JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 08, 2020

Diagnosis of Squamous Erythematosus Using KNN and Genetic Algorithms

Soraya Gharavi*

Esfarayen Higher Education Complex, North Khorasan, Esfarayen
Instructor, Faculty of Computer and Electrical Engineering, Department of Computer
Bojnourd, Nader St., Daqiqi Alley, No. 32, Postal code: 9414965395, ORCID: 9065-7451-0001-0000
Iran, gharravi@esfarayen.ac.ir

ABSTRACT

Background and purpose: In Dermatology, the correct prediction of Erythemato-Squamous disease is really significant. The presence of different signs and features of the disease has made it difficult for physicians to diagnose. Data mining allows the analysis of the patients' clinical data for medical decisions. The purpose of the paper is to present an accurate model for predicting Erythemato-Squamous disease.
Materials and methods: In this study, the medical cases of 366 patients with Erythemato-Squamous disease have been studied with 12 features related to clinical findings and 22 pathological features. Patients' information was selected from the Standard database of California University. Genetic Algorithm and Data mining were used to present a model for predicting Erythemato-Squamous disease.
Results: The proposed model was compared with the Decision Tree, Naïve Bayes and Nearest Neighbor methods. The results show that the prediction accuracy of the proposed model was 0.992. Also, for Naïve Bayes, Decision Tree and Nearest Neighbor methods, the prediction accuracy equals 0.937, 0.948 and 0.970, respectively.
Conclusion: In the prediction of Erythemato-Squamous disease, the proposed model includes the minimum error and the maximum accuracy and validation compared to the other models. The Naïve Bayes method has the maximum error and minimum accuracy.
Keywords: Erythemato-Squamous disease, Genetic Algorithm, Data Mining, Prediction.

1. INTRODUCTION

Dermatology is associated with the diagnosis and treatments of the skin diseases [1]. Skin diseases and its adnexa including hair, nails, sweat glands and membranes of the oral and external genitalia belong to the subset of dermatology. Differential diagnosis of squamous erythematosus is one of the most important issues in dermatology. There is little difference between squamous erythematosus diseases regarding size and clinical features. Because of many similarities in the clinical and laboratory symptoms of squamous erythematosus diseases the likelihood of misdiagnosis will increase. Psoriasis, Seboreic Dermatitis, Lichen Planus, Pityriasis Rosea, Croni Dermatitis and Pityriasis Rubra Pilaris are among the subsets of squamous erythematosus diseases [2].

Psoriasis will stop the life cycle of skin cells. It increases the rate of cell formation on the skin surface and this will lead the formation of thick, itchy or dry red or silver patches [3]. Seboreic Dermatitis will spread in different body regions in the form of a scaly red rash with itching. The scalp, side of the nose, eyebrows, eyelids, behind the ears and middle of the chest are the most common regions involved [4].

Lichen Planus will genetically develop having a chronic itch on the skin. The disease will spread in the form of Papule on the wrists, feet, skin of the trunk and genitals [5]. Pityriasis Rosea is benign and develop in the abdomen and chest in the form of scaly pimples in adults [6]. Croni Dermatitis will spread in the form of long-term inflammation and irritation that will be developed darker and thicker around the skin [7]. Pityriasis Rubra Pilaris s is a chronic inflammatory disease of the skin that is less common than other squamous erythematosus diseases [8].

Because of many similarities in the clinical and laboratory symptoms of squamous erythematosus diseases the likelihood of misdiagnosis will increase. It is possible to diagnosis and prognosis of various diseases using data mining techniques. In medicine, data mining means the process of extracting valid, previously unknown, understandable and reliable information from a medical database and using it to prognosis, diagnose and assist in the treatment of disease. One of the applications of data mining in medicine is the discovery of useful patterns between disease and clinical and laboratory symptoms of the patient. A useful pattern is a model in data, which expresses the relationship between a subset of patient data and disease diagnosis [9].

