Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi

Afyon Kocatepe University Journal of Science and Engineering

AKÜ FEMÜBİD **19** (2019) 037201 (894-901) AKU J. Sci. Eng. **19** (2019) 037201 (894-901)

DOI: 10.35414/akufemubid.591816

Araştırma Makalesi / Research Article

Classification of Dermatological Data with Self Organizing Maps and Support Vector Machine

Uğur FİDAN[1*], Esma UZUNHİSARCIKLI[2], İsmail ÇALIKUŞU[3]

[1]Afyon Kocatepe Üniversitesi, Mühendislik Fakültesi, Biyomedikal Mühendisliği Bölümü, Afyonkarahisar
[2] Kayseri Üniversitesi, Meslek Yüksekokulu, Biyomedikal Cihaz Teknolojisi, Kayseri
[3] Nevşehir Hacıbektaş Veli Üniversitesi, Hacıbektaş MYO, Biyomedikal Cihaz Teknolojisi, , Nevşehir

Sorumlu yazar e-posta: ufidan@aku.edu.tr *[1]ORCID ID: http://orcid.org/0000-0003-0356-017X*
uzunhise@kayseri.edu.tr *[2]ORCID ID: http://orcid.org/0000-0003-2821-4177*
ismailcalikusu@nevsehir.edu.tr *[3]ORCID ID: http://orcid.org/0000-0002-6640-7917*

Geliş Tarihi: 14.07.2019; Kabul Tarihi: 17.10.2019

Keywords
Dermatology; Erythematous Squamous; Clustering; Classification; Self Organizing Maps; Support Vector Machine

Abstract

The frequency incidence of dermatological diseases is increasing in parallel with the fact that human skin is exposed to different chemicals. Examined many skin diseases, many of them are similar in shape and appearance, although the reasons for their appearance are different. In dermatology, the differential diagnosis of Erythemato-squamous diseases is frequently encountered by doctors. Doctors try to differentiate and diagnose diseases by evaluating clinical findings and histopathological parameters together. Many researchers have developed different algorithms on the classification and clustering of diseases and data that have been diagnosed from the UCI database. In the present study, unlike previous studies, clinical and histopathological findings of 6 different Erythamo Squamos skin diseases were clustered by applying to SOM network separately. As a result of this clustering process, it is determined that Psoriasis - Cronic Dermatitis and Seborreic Dermatitis - Pitriasis Rosea diseases were found in the same cluster and the diagnoses are confused. In order to prevent this confusion, clinical and histopathological findings of the diseases were clustered by SOM method. Clustering parameters of clinical and histopathological findings were classified with SVM. As a result of the study, it was achieved that the classification of Psoriasis - Cronic Dermatitis diseases was classified as 0.89 with an accuracy of 0.93 and that of Seborreic Dermatitis - Pitriasis Rosea with an accuracy of 0.79 and 0.80.

Dermatolojik Verilerin Öz Düzenleyici Harita ve Destek Vektör Makinaları ile Sınıflandırılması

Anahtar Kelimeler
Dermatoloji; Eritematöz Skuamöz; Kümeleme; Sınıflandırma; Öz Düzenleyici Harita; Destek Vektör Makinaları

Öz

İnsan derisinin özellikle farklı kimyasallara maruz kaldığı günümüzde dermatolojik hastalıkların görülme sıklığı da buna paralel olarak artış göstermektedir. Birçok deri hastalığı incelendiğinde birçoğu ortaya çıkış sebepleri farklı olmasına karşın şekil ve görünüş açısından benzerlik taşımaktadır. Dermatolojide, Erythemato-squamos hastalıklarına ayırt edici tanı koyulması doktorların sıkça karşılaştığı bir durumdur. Doktorlar klinik bulgular ile histopatolojik parametreleri birlikte değerlendirerek hastalıkları birbirinden ayırt etmeye ve teşhis koymaya çalışmaktadır. Konu ile ilgili birçok araştırmacı UCI veri tabanından alınan ve tanısını konmuş veriler ile hastalıkların sınıflandırılması ve kümelenmesi üzerine farklı algoritmalar geliştirmiştir. Bu çalışmada önceki çalışmalardan farklı olarak 6 farklı Erythamo Squamos deri hastalığına ait klinik ve histopatolojik bulgular SOM ağına ayrı ayrı uygulanarak kümelenmiştir. Bu kümeleme işleminin sonucunda Psoriasis - Cronic Dermatitis ve Seborreic Dermatitis - Pitriasis Rosea hastalıkları aynı küme içerisinde kaldığı ve tanıların karıştırıldığı tespit edilmiştir. Bu karışmayı önlemek için hastalıkların klinik ve histopatolojik bulguları ayrı ayrı SOM yöntemi ile kümelenmiştir. Klinik ve histopatolojik bulgulara ait kümelenme parametreleri kullanılarak SVM ile sınıflandırılma yapılmıştır. Yapılan çalışma sonucunda karıştırılan Psoriasis - Cronic Dermatitis hastalıkları arasında F1 sokuru 0.89 doğruluğu 0.93 olarak ve Seborreic Dermatitis - Pitriasis Rosea hastalıkları arasında F1 sokuru 0.79 doğruluğu 0.80 olarak sınıflandırma başarımı sağlanmıştır.

