Skin Disorder

03.11.2022

Vivek Viradia
Rubixe
1st Cross Road, Kudlu Gate, 560068
Bangalore, Karnataka

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. In 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 the importance of reflecting specifications of diseases we especially focused on dataset features named 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, KNeihgbours 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.

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. Erythemato-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)

1. Psoriasis

Psoriasis is a skin disease that causes a rash with itchy, scaly patches, most commonly on the knees, elbows, trunk and scalp. Psoriasis is a common, long-term (chronic) disease with no cure. It can be painful, interfere with sleep and make it hard to concentrate.
Symptoms: Arthralgia; Erythema; Itch; Plaque,...

2. Seboreic Dermatitis

Seborrheic (seb-o-REE-ik) dermatitis is a common skin condition that mainly affects your scalp. It causes scaly patches, inflamed skin and stubborn dandruff. It usually affects oily areas of the body, such as the face, sides of the nose, eyebrows, ears, eyelids and chest.

3. Lichen Planus

Lichen planus (LIE-kun PLAY-nus) is a condition that can cause swelling and irritation in the skin, hair, nails and mucous membranes. On the skin, lichen planus usually appears as purplish, itchy, flat bumps that develop over several weeks.

Symptoms: Itch,...

4. Pityriasis Rosea

Pityriasis rosea is a rash that often begins as an oval spot on the face, chest, abdomen or back. This is called a herald patch and may be up to 4 inches (10 centimeters) across. Then you may get smaller spots that sweep out from the middle of the body in a shape that looks like drooping pine-tree branches.

Symptoms: Itch; Rash,....

5. Cronic Dermatitis

Atopic dermatitis (eczema) is a condition that causes dry, itchy and inflamed skin. It's common in young children but can occur at any age. Atopic dermatitis is long lasting (chronic) and tends to flare sometimes. It can be irritating but it's not contagious.

Symptoms: Itch; Xeroderma; Inflammation,...

6. Pityriasis Rubra Pilaris

Pityriasis rubra pilaris (PRP) is a rare condition that causes an orange-red, scaly rash on the skin with thickening and scaling of the palms and soles.

There are often small scaly bumps surrounding the hair follicles, described as nutmeg grater.

Clinical Features Specifications

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 (Swellings 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.

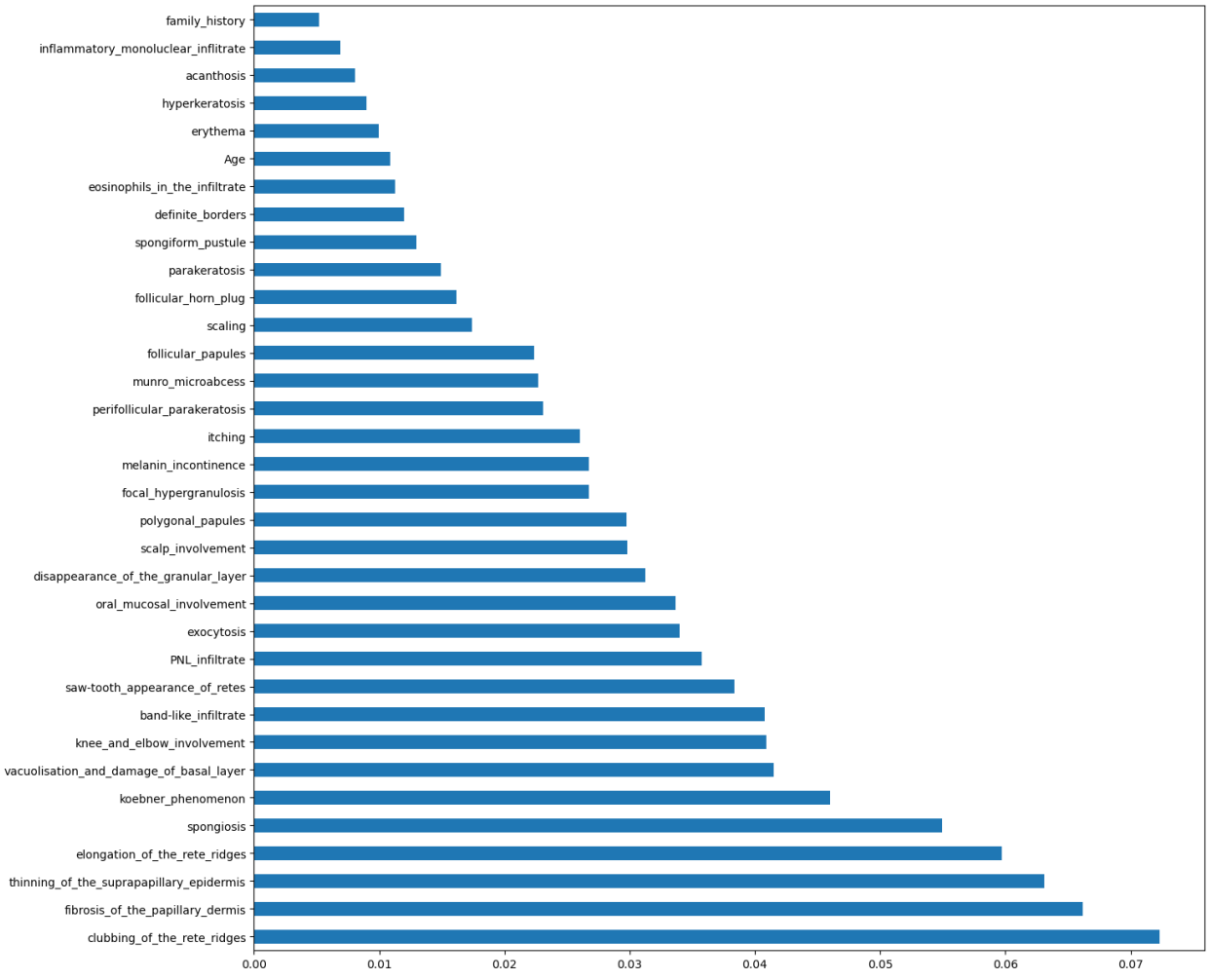
32. Inflammatory mononuclear infiltrate : Migration of mononuclear inflammatory cells.