In [10], we have suggested the use of SVM with a hybrid approach to select the features in order to diagnose the squamous erythematosus. In this method, the accuracy of the diagnosis is increased by filtering the inappropriate features. In [11], Adaptive Neuro-Fuzzy Inference Systems (ANFIS) to diagnose squamous erythematosus was introduced. In this method, ANFIS classifier was used to increase the accuracy of

ISSN- 2394-5125 VOL 7, ISSUE 08, 2020

diagnosis. Soltani et al. have proposed the Nearest Neighbor Approach to diagnose automatically the squamous erythematosus. This method was presented for data mining with the aim classifying and the differential diagnosis of the disease [12].

The accuracy of diagnosis and prognosis of data mining methods can be increased by using genetic algorithms [13]. Genetic algorithm is one of the subgroups of meta-heuristic computing, which follows the laws of natural biological evolution. Using the Law of the Survival of the fittest in a subset of problem answers, the genetic algorithm seeks to find the best answers. The initial population is made by a subset of possible answers to the problem. Proportional to the value of each answer, the selection process from the initial population and reproduction is done to create a new generation.

In each generation, by combining and reproducing the selected responses with the help of agents that follow natural genetics, better approximations are obtained from the final answer. This process makes the new generations more adapted to the problem conditions [13].

2 Materials and Methods

1.2 Datasets Selections

We have selected the patients' medical records and dermatology datasets from the standard database of the University of California Irvine (UCI) for data mining [14]. This dataset contains 336 samples, among which 8 samples do not have complete information. 34 features are registered for each patient, 12 features are related to the clinical findings and 22 are the pathological features. In Table 1, the values of each feature are shown. It is worth mentioning that the samples have no complete information were estimated using the maximum frequency method. Range of initial values of patient clinical features is presented in Table 2. Table 3 shows the features of the patient's pathological information. The class names and the number of cases in each class are shown in

Table 1: Description of the values of clinical and pathological features.

Feature value	Description
0	No complication
1	Complication with low probability
2	Complication with moderate probability
3	High possibility of a complication
?	No value recorded.

Table 2: Range of initial values of patient clinical features.

Clinical Features	The range of initial values
Family History	0-does not have
Age	[7-75]
Erythema	{0،1،2،3}
Scaling	{0،1،2،3}
Definite Borders	{0،1،2،3}
Itching	{0،1،2،3}
Koebner Phenomenon	{0،1،2،3}
Polygonal Papules	{0،1،2،3}
Follicular Papules	{0،1،2،3}
Oral Muscosal Involvement	{0،1،2،3}
Knee And Elbow Involvement	{0،1،2،3}
Scalp Involvement	{0،1،2،3}

Table 3: characteristics of pathological information.

Features
1. Melanin Incontinence 1-has it
2. Eosinophils in the Infiltrate
3. PNL Infiltrate
4. Fibrosis of the Papillary Dermis
5. Exocytosis
6. Acanthosis
7. Hyperkeratosis
8. Parakeratosis

ISSN- 2394-5125 VOL 7, ISSUE 08, 2020

9. Clubbing of the Rete Ridges
10. Elongation of the Rete Ridges
11. Thinning of the Suprapapillary Epidermis
12. Spongiform Pustule
13. Munro Microabcess
14. Focal Hypergranulosis
15. Disappearance of the Granular Layer
16. Vacuolisation and Damage of Basal Layer
17. Spongiosis
18. Saw-Tooth Appearance of Retes
19. Follicular Horn Plug
20. Perifollicular Parakeratosis
21. Inflammatory Monoluclear Inflitrate
22. Band-Like Infiltrate

Appropriate data format as data mining input will influence on the results and output. If the dataset features' values are in a different range, the probability of error in the findings increases. Normalization means "putting the data of a statistical community in a similar range" [15]. Normalization is done by Max / Min method and in the range of [1-0] [16].

Table 4: names and number of class instances.

Class Name	Number Of Sample
Pesoriasis	112
Seboriec Dermatitis	61
Lichen Planus	72
Pityriasis Rosea	49
Croni Dermatitis	52
Pityriasis Rubra Pilaris	20

1.3 Genetic Algorithm

The genetic algorithm was proposed by John Holland in 1962. This algorithm belongs to the group of random optimization algorithms and is suitable for optimizing complex problems with unknown search space [17].