1. Introduction

The frequency of dermatological diseases increases in parallel with the recent exposure of human skin to different chemicals. It is clear that most of these are similar in their shapes and appearances although they are originally different from each other. In dermatology, diagnosing of Erythematous-Squamous skin diseases is a problem frequently encountered unclear by doctors(Martinelli, El Hachem, Bertini, & Dionisi-Vici, 2017). Doctors generally try to differentiate and diagnose diseases by evaluating clinical and pathological parameters together. With the advancements in the biomedical science, researchers are trying to help dermatologists in diagnosing the disease by developing various algorithms in the classification of these diseases(Esteva et al., 2017; Karaca, Sertbaş, & Bayrak, 2018; Xie, Ji, & Wang, 2018) However, only three studies on SOM and SVM have been considered here. In these studies, in 2014 Haryanto and colleagues used the Self-Organizing Map (SOM) method to identify Erythematous-Squamous dermatological diseases by working on the same dataset, which included 231 data sets used for classification purposes as 30 training and 201 test data. The SOM method was used for each of six class of der datas with som-svm different dermatological diseases for nine times and then the accuracy percentages of experience for each disease were determined. As a result, each patient had the best accuracy in the experience using the SOM structure(Haryanto et al., 2015). In 2016, the study conducted by Fidan U. and et al. 6 different Erythematous-Squamous skin diseases clustered as clinically and pathologically separately differently from the previous studies. The study showed that the diagnosis of Psoriasis - Chronic Dermatitis and Seborrheic Dermatitis - Pityriasis Rosea confused each other due to being found on the same clustering. For this reason, the results of SOM clustering were obtained using clinical and pathological parameters separately(Fidan, Ozkan, & Calikusu, 2016). Karaca Y. and his colleagues in 2018 developed a wavelet, probabilistic and information data set by applying 1-D continuous wavelet coefficient analysis to the data set, Basic Component Analysis and Linear Discriminant Analysis to distinguish the important characteristics of the dermatological data set for successful classification. Then, by applying the Support Vector Machine kernel algorithms (Linear, Quadratic, Cubic, Gaussian) to these data sets, the accuracy ratios were obtained. Finally, the wavelet data set was obtained with the highest accuracy (Karaca, Sertbaş et al. 2018).

In the present study, clinically and histopathologic findings of six different Erythematous-Squamous skin diseases were clustered separately and differentiated in groups and identified as a result of clustering. The mixed disease groups were then attempted to be classified by the SVM method in the way of classifying as clinical findings in one axis histopathological finding as another axis. In this respect, it was expected that the classification success increased in the disease groups, which were mixed in classification.

2. Materials and Methods

The data used in the current study were taken from 357 different adults and children having six types of Erythematous-Squamous dermatologic diseases ranging from 7 to 75 years from the UCI database. The percentages of distribution for these skin diseases are shown in Fig 1.

