). Pityriasis rubra pilaris. *Dermatology Online Journal*, 9(4), 6. <https://doi.org/10.5070/D365618361>
18. Keskenler, M. F., Dal, D., & Aydın, T. (2021). Yapay Zeka Destekli ÇOKS Yöntemi ile Kredi Kartı Sahtekarlığının Tespiti. *El-Cezeri Fen ve Mühendislik Dergisi*. <https://doi.org/10.31202/ecjse.908260>
19. Kim, B. Y., Choi, J. W., Kim, B. R., & Youn, S. W. (2015). Histopathological findings are associated with the clinical types of psoriasis but not with the corresponding lesional psoriasis severity index. *Annals of Dermatology*, 27(1), 26–31. <https://doi.org/10.5021/ad.2015.27.1.26>
20. Nanni, L. (2006). An ensemble of classifiers for the diagnosis of erythematous diseases. *Neurocomputing*, 69(7-9 SPEC. ISS.), 842–845. <https://doi.org/10.1016/j.neucom.2005.09.007>

21. Özyürek, G. D., Alan, S., & Çenesizolu, E. (2014). Evaluation of clinico-epidemiological and histopathological features of pityriasis rosea. *Postepy Dermatologii i Alergologii*, 31(4), 216–221. <https://doi.org/10.5114/pdia.2014.40641>
22. Park, J. H., Park, Y. J., Kim, S. K., Kwon, J. E., Kang, H. Y., Lee, E. S., Choi, J. H., & Kim, Y. C. (2016). Histopathological differential diagnosis of psoriasis and seborrheic dermatitis of the scalp. *Annals of Dermatology*, 28(4), 427–432. <https://doi.org/10.5021/ad.2016.28.4.427>
23. Rıfaioğlu, E. N., Ş En, B. B., & Ekiz, Ö. (2014). Tatuaj Komplikasyonu Olarak Koebner Fenomeni; Psoriasis Tanılı Bir Olgu. *Turk Dermatoloji Dergisi*, 8(4), 244–245. <https://doi.org/10.4274/tdd.2003>
24. Thomsen, K., Iversen, L., Titlestad, T. L., & Winther, O. (2020). Systematic review of machine learning for diagnosis and prognosis in dermatology. In *Journal of Dermatological Treatment* (Vol. 31, Issue 5, pp. 496–510). Taylor and Francis Ltd. <https://doi.org/10.1080/09546634.2019.1682500>
25. Übeyli, E. D., & Güler, I. (2005). Automatic detection of erythematous-squamous diseases using adaptive neuro-fuzzy inference systems. *Computers in Biology and Medicine*, 35(5), 421–433. <https://doi.org/10.1016/j.compbiomed.2004.03.003>
26. Wang, H., Bah, M. J., & Hammad, M. (2019). Progress in Outlier Detection Techniques: A Survey. *IEEE Access*, 7, 107964–108000. <https://doi.org/10.1109/ACCESS.2019.2932769>
27. Xie, J., & Wang, C. (2011). Using support vector machines with a novel hybrid feature selection method for diagnosis of erythematous-squamous diseases. *Expert Systems with Applications*, 38(5), 5809–5815. <https://doi.org/10.1016/j.eswa.2010.10.050>

Diagram

Diagram 1-2 are available in Supplementary Files section.