The main idea of the genetic algorithm is taken from Darwin's theory of evolution. Darwin's theory states that those natural features that are more compatible with natural laws have a better chance of survival. It is worth noting that; but it has been confirmed empirically and statistically [18].

New members of a society are created through procreation. The chance of a person surviving in the new generation depends on the specific chromosomal composition. During the procreation stages, there may be mutations in the traits of a new generation and as a result a person with excellent traits and high compatibility is created. During the process of procreation, the best species in each generation are allowed to reproduce and the undesirable species will gradually disappear and people of new generations will evolve over the time. The genetic algorithm is summarized below.

1. Random sets of answer candidates are generated as the initial population, and are replaced by new candidates in each generation.
2. In each reiteration of the algorithm, the population is evaluated by the fitness function. Then some of the best candidates are selected for the next generation and the new population will be formed.
3. A number of this population is used to produce new offspring using genetic operators such as Crossover and Mutation.
4. The above steps continue until an appropriate answer is reached.

In Figure 1, the steps for implementing the genetic algorithm are presented in the form of a routine.

1.4 The Proposed Model

Each of the features and findings is of particular importance in the diagnosis and prognosis of skin diseases. In other words, not all features have the same value. For example, in diagnosing the disease, two features of itching in some part of the body and the patient's family history are of different importance.

ISSN- 2394-5125 VOL 7, ISSUE 08, 2020

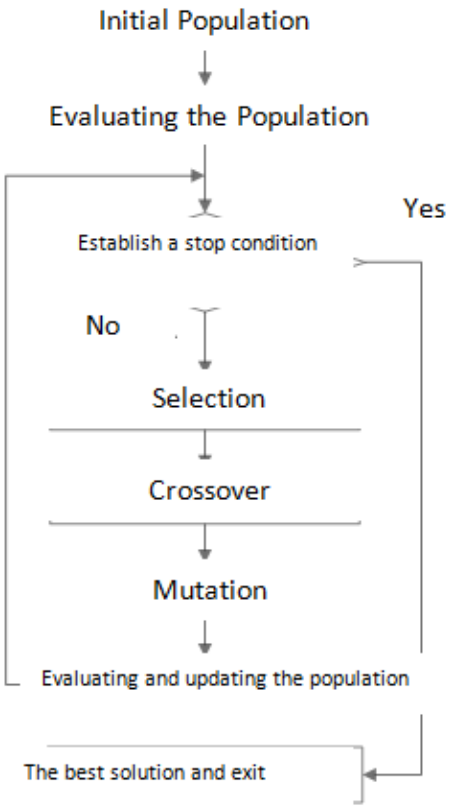

Figure 1: Genetic Algorithm Flowchart

In the proposed method, the value and role of each feature will be precisely determined by the genetic algorithm to diagnose the disease. A Gene is defined for each feature and a number in the range of [1-0] is randomly assigned to quantify gene, which indicates the importance of the feature corresponding to that gene. The larger the value of the gene is, the greater the value and importance of the corresponding feature will be. Table 6 suggests how to weight the features based on the corresponding gene.
For each chromosome in the population, the following steps are repeated:

1. The dataset is called.
2. A record is selected from the dataset.
3. The values of the record's features are normalized.
4. The value of gene associated with each feature is extracted from the chromosome.
5. Correspondingly, each feature is multiplied in its own gene to weight the feature value.
6. Steps one to five are repeated for all records.
7. Based on the weighted values of the features, the classification is performed.

Table 6: names and number of class instances.

Property		Initial Value	Normal Value	The Amount O Gene	Weighed Amount
1	Erythema	2	0.66	0.3	0.18
2	Scaling	3	1	0.8	0.8
3	Definite Borders	1	0.33	0.2	0.066
-	-	-	-	-	-
34	Band-Like Infiltrate	3	1	0.7	0.7

The fitness function indicates the suitability or capability of each chromosome. Assuming the variables defined in Table 7,the fitness function is shown in Figure 2.