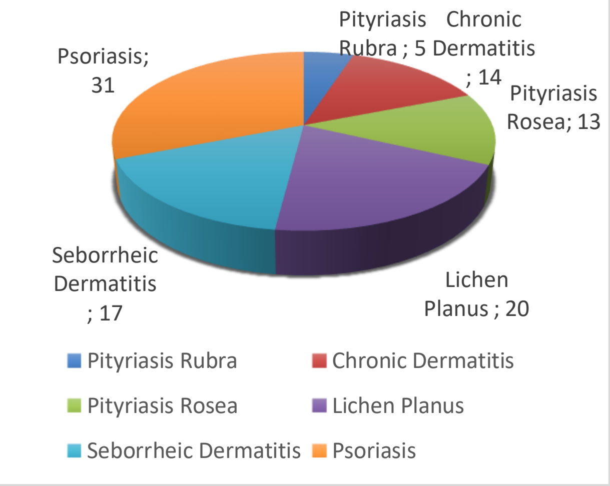


Fig 1. Percentage distribution of dermatologic patients in the used data

A total of 34 parameters were used as the evaluation criteria for classification and clustering of six different Erythemato-Squamos diseases. These parameters consist of 22 histopathological and 12 clinical findings for the diagnosing of Erythematous-Squamous skin diseases (Djamaludin, Haryanto, & Hasim, 2018)

2.1 Self organizing map (SOM)

SOM is an artificial neural network structure that places input data into a geometric structure and expresses it with a limited number of units. Updating the weight vectors that define these units in the neighbours' relation of the units arranged in the geometric structure is obtained. The SOM differs from other unsupervising learning schemes because it maintains the topological properties of the input space with its self-organizing structure and neighbourhood function. Thus, the topological properties of the input space formed by the data are transferred to the one- or two-dimensional geometric structure(D'Urso, De Giovanni, & Massari, 2019).

2.2 SOM network algorithm.

With the introduction of the data in the training cluster, the weights of the neurons are updated accordingly and here clusters of the neurons constituting the SOM network represent the clusters in each update. Since the weights of neighbouring neurons change in a similar way, the properties of the input space in which the data are stored are topologically conserved and transformed into discrete neuron space. The SOM flow diagram (Fig. 2) consists of assigning the number of iterations, determining the neighbours, presenting the data, determining the winning neuron, and updating the neighbourhood of the winning neuron. Firstly, weight vectors wj ϵ Rm, j=1,….,l are randomly assigned. Thus, the first value of l neurons that will fall against the centres of the data is determined. During the determination of neighbourhoods, n neurons are placed in a desired one or two-dimensional geometry to determine the neighbourhood pattern and the physical distance between each neuron which is calculated by Equation 1.

$$d_{ij}^{distance} = \| d_i - d_j \| \text{, j=1…, } l \text{, i=1…, } l \qquad (1)$$

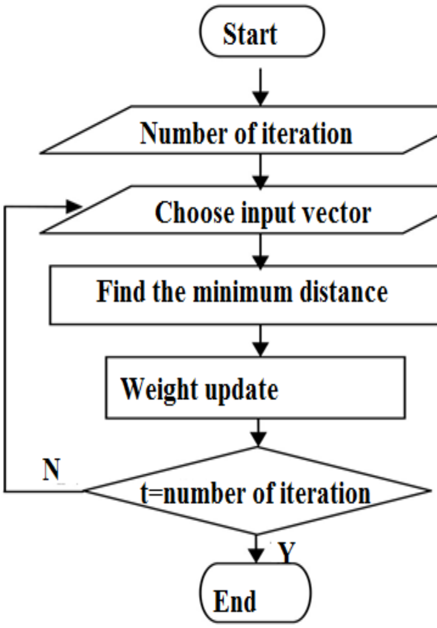


Fig 2. Flow diagram of the classical SOM algorithm

From the training set, a data set vector is selected such that xk ∈ Rm, k = 1, ..., p. To determine the winner neuron, kth data which is the most similar of the l neurons is determined with Equation 2 and the winning neuron index for k. data with the Equation 3.

$$d_{ki}^{similarity} = \left(\sum_{j=1}^{m}(x_{kj} - w_{ij})^2\right)^{1/2}, i = 1, \dots l \qquad (2)$$

$$index^{winner}(k) = min_i\{d_{ki}^{similarity}\}, i = 1, \dots l \qquad (3)$$

The weight of the winning neuron neighbourhoods according to the neighbourhoods function Q(d_ij^distance,n) determined to be the largest winning neuron weight is updated with Equation 4.

$$\mathrm{w(n+1) = w(n) + Q}\left(d_{ij}^{distance}, n\right)\mu(n)\left(x(n) - w(n)\right) \qquad (4)$$

Here, μ (n) corresponding to the learning rate is expressed as in Equation 5 and exponentially decreases over successive incremental updates.

$$\mu(n) = \mu_0 \exp\left(\frac{-n}{\lambda}\right) \qquad (5)$$

Clustering occurred when the same data began to acquire the same neurons in the training set. To acquire a representation feature of an active cluster, the neurons corresponding to cluster centers are maintained for four times the number of steps taken until they reached that stage(Shen et al., 2016)

2.3 Support vector machine (SVM)

The SVM algorithm, a supervised statistical learning method developed by Vapnik (1995) for solving classification and curve fitting problems is based on the principle of minimizing the structural risk. The purpose of the SVM is to separate the input data whose classes are defined by class label into two separate classes by specifying the optimal hyperplane (Fig.3)(Yang, Liu, Li, Li, & Ma, 2015).