Figures

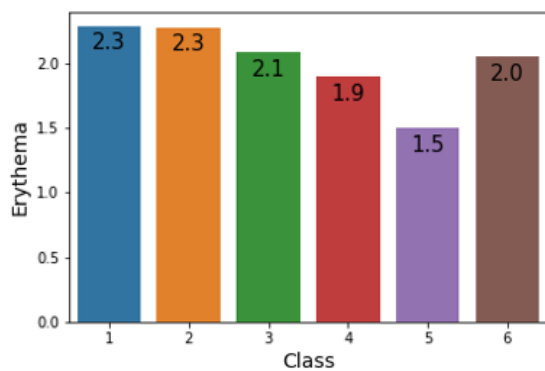


Figure 1

Erythema feature

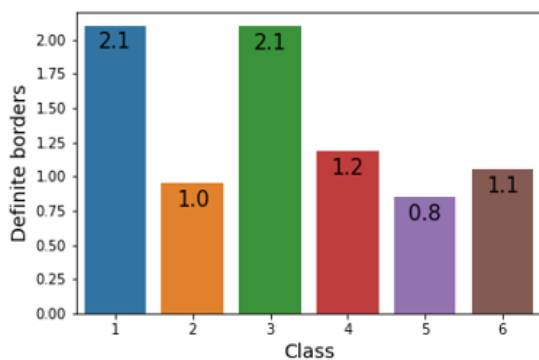


Figure 2

Class feature

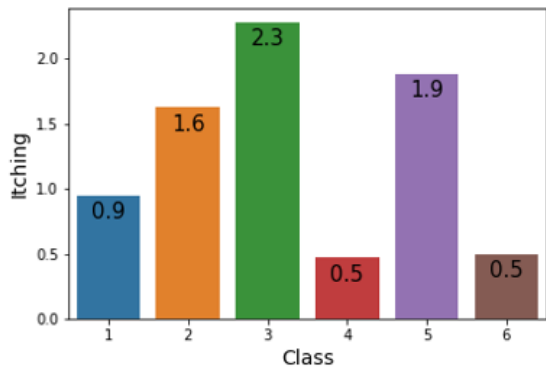


Figure 3

Itching feature

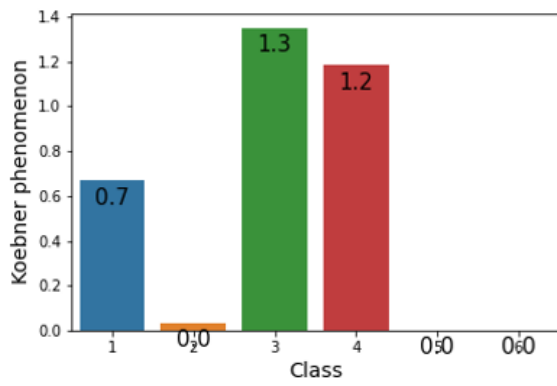


Figure 4

Koebner phenomenon feature

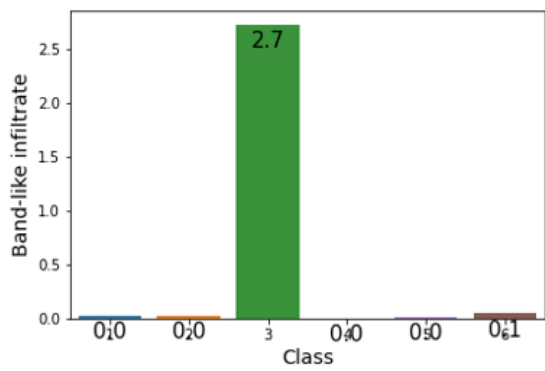


Figure 5

Band-like infiltrate feature

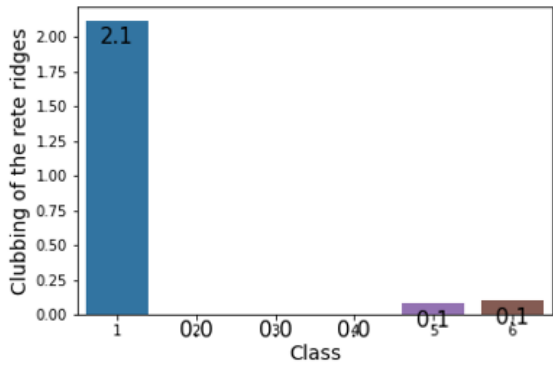