ISSN- 2394-5125 VOL 7, ISSUE 08, 2020

ALGORITHM

1) Read the training data from a file
2) Read the testing data from a file
3) Normalize the attribute values in the range of 0 to 1.
4) Let $x_1, x_2 \ldots . x_m$ denote the m instances from data set
$\{f_1, f_2, \ldots, f_n\}$, n = Number of features
5) Let *Ch* denote the current chromosome from population
$\{g_1, g_2, \ldots, g_r\}$, r = Number of genes
6) Assign weight *Ch* to each instance x_i in the training set
7) Train the weights on the whole training data set

For every training instance

- Calculate the weighted value as
- $Ch_j * x_{ij}$, where j is the attribute
- Find the K nearest neighbors based on the Euclidean distance
- Calculate the class value

End for

8) For each testing instance in the testing data set
- Find the K nearest neighbors in the training data set based on the Euclidean distance
- Predict the class value by finding the maximum class represented in the K nearest neighbors

End for

9) Calculate the error rate as

Error Rate =
*1- (# of correctly classified examples / All) * 100*

10) Fitness Function=Minimize (*Error Rate*)

Figure 2: Fit Function

The classification with the values of weighted features, which are calculated based on each of the chromosomes, is performed by the nearest neighbor method [19]. Any chromosome that performs the classification with less prognosis error will be more suitable. Error estimation based on sampling method is done using 10-Fold Cross Validation [20].

Selection operator, will select a number of chromosomes from the population to reproduce. In this model, the Elitist Selection is used. In this method, the best chromosomes in each generation are selected for reproduction. Elites' selection will considerably increase the efficiency of genetic algorithm [21]. In the Crossover method, sections of chromosomes are randomly combined and a new child is born. This causes the children to inherit the characteristics of their parents. Figure 3 shows how to reproduce a new child.

1-	The value of the r vector in dimensions (34*1) is randomly determined in the range of [0-1]
2-	The two chromosomes, Ch_1 and Ch_2, are randomly selected from the population.
3-	$r * Ch_{1 + (1-r) *} Ch_2$ = First new child
4-	$(1-r) * Ch_1 + r * Ch2$ = Second new child

Figure 3: How to reproduce a new child

The mutation operator r is applied to new children after the crossover is completed. This operator will select a gene from a chromosome and then changes the content of that gene. Figure 4 shows how to mutate.

1-	The r vector r is defined in the dimensions of a chromosome (34 × 1).
2-	The value of the r vector is determined by the normal distribution in the range [-1.1].
3-	A chromosome is randomly selected from the population.
4-	A mutated chromosome is created from the sum up of selected chromosomes with r.

Figure 4: How to apply the mutation operator

When the mutation is finished, the produced chromosomes are considered as a new generation. After several generations, the chromosome values converge and the final answer is obtained. Figure5 shows the error percentage for disease prognosis in the proposed algorithm in 40 generations.

Data in patients' records were described, simulated and analyzed using Matlab software (R2013b, The Mathworks Inc., USA).

ISSN- 2394-5125 VOL 7, ISSUE 08, 2020

To compare the proposed model with other methods, the criteria of Accuracy, Sensitivity, Specificity, Precision and F-Measure are used according to the following relations [16]:

$$Accuracy = (TP + TN)/All, \quad (1)$$
$$Sensitivity = TP/(TP + FN), \quad (2)$$
$$Specificity = TN/(FP + TN), \quad (3)$$
$$Precision = TP/(TP + FP), \quad (4)$$
$$Recall = TP/(TP + FN), \quad (5)$$
$$F\ Measure\ = \frac{2 \times Precision \times Recall}{Precision + Recall}. \quad (6)$$

Figure 7 shows a graph of the results of the diagnosis of different methods with the Accuracy criterion. As it can be seen, the proposed model is more accurate than other methods. Comparison of disease prognosis results with Sensitivity and Specificity criteria is also shown in Tables 11 and 12, respectively. In Tables 13 and 14, the results of the methods are compared with the Precision and F-Measure criteria, respectively. The comparison results show the performance superiority of the proposed model. In Table 15, the proposed method is compared with other methods by the criteria of F-Measure, Precision, Sensitivity, Specifity. The values in the table indicate better performance of the proposed method.