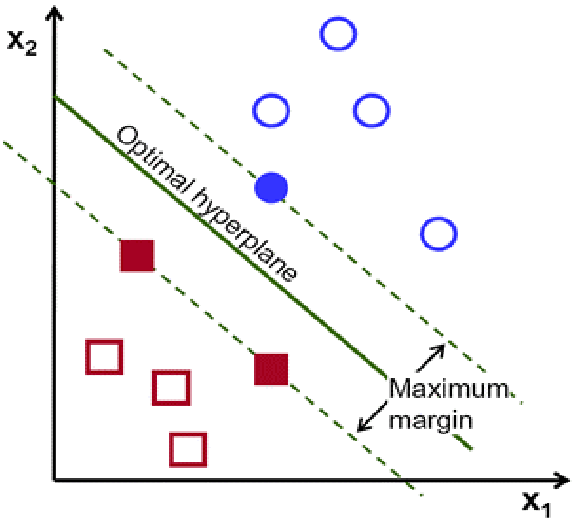
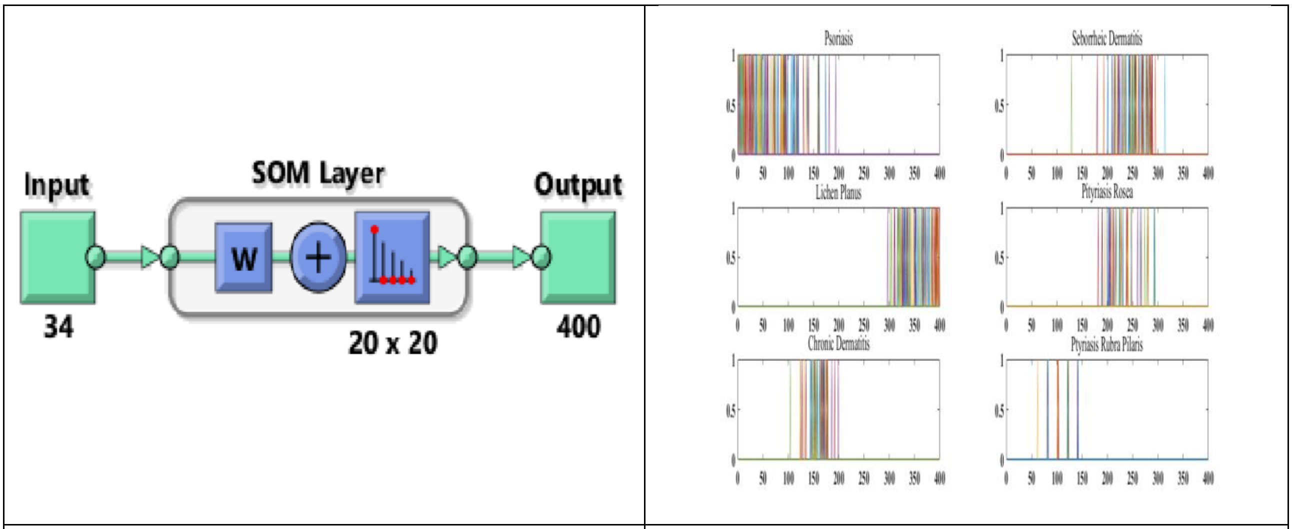