Figure 6

Clubbing of the rete ridges feature

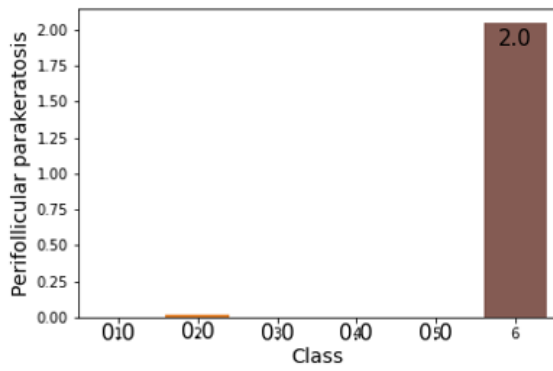


Figure 7

Perifollicular parakeratosis feature

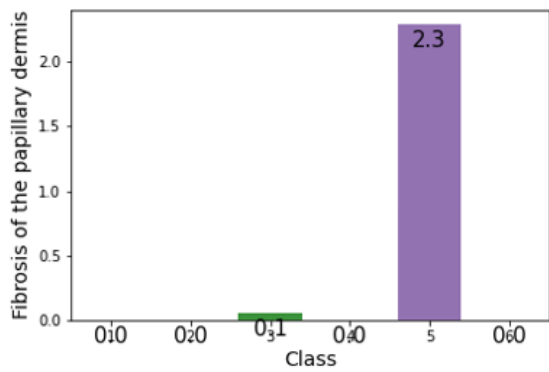


Figure 8

Fibrosis of the papillary dermis feature

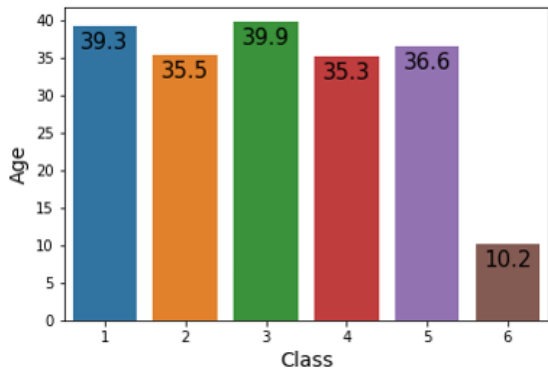


Figure 9

Age feature

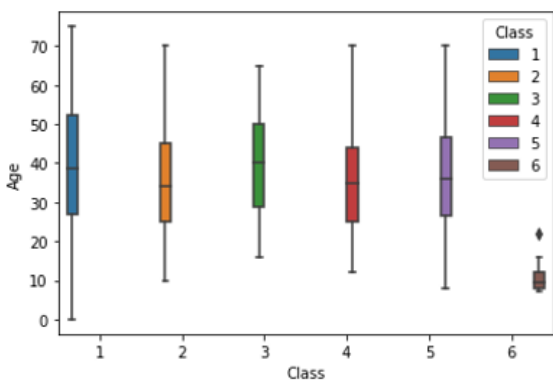


Figure 10

Statistical average value of Age feature

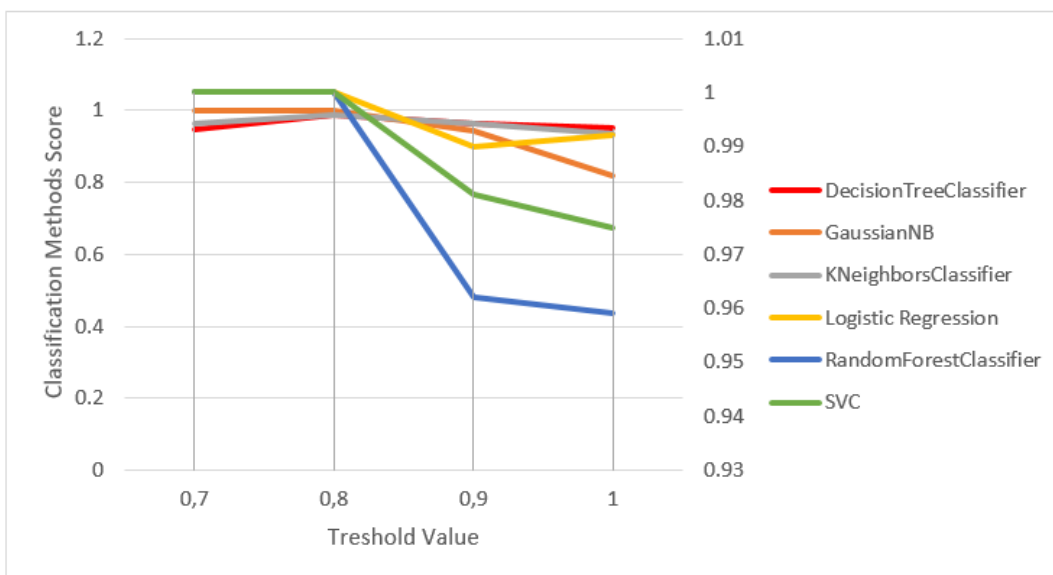