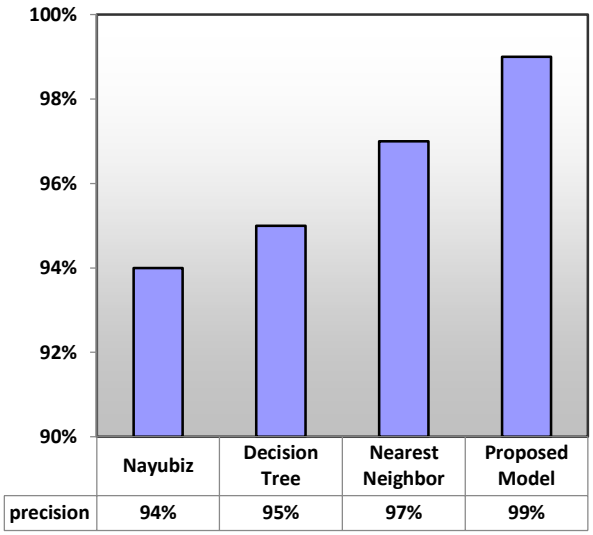

Figure 7: Graph the results with accuracy criteria

Table 11: Comparison of results with sensitivity criteria

	Nayubiz	Decision Tree	Nearest Neighbor	Proposed Model
Seboriec Dermatitis	0.852	0.902	0.869	0.967
Pesoriasis	1.000	0.973	0.991	1.000
Lichen Planus	0.944	0.972	0.986	1.000
Cronic Dermatitis	0.942	0.942	1.000	1.000
Pityriasis Rosea	0.918	0.918	0.980	0.980
Pityriasis Rubrapilaris	0.850	0.950	1.000	1.000

Table12: Comparison of results with specificity criteria

	Nayubiz	Decision Tree	Nearest Neighbor	Proposed Model
Seboriec Dermatitis	0.987	0.984	0.993	0.997
Pesoriasis	0.953	0.994	1.000	1.000
Lichen Planus	0.997	0.993	1.000	1.000

ISSN- 2394-5125 VOL 7, ISSUE 08, 2020

Cronic Dermatitis	1.000	0.997	1.000	1.000
Pityriasis Rosea	0.987	0.981	0.972	0.994
Pityriasis Rubrapilaris	0.994	0.997	1.000	1.000

Table13: Comparison of results with precision criteria

	Nayubiz	Decision Tree	Nearest Neighbor	Proposed Model
Seboriec Dermatitis	0.929	0.917	0.964	0.983
Pesoriasis	0.903	0.965	1.000	1.000
Lichen Planus	0.986	0.972	1.000	1.000
Cronic Dermatitis	1.000	0.980	1.000	1.000
Pityriasis Rosea	0.918	0.882	0.842	0.960
Pityriasis Rubrapilaris	0.895	0.950	1.000	1.000

Table 14: Comparison of results with F-Measure criteria

	Nayubiz	Decision Tree	Nearest Neighbor	Proposed Model
Seboriec Dermatitis	0.889	0.909	0.914	0.975
Pesoriasis	0.949	0.969	0.996	1.000
Lichen Planus	0.965	0.972	0.993	1.000
Cronic Dermatitis	0.970	0.961	1.000	1.000
Pityriasis Rosea	0.918	0.900	0.906	0.970
Pityriasis Rubrapilaris	0.872	0.950	1.000	1.000

Table 15: General comparison of data mining methods

	F-Measure	Specifity	Sensitivity	Precision
Nayubiz	0.927	0.986	0.918	0.938
Decision Tree	0.943	0.989	0.943	0.994
Nearest Neighbor	0.968	0.994	0.971	0.968
Proposed Model	0.991	0.998	0.991	0.991

1 Findings

The proposed model is compared with the three methods of the Naive Bayesian [22], the Decision Tree [23] and the Nearest Neighbor. The relationship between the real classes and the predicted classes can be calculated using the Confusion matrix. Figure6 lists the required parameters of the Confusion matrix.