Fig 3. Optimal Hyperplane and Margin of SVM

A linear SVM classification was used in this study. If a set of training data consisting of N elements is assumed to be $K = \{(x_i, y_i), i = 1 \ldots N\}$ then $y_i \in \{-1, +1\}\}$ class label, $X_i \in R^n$ and any data in n dimensional space. In the function $f(x) = w^Tx + b$, w^T is the normal of the decision function, x is the point on this line, and b show the bias. The aim is to train the system by finding w^T and b with the help of training data. In all SVMs, the purpose is to separate a data group into two different classes. In Fig.3, the vectors on the lines denoted by the broken lines are called support vector and the soft discrimination line passes over these vectors. The middle of the two soft segment lines represents the hard separation and is plotted as $f(x) = w^Tx + b = 0$. w^T and x are the vector magnitudes in the expression, $f(x) = w^Tx + b$ where $y_i = 1$ in the case of $f(x) = w^Tx + b \leq 1$ and $y_i = -1$ in the case of $f(x) = w^Tx + b \leq 1$. It can be shortened to two functions by a single equation $y_i(w^Tx + b) \geq 1$. The most important point of this separation is to have the best separation by making the maximum value of the limit. Although it is possible to draw infinite number of multiple planes that can divide the data set into classes, the goal is to select the hyperplane to make the classification error had the smallest value when encountering an unknown set of data. For this purpose, the maximum limited technique is proposed. The size of the limit value increases the generalization ability. x_1 is a point on the function $f(x) = w^Tx + b = 1$ and x_3 value is another point on the function $f(x) = w^Tx + b = 1$. To find the limit value, if $w^Tx_1 + b = +1$ is multiplied by (-1) and $w^Tx_3 + b = -1$ is added and $x_1 = x_3 + \lambda w$ is written in the equation, then $\lambda = {}^{2}/_{w^2}$ expression is obtained. Since the target is to maximize the value of λ, ${}^{1}/_{\lambda}$ must be the minimum one. Therefore, the limitation is$y_i(w^Tx + b) - 1 \geq 0, y_i \in \{-1, +1\}$. The task of the optimization problem is to find the maximum or minimum of a function under given constraints and to perform this operation in minimum time (Deo, Wen, & Qi, 2016; Guenther & Schonlau, 2016; Suthaharan, 2016).

3. Results

Clinical and histopathological findings of 300 patients with six different erythematous-squamous diseases were applied with 34 different parameters input to a 20 × 20 SOM network (Figure 4a). The data of 300 patients who were diagnosed according to clinical and pathological findings were applied to 20x20 SOM network as shown in Figure 4.a. Each disease group was observed to be clustered in certain regions of the SOM map.



a) 20x20 SOM Network

b) Classification of clinical and histopathological parameters together.

Fig 4. Clustering of Erythemato-Squamos diseases using clinical and histopathological findings with SOM.

Clinical and pathological findings were applied separately to the SOM network to classify the diseases to overcome this problem. First, 12 clinical parameters were clustered with 20x20 SOM network. As a result of clustering using clinical parameters, diseases clustered as a result of SOM network as in Fig 5. It can be seen here that the diagnoses of Psoriasis - Chronic Dermatitis and Seborrheic Dermatitis - Pityriasis Rosea diseases, which are confused with each other, become more understandable.

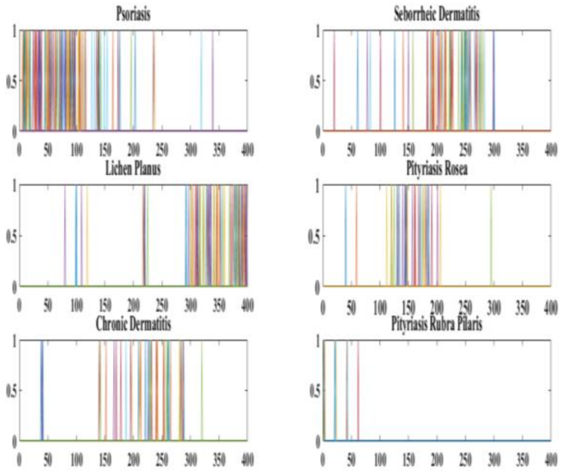


Fig 5. Clustering of six types diseases according to clinical parameters by SOM.

Similarly, 22 different histopathological findings used in the diagnosis of diseases were clustered with a 20x20 SOM network. As a result of this clustering, it is seen in Figure 6 that Psoriasis - Chronic dermatitis diseases clustered in different regions of the SOM network, while Seborrheic Dermatitis - Pityriasis Rosea diseases clustered in the same SOM network regions.