33. Band-like infiltrate : Migration of white blood cells in band appearance.

Feature Selection

Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data. It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you are trying to solve.

Values between 1 and 6 on the x-axis of the figures represent 1-Psoriasis, 2-Seberoic dermatitis, 3-Lichen planus, 4-Pityriasis rosea, 5-Cronic dermatitis and 6-Pityriasis rubra pilaris diseases, respectively. The y-axis in the figures show the values of these diseases in the data set of the related feature.



We can deduce from the preceding figure that there are several independent features with negligible influence on the dependent features. In light of this, we will remove the features from our model.

```
#Creating new data set by removing the less important features.
final_data = data.drop(['family_history','inflammatory_monoluclear_inflitrate','acanthosis'
                        ,'hyperkeratosis','erythema','Age','eosinophils_in_the_infiltrate'
                        ,'definite_borders','spongiform_pustule','parakeratosis','follicular_horn_plug'
                        ,'scaling','follicular_papules'], axis=1)
```

Performance Analysis of Classification Methods

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.

```
y_predict = log_reg_hypertuned.predict(X_test)
print('Accuracy of Logistic Regression',accuracy_score(y_test, y_predict))
```

```
Accuracy of Logistic Regression 0.9814814814814815
```

KNeighbours Classifier

In the Kneighbours 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.

```
y_predict = modelgridsearch.predict(X_test)
print('Accuracy of K-Nearest Neighbor',accuracy_score(y_test, y_predict))#checking performance
```

```
Accuracy of K-Nearest Neighbor 0.9444444444444444
```

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

```
grid_predictions = grid.predict(X_test)
print('Accuracy of Support Vector Machine',accuracy_score(y_test, grid_predictions))#checking performance
```

```
Accuracy of Support Vector Machine 0.9444444444444444
```

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.

```
rf_clf2 = RandomForestClassifier(**rf_best_params)#passing best parameter to randomforest
rf_clf2.fit(X_train, y_train)#training
y_predict=rf_clf2.predict(X_test)#testing
print('Accuracy of Random Forest',accuracy_score(y_test, y_predict))#checking performance
```

```
Accuracy of Random Forest 0.9537037037037037
```

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 data should be corrected as much as possible and these data 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.

Logistic regression, KNeighbours Classifier, Support Vector Classification and Random Forest Classifier methods were used as classification methods. The machine learning method that provides the highest classification rate was determined.

	MLA_Names	Train_Score	Test_Score	Precision	Recall_Score
0	Logistic Regression	0.980	0.9815	0.981481	0.981481
1	Random Forest Implementation	0.992	0.9537	0.953704	0.953704
2	K-Nearest Neighbor Algorithm	0.976	0.9444	0.944444	0.944444
3	Support Vector Classification	0.992	0.9444	0.944444	0.944444

1. **Accuracy of Logistic Regression 0.9814814814814815**
2. **Accuracy of Random Forest 0.9537037037037037**
3. **Accuracy of K-Nearest Neighbor 0.9444444444444444**
4. **Accuracy of Support Vector Machine 0.9444444444444444**

Based on the observations above, we can conclude that Logistic Regression is the best-fitting model with 98% accuracy for the given problem.

GitHub Link for Source Code:-
https://github.com/VivekViradia/Skin_Disorder_Prediction