In tables 7-10, the Confusion matrix for the Naive Bayesian, the Decision Tree and the Nearest Neighbor and the proposed model is shown, respectively. A comparison of the values in the Confusion matrix tables shows that the proposed method has better results than the other methods for the number of records that have been correctly diagnosed as positive; the correct diagnosis of negative records is also more accurate.

ISSN- 2394-5125 VOL 7, ISSUE 08, 2020

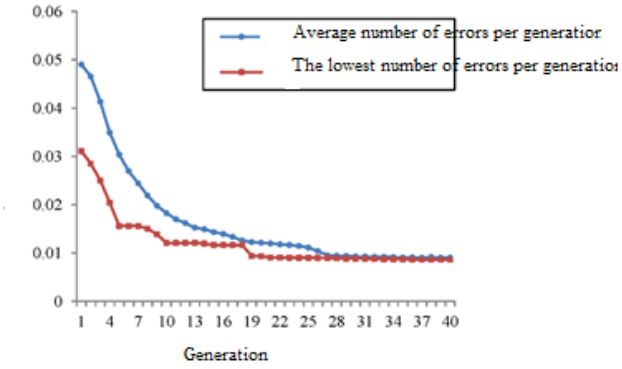

Figure 5: Error percentage for disease prognosis in the proposed algorithm

TP: The number of records that are correctly diagnosed as positive.
TN: The number of records that are correctly diagnosed as negative.
FP: The number of records that are not correctly diagnosed as positive.
FN: The number of records that are not correctly diagnosed as negative.

Figure 6: Parameters required for the relationship between the real classes and the predicted classes

Table 7: Confusuin matrix for neobase method

Class	TP	FP	TN	FN
Seboreic Dermatitis	52	4	301	9
Psoriasis	112	12	242	0
Lichen Planus	68	1	293	4
Cronic Dermatitis	49	0	314	3
Pityriasis Rosea	45	4	313	4
Pityriasis Rubra Pilaris	17	2	344	3

Table 8: Confusion matrix for decision tree method

Class	TP	FP	TN	FN
Seboreic Dermatitis	55	5	300	6
Psoriasis	109	4	250	3
Lichen Planus	70	2	292	2
Cronic Dermatitis	49	1	313	3
Pityriasis Rosea	45	6	311	4
Pityriasis Rubra Pilaris	19	1	345	1

Table 9: Confusion matrix for nearest neighbor method

Class	TP	FP	TN	FN
Seboreic Dermatitis	53	2	303	8
Psoriasis	111	0	254	1
Lichen Planus	71	0	294	1
Cronic Dermatitis	52	0	314	0
Pityriasis Rosea	48	9	308	1
Pityriasis Rubra Pilaris	20	0	346	0

Table 10: Confusion matrix for the proposed model

ISSN- 2394-5125 VOL 7, ISSUE 08, 2020

Class	TP	FP	TN	FN
Seboreic Dermatitis	59	1	304	2
Psoriasis	112	0	254	0
Lichen Planus	72	0	294	0
Cronic Dermatitis	52	0	314	0
Pityriasis Rosea	48	2	315	1
Pityriasis Rubra Pilaris	20	0	346	0

5 Conclusion

Searching medical databases will be done to gain knowledge and information to prognosis and diagnosis the diseases and decide on data mining applications in medicine. Inheritance algorithms such as genetic algorithms can be used to optimize data mining techniques. In this paper, we have used the genetic algorithm for prognosis and diagnosis of the squamous erythematosus in dermatology and the nearest neighbor k algorithm was used to optimize the data mining results and a new model was presented, this proposed model has the least error rate the highest accuracy compared to other models. The prognosis and diagnosis of the skin disease using artificial intelligence and machine learning increase the chances of successful treatment.