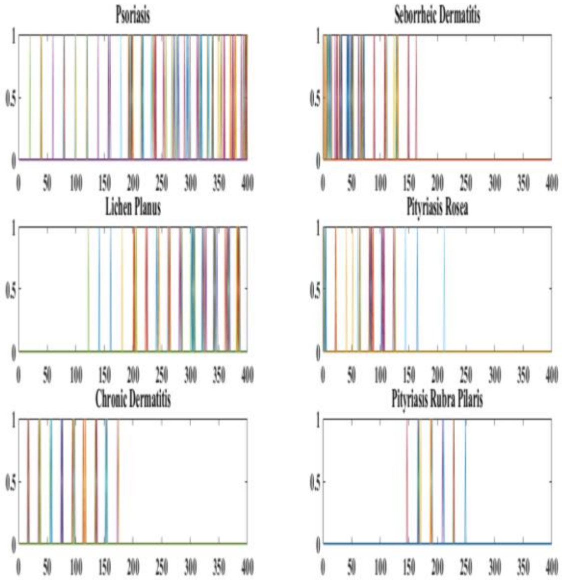


Fig 6. Clustering of diseases according to histopathologic parameters by SOM.

In conclusion, it can be concluded that Psoriasis - Chronic Dermatitis and Seborrheic Dermatitis - Pityriasis Rosea diseases can not be distinguished completely from clinical or histopathologic parameters only when Figure 4, Figure 5 and Figure 6 are taken together. For this purpose, SVM(using Matlab Toolbox) was used to distinguish between these two disease groups. The values obtained according to clinical and histopathological parameters were used as input parameters of SVM. In this classification, clinical SOM results were classified on one axis and histopathologic results on the other axis.

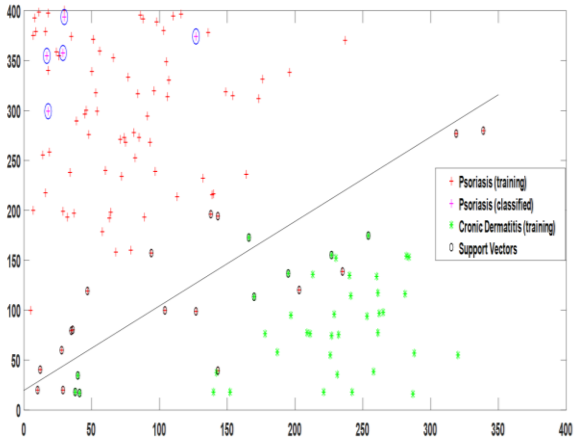
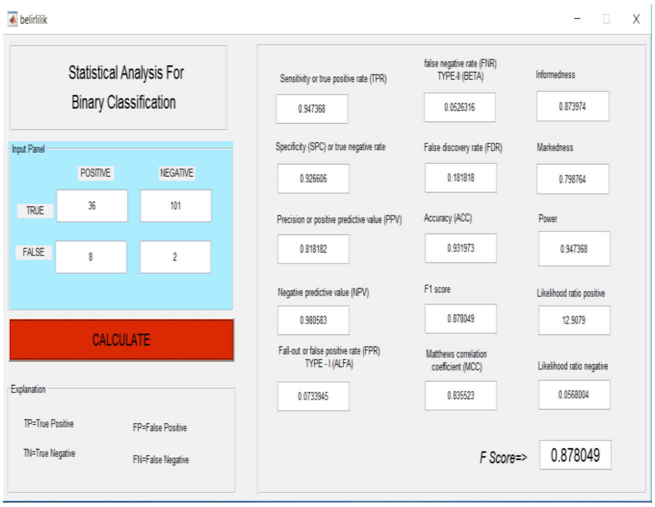


Fig 7. Classification of Psoriasis - Chronic Dermatitis with SVM.

A binary classification test is used to determine the success of the classification. For this purpose, 103 Psoriasis and 44 Chronic Dermatitis patients is used as test data. The statistical test results of the classification are shown in Fig. In the classification, the level of significance between Psoriasis and chronic dermatitis diseases (F1 score) was 0.89 and accuracy was 0.93.