Figure 11

The success graph obtained with the threshold value of machine learning methods

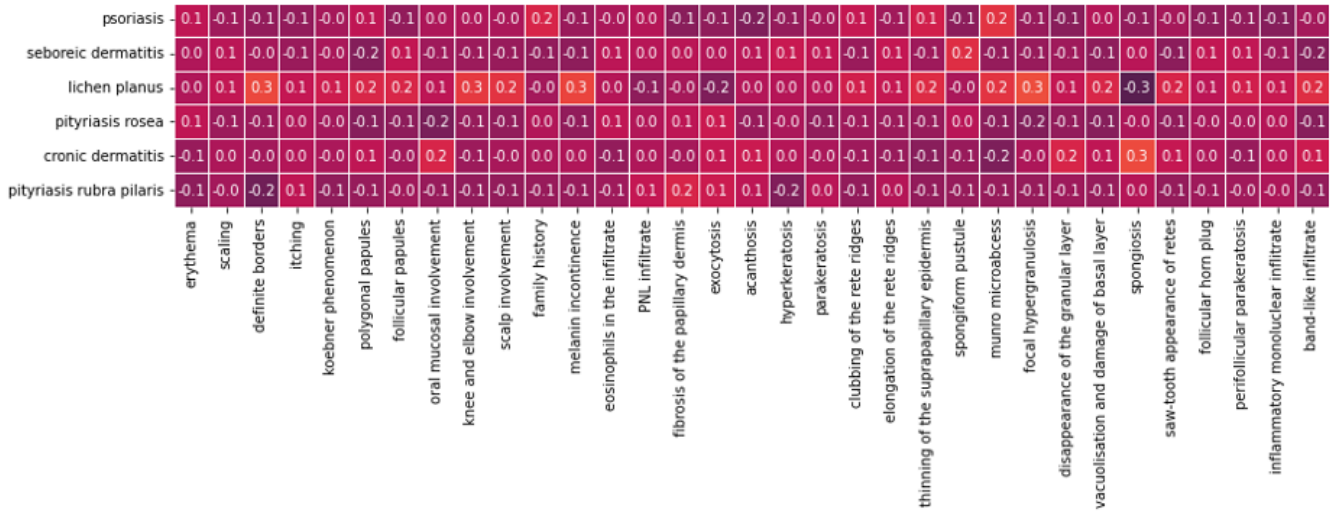


Figure 12

Dataset features correlation graph before the outlier data was cleaned

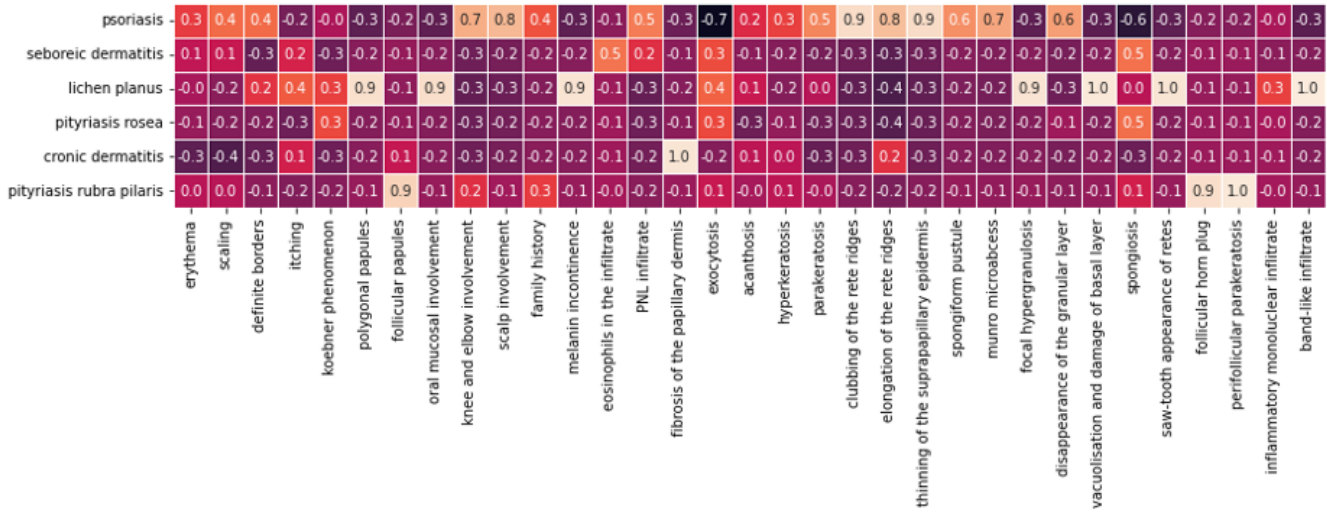


Figure 13

Dataset features correlation graph after the outlier data was cleaned

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [D1.png](#)
- [d2.png](#)