The simulation results suggest that the proposed model with a prognosis accuracy of 0.992 is more accurate than the Naive Bayesian, the Decision Tree and the Nearest Neighbor. The high in the diagnosis of the squamous erythematosus in dermatology indicates the superiority of the proposed approach. The weaknesses of this method are the complexity and time consuming nature of the implementation. Increasing the efficiency and reducing the temporal complexity of the proposed model using other data mining algorithms are the future propositions of this study.

References

[1] B. DA, C. NH., Introduction and historical bibliography, Wiley Blackwell; 2010

[2] G. Demiroz, H. A. Govenir, and N. Ilter, "Learning Differential Diagnosis of Eryhemato-Squamous Diseases using Voting Feature Intervals", Aritificial Intelligence in Medicine, 1998; Vol. 13,No. 3: 147-165

[3] M. Gupta, A. Gupta, Depression and suicidal ideation in dermatology patients with acne, British Journal of Dermatology, 1998; Vol. 139: 846-850,

[4] B. Mathes, M. Douglass, Seborrheic dermatitis in patients with acquired immunodeficiency syndrome, Journal of the American Academy of Dermatology, 1985; Vol. 13, No. 6: 947-951

[5] S. Breathnach, L. Planus and L. Disorders, Rook's Textbook of Dermatology, Eighth Edition, 2010; 1-28

[6] F. Drago, E. Ranieri, F. Malaguti, Human herpesvirus in patients with pityriasis rosea, Dermatology, 1997; Vol. 195, No. 4: 374-378

[7] E. D. Übeylı, I. Güler, Automatic detection of erythemato-squamous diseases using adaptive neuro-fuzzy inference systems, Computers in Biology and Medicine, 2005; Vol. 35, No. 5: 421-433,

[8] W. Griffiths, P. Pilaris, Clinical and experimental dermatology, 1980; Vol. 5, No. 1: 105-112

[9] K. J. Cios, G. W. Moore, Uniqueness of medical data mining, Artificial intelligence in medicine, 2002; Vol. 26, No. 1: 1-24,

[10] J. Xiea, C. Wangc, Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases, Expert Systems with Applications, Vol. 38, No. 5, 2011: 5809–5815

[11] E. Übeylı, İ. Güler, Automatic detection of erythemato-squamous diseases using adaptive neuro-fuzzy inference systems, Computers in Biology and Medicine, 2005; Vol. 35, No. 5: 421–433

[12] T. S. Soltani, et al, Automatic Detection of Erythemato-Squamous Diseases Using K-Nearest Neighbor Algorithm, Iranian Journal of Medical Informatics, 2012; Vol 2, No. 2: 15-18

[13] J. H. Holland, Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence, U Michigan Press; 1975

[14] Frank, Asuncion, UCI Machine Learning Repository, University of California;2010,http://archive.ics.uci.edu/ml/machine-learning-databases/audiology/audiology.data, 30-May-1989.

[15] Dodge, Y The Oxford Dictionary of Statistical Terms, OUP.; 2003

[16] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd edition, Morgan Kaufmann; 2011

[17] Goldberg, D., Holland, J. "Genetic algorithms and machine learning", Machine learning, Vol. 3, No. 2: 95-99

[18] E. G. Talbi, Metaheuristics: From Design to Implementation, Wiley, 2009

[19] Alpaydin, Voting over Multiple Condensed Nearest Neighbors, Artif. Intell. Rev., 1997; 11: 115-132,

ISSN- 2394-5125 VOL 7, ISSUE 08, 2020

[20] Kohavi, Ron "A study of cross-validation and bootstrap for accuracy estimation and model selection". Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (San Mateo, CA: Morgan Kaufmann), 1995; Vol. 2, No. 12: 1137–1143
[21] D. Jong, K., An Analysis of the Behavior of a Class of Genetic Adaptive Systems, Ph.D. dissertation, University of Michigan, Ann Arbor, MI, 1975
[22] Rish, I.: An empirical study of the naive Bayes classifier, IJCAI-01 Workshop on Empirical Methods in Artificial Intelligence; 2001
[23] Rokach, L., Maimon, O., Data Mining with Decision Trees: Theory and Applications,World Scientific Publishing; 2008