Statistical Analysis For Binary Classification

Input Panel

	POSITIVE	NEGATIVE
TRUE	36	101
FALSE	8	2

Explanation

TP=True Positive FP=False Positive

TN=True Negative FN=False Negative

Sensitivity or true positive rate (TPR) 0.947368	false negative rate (FNR) TYPE-II (BETA) 0.0526316	Informedness 0.873974
Specificity (SPC) or true negative rate 0.926606	False discovery rate (FDR) 0.181818	Markedness 0.798764
Precision or positive predictive value (PPV) 0.818182	Accuracy (ACC) 0.931973	Power 0.947368
Negative predictive value (NPV) 0.980583	F1 score 0.878049	Likelihood ratio positive 12.9079
Fall-out or false positive rate (FPR) TYPE - I (ALFA) 0.0733945	Matthews correlation coefficient (MCC) 0.835523	Likelihood ratio negative 0.0568004

F Score=> 0.878049

Fig 8. Statistical analysis of Psoriasis - Chronic Dermatitis classification results.

Fig 9 shows the classification of Seborrheic Dermatitis - Pityriasis Rosea by SVM algorithm. The horizontal axis shows the clustering locations in the SOM network of clinical findings while the histopathological findings of vertical axis diseases indicate the clustering locations in the SOM network. Fig.9 shows that clinical features of Seborrheic Dermatitis - Pityriasis Rosea are more effective than histopathological findings in differential diagnosis.

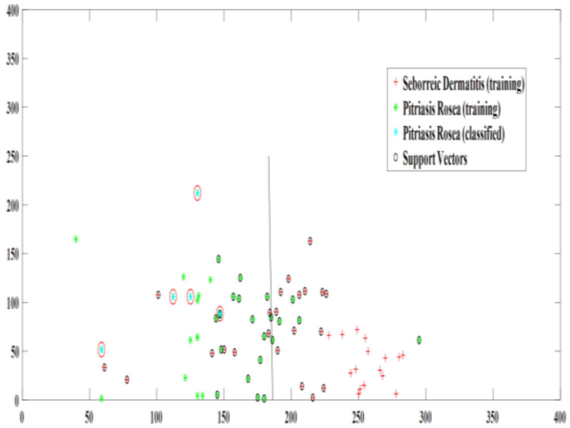
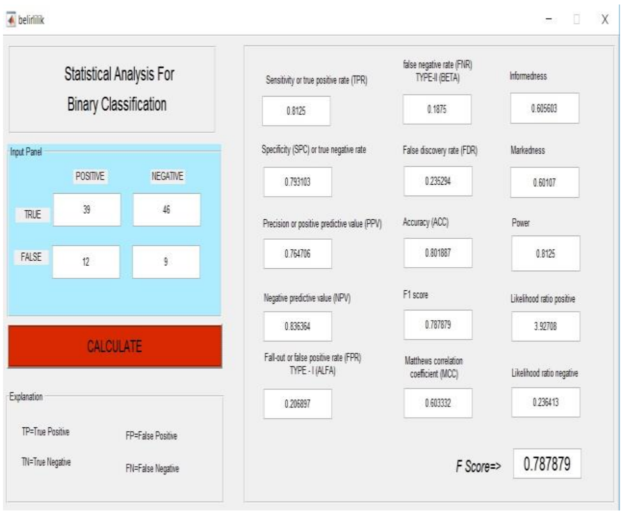


Fig 9. Classification of Seborrheic Dermatitis - Pityriasis Rosea with SVM.

55 Seborrheic Dermatitis and 51 Pityriasis Rosea disease data were used in the test data to determine the success of the binary classification. The statistical significance of the classification is shown in Fig.10. The level of significance (F1 score) between Seborrheic Dermatitis - Pityriasis Rosea diseases was 0.79 while the accuracy was 0.80.



Statistical Analysis For Binary Classification

	POSITIVE	NEGATIVE
TRUE	39	46
FALSE	12	9

Sensitivity or true positive rate (TPR): 0.8125
false negative rate (FNR) TYPE-II (BETA): 0.1875
Informedness: 0.605603
Specificity (SPC) or true negative rate: 0.793103
False discovery rate (FDR): 0.235294
Markedness: 0.60107
Precision or positive predictive value (PPV): 0.764706
Accuracy (ACC): 0.801887
Power: 0.8125
Negative predictive value (NPV): 0.836364
F1 score: 0.787879
Likelihood ratio positive: 3.92708
Fall-out or false positive rate (FPR) TYPE - I (ALFA): 0.206897
Matthews correlation coeficient (MCC): 0.603332
Likelihood ratio negative: 0.236413

TP=True Positive
FP=False Positive
TN=True Negative
FN=False Negative

F Score=> 0.787879

Fig 10. Statistical analysis of Seborrheic Dermatitis-Pityriasis Rosea classification results.

