Machine Learning Research
2020; 5(3): 39-45
http://www.sciencepublishinggroup.com/j/mlr
doi: 10.11648/j.mlr.20200503.11
ISSN: 2637-5672 (Print); ISSN: 2637-5680 (Online)

Applying Different Pattern Recognition Methods for Identifying Skin Diseases

Amir Zirjam[1, *], Saman Rajebi[2]

[1]Department of Biomedical Engineering, Tabriz Branch, Islamic Azad University, Tabriz, Iran

[2]Faculty of Electrical Engineering, Siraj Institute of Higher Education, Tabriz, Iran

Email address:
stu.amirzirjam@iaut.ac.ir (A. Zirjam), s.rajebi@seraj.ac.ir (S. Rajebi)

[*]Corresponding author

To cite this article:
Amir Zirjam, Saman Rajebi. Applying Different Pattern Recognition Methods for Identifying Skin Diseases. *Machine Learning Research.* Vol. 5, No. 3, 2020, pp. 39-45. doi: 10.11648/j.mlr.20200503.11

Received: October 25, 2020; **Accepted**: November 10, 2020; **Published**: November 19, 2020

Abstract: According to the WHO (World Health Organization) (2015), cancer is the first or second major cause of death before the age of 70 in 91 of 172 countries, and it is ranked third or fourth in 22 other countries. In 2018, out of 1042056 new non-melanoma skin cancer cases in the world, 6.25% of them had been reported to have died. The most effective method to reduce disease mortality is early diagnosis, which requires a precision and reliable diagnosis. Automatic diagnosis is speedy and far from human error and reduces the workload and warns about patients who need more attention, and allows physicians to focus on diagnosis and prognosis. For automatic classification, six K-NN methods, weighted K-NN, Bayesian, perceptron artificial neural network, RBF neural network, SVM are used, and the results of the correct classification rate are compared. Then the correct classification rate is significantly increased using the FDR formula and genetic algorithm. RBF, perceptron artificial neural network, and weighted K-NN methods had the best precision of classification, respectively. After applying the genetic coefficients, RBF weighted K-NN and K-NN methods are reached to a precision of 100%. After them, SVM and perceptron artificial neural network methods are reached to a precision of 99%.

Keywords: Neural Network, RBF, Perceptron, K-NN, Bayesian, Melanoma, Eczema, Psoriasis, Genetic Algorithm, FDR

1. Introduction

The skin is the largest organ in the body. The skin protects against heat, sunlight, damage, and infection. It also helps control body temperature and stores water, fat, and vitamin D. The skin has several layers. However, the two main layers are the epidermis (top or outer layer) and the dermis (bottom or inner layer). Today, skin diseases are widespread. Some of them are simple and easy to treat, but some are very harmful and may not cure them. Therefore, this important organ in the body should be taken care of. Diagnosis of skin diseases is very complex, especially when the symptoms of more than one disease are almost the same.

Therefore, a dermatologist with a wide range of skin diseases experience is needed. According to the WHO (2015) estimates, cancer is the first or second leading cause of death before the age of 70 in 91 of 172 countries, and in 22 other countries, it is ranked third or fourth. The outbreak and mortality of cancer in the world by continents are as follows [1, 2]:

Table 1. *Outbreak and mortality in the world.*

Continent	Outbreak rate	Mortality rate
America	21.0%	14.4%
Africa	5.8%	7.3%
Europe	23.4%	20.3%
Asia	48.4%	57.3%
Oceania	1.4%	0.7%

The main types of skin cancer are SCC, BCC, and melanoma. Squamous Cells (S.C.s): The thin, flat cells make up the top layer of the epidermis. Cancer that makes in S.C.s is called SCC of the skin.

Basal cells (B.C.s): These cells are round cells beneath S.C.s. BSC is carcinoma in stem cells.

Melanocytes: It is found in the lower part of the epidermis; these cells make melanin, a pigment that gives the skin its natural color. When the skin is exposed to the sun, melanocytes produce more pigment, causing the skin to stain

or darken. Cancer that makes in melanocytes is called melanoma. Melanoma is much less common than other types but is much more likely to attack the surrounding tissue and spread to other parts of the body. Most deaths from skin cancer are due to melanoma "figure 1a".

BCC and SCC of the skin are also called non-melanoma skin cancer and are the most common skin cancer forms. Most BSCs and SCCs are curable. Melanoma is more likely to spread to surrounding tissues and other parts of the body and is more difficult to treat. Melanoma becomes easier if the tumor is found before it spreads to the dermis (the skin's inner layer). Early diagnosis of melanoma and early treatment reduces the likelihood of mortality [1, 3].

The most effective method to reduce mortality from disease and cancer is to diagnose it early. Early diagnosis requires an accurate and reliable diagnosis. Besides, the non-automatic diagnosis will be very time consuming, tedious, and associated with human error, so the use of machine learning tools in medical diagnosis is gradually increasing. An automated diagnostic computer system allows physicians to focus on diagnosis and prognosis by reducing workload and alerting patients who need more attention.

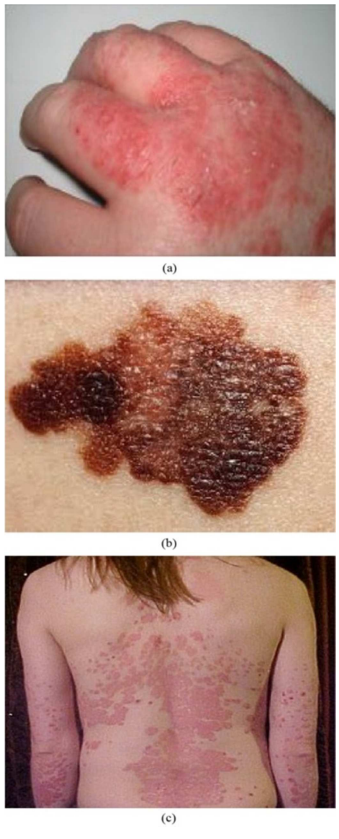

Figure 1. *(a) Melanoma skin disease, (b) Eczema skin disease, (c) Psoriasis skin disease.*

In this paper, 4 data classes are classified: melanoma, eczema "figure 1b", psoriasis "figure 1c", and healthy. By using image processing on skin images forty-eight features were collected from each class include 64 sample. KNN, Bayesian, neural networks, and SVM have been used to analyze and classify these diseases.

2. Methodology

2.1. K-Nearest Neighbor

K-NN algorithm is widely used in pattern recognition and data mining for classification due to its simple implementation and outstanding performance. The main idea of the standard K-NN method is to predict the label of a test data according to the law of majority so that the Euclidean distance of the test data from the training data is calculated and the test data class is determined based on the class with the largest number of neighbors [4].

Steps of identifying and determining the class in K-NN method:

Creating training and test data
Calculating the Euclidean distance
Specifying the class of neighbors with a short distance
Determining the maximum class as the test data class
End of classification
Calculating the correct classification rate
The end

In this method, the correct classification rate of 87.5% in 160ms has been calculated, and also in the number of different neighborhoods is calculated as follows.

Table 2. *The correct classification rate in the number of different neighborhoods.*

Number of neighbors	Correct classification rate
3	88.09%
5	87.5%
7	87.5%
9	88.09%
11	89.28%

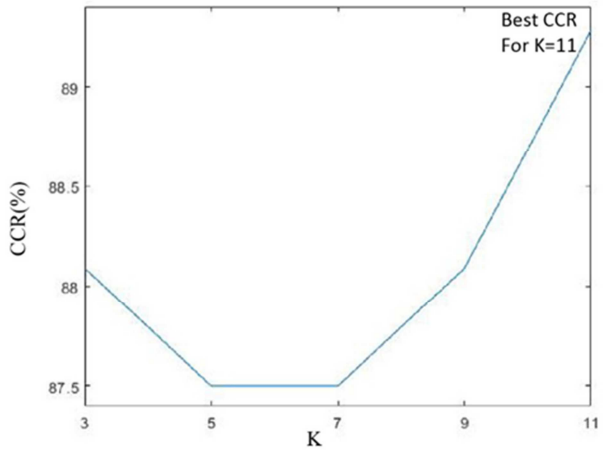

Figure 2. *The correct classification rate in the number of different neighborhoods.*

2.2. Weighted K-Nearest Neighbor

First, a weighting method is introduced for K-NN, which is called the weight distance of k of K-NN. Using the weight distance function, more weight is allocated to closer neighbors than farther neighbors.

Weight w_i for the i^{th} nearest neighbor:

$$w'_i = \begin{cases} \frac{d(x', x_k^{NN}) - d(x', x_i^{NN})}{d(x', x_k^{NN}) - d(x', x_i^{NN})}, if\ d(x', x_k^{NN}) \neq d(x', x_1^{NN}) \\ 1, if\ d(x', x_k^{NN}) = d(x', x_1^{NN}) \end{cases} \quad (1)$$

Which is x' data for weight change and x_i^{NN} is the weight training data of class i. In other words, the class can be defined by the following relation. Then weighted majority of the neighbors then determines the classification result. In the conditions of Equation 1, one of the values will be one or zero.

$$classification\ resatio = argmax \sum_{(x_i^{NN}, y_i^{NN}) \in T'} w'_i \times \delta(x = x_i^{NN}) \quad (2)$$

According to (1), it can be seen that a neighbor at a close distance has more weight than farther distance. The nearest neighbor weighs 1, and the farthest neighbor weighs 0, and the other neighbors are divided linearly between them.

This method, was used in data classification and the correct classification rate 92.26% in 250ms has been calculated [5].

2.3. Bayesian Method

In the Bayesian method, according to the Gaussian distribution for all data in each class, based on the density function, the probability of each class is calculated

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (3)$$

By calculating the mean values and the variance of features of each class. The probability of attendance of test data in each class is calculated. The class of test is selected as the class that has shown the highest probability [6].

Steps of identifying and determining the class in Bayesian method [6, 7]:

Putting test samples in the Gaussian Equation

Calculating the probability function of each test sample

Determining the class with the highest probability of attendance for the test data

The end and calculating the correct classification rate

In this method, the correct classification rate 91 in 110ms has been calculated.

2.4. Perceptron Artificial Neural Network

Artificial neural networks that are (a very popular classification method) (neural network inspires the learning technique in the human brain.) After processing with the inputs of the neurons of the previous layer, this network sends the weight values of each of the artificial neurons as output to the next layer. This change in weight between the neurons minimizes the error. An important issue in this network is determining the number of layers and the number of neurons in the hidden layer and its relationship. These issues and parameters have a significant effect on neural network performance. The results may be very different in each of these parameters. Different architectures will have different results for different problems. However, achieving a desirable architecture with trial and error is very important. The training data set were used to determine the bias and weight values. The training is iterated to obtain the lowest level of error by changing the number of iterations and neurons. The trained algorithm is then used on the test data set [8, 9].

The perceptron neural network algorithm for the data set in this paper is designed according to the figure 3.

A graph called the regression line is drawn using the confusion matrix, which shows the correct classification rate. This graph includes bisector of the first and third quarters, and according to the number of data classes used, it will have the number of classes that in this article will have numbers from 1 to 4 in each axis so that each detected class is compared to the real class and is marked at the same point. If the correct classification rate is 100%, all marks will be on the bisector of the first and third quarters. If the points are outside this line, it indicates that the correct classification rate is not 100% [10].

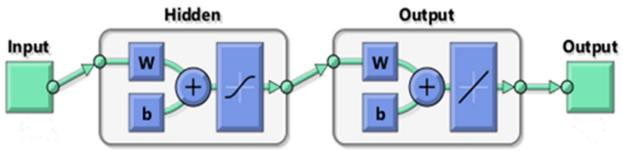

Figure 3. *Designed perceptron neural network with* 10 *neurons in the hidden layer and use of conversion functions tansig and linear respectively in the hidden and output layers.*

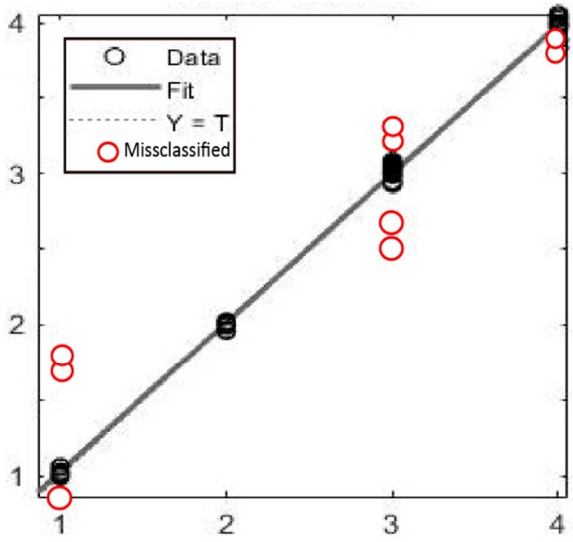

Figure 4. *Perceptron neural network regression.*

The correct classification rate of 99.72% in 360 ms has been calculated by the perceptron neural network.

2.5. Radial Basis Function

RBF neural network is one of the types of artificial neural networks with its own advantages, including approximate capabilities, simpler network structures, and faster learning algorithms. The RBF network is a fully connected three-layer network forward, which uses RBFs as the only nonlinearity in hidden layer neurons. The output layer is nonlinear, and the output layer connections have weights, but the connections from the input to the hidden layer have no weight. Due to better capabilities, simpler network structures, and faster

network learning algorithms have wide applications in engineering sciences and disciplines. RBF neural networks are based on interpolation. Theories consist of three leading layers. The first layer receives a vector and propagates to the middle layer, which is made up of neurons that use RBF. The output is received in the last layer [11, 12].

The RBF neural network error can be seen in the "figure 5" which after performing 130 Radial Basis functions for its layer the error value for this number of function is almost zero and does not need to be iterated. The correct classification rate with this neural network of 100% in 14.36 s has been calculated.

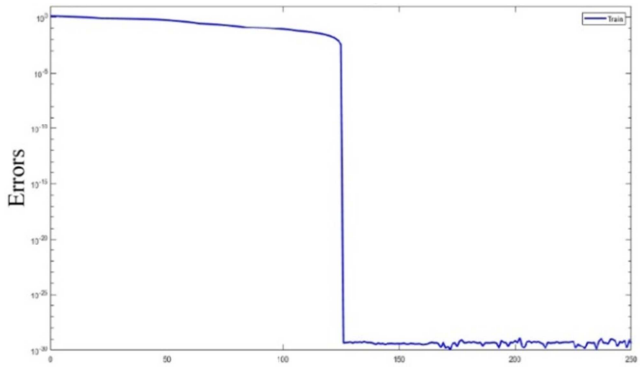

Figure 5. *RBF neural network error.*

2.6. Support Vector Machine Neural Network

SVM is a powerful data classification tool that tries to find a linear separator between data so that it has the greatest distance from all classes. This method is used for data of two classes, and for data in more than two classes should be calculated in pairs [11, 13].

The goal is to find the best line to separate the two classes, which its Equation is (4