Classification accuracy seemed to be between 90% and 98.62% when the clinical and histopathologic findings are evaluated together in the previous studies. In the current study, firstly, confusing skin disease groups were determined using clinical and histopathological findings separately, and these data were classified with classification algorithms.

4. Discussion and Conclusion

Nowadays, the frequency of dermatological diseases has been increasing in parallel with the recent exposure of human skin to different chemicals. When we examined many skin diseases, most of them have similar shapes and appearances although the reasons for their emergency are different. In dermatology, differential diagnosis of Erythematous-Squamous diseases is a frequently encountered problem for doctors. Doctors try to differentiate and diagnose diseases by evaluating clinical and histopathological parameters together.

When the studies done with this disease group in the literature are examined, it is generally observed that classifications are obtained by using all clinical and pathological findings together. In the literature, although the other studies have the highest success rate, the evaluation of clinical and pathological together that decreases the reliability of the diagnosis(D’Urso et al., 2019; Haryanto et al., 2015). On the other hand, this makes difficult to diagnosis of skin diseases. For this purpose, in this study clinical and histopathological findings of six different Erythematous-Squamous skin diseases were clustered separately with SOM network different from previous studies. As a result of this clustering, the findings of Psoriasis - Chronic Dermatitis and Seborrheic Dermatitis - Pityriasis Rosea disease pairs were observed in the same cluster. In real life, it has been understood that doctors may have difficulty in recognizing these disease couples. To achieve this confusion, the clustering parameters of clinical and histopathological findings were used to classify with SVM. If the classification accuracy is low, the hyperplane in the SVM algorithm is linear and only the study of confused groups is effective. For this reason, it is seen that an algorithm can increase the accuracies with high selectivity such as Kernel SVM. Future studies will attempt to improve the classification and clustering accuracy with appropriate optimization techniques such as PSO and ABC.

5. References

D’Urso, P., De Giovanni, L., & Massari, R., 2019. Smoothed Self-Organizing Map for robust clustering. *Information Sciences*, **3,** 7-51

Deo, R. C., Wen, X., & Qi, F.,2016. A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset. *Applied Energy,* **168,** 568-593.

Djamaludin, D., Haryanto, H., & Hasim, Y. K., 2018. Expert System Of Dental And Diagnosis Diseases Using Forward Chaining Method Based Android. *Paper Presented At The International Seminar on Education and Development of Asia,* **1,** 37-42.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature,* **542,** 115-118.

Fidan, U., Ozkan, N., & Calikusu, I., 2016. Clustering and classification of dermatologic data with Self Organization Map (SOM) method. *Medical Technologies National Congress*, Tıptekno**,** 1-4.

Guenther, N., & Schonlau, M., 2016. Support vector machines. *The Stata Journal,* **16(4),** 917-937.

Haryanto, H., Ulum, M., Rahmawati, D. R., Joni, K., Ubaidillah, A., Alfita, R., Khotimah, B. K., 2015. The Erythemato-Squamous Dermatology Diseases Severity Determination using Self-Organizing Map. *IPTEK Journal of Proceedings Series,* **1,** 279-284.

Karaca, Y., Sertbaş, A., & Bayrak, Ş., 2018. Classification of Erythematous-Squamous Skin Diseases Through SVM Kernels and Identification of Features with 1-D Continuous Wavelet Coefficient. *International Conference on Computational Science and Its Applications,* Iccsa *107-120.*

Martinelli, D., El Hachem, M., Bertini, E., & Dionisi-Vici, C., 2017. Skin and Hair Disorders 31. Springer, 341-370

Shen, J., Chen, P., Su, L., Shi, T., Tang, Z., & Liao, G., 2016. X-ray inspection of TSV defects with self-organizing map network and Otsu algorithm. *Microelectronics Reliability,* **67,** 129-134.

Suthaharan, S., 2016. Support vector machine Machine learning models and algorithms for big data classification, 36. Springer, 207-235.

Xie, J., Ji, X., & Wang, M., 2018. Extreme Learning Machine Based Diagnosis Models for Erythemato-Squamous*. International Conference on Health Information Science*, **11148,** 61-74

Yang, D., Liu, Y., Li, S., Li, X., & Ma, L., 2015. Gear fault diagnosis based on support vector machine optimized by artificial bee colony algorithm. *Mechanism and Machine Theory,* **90,** 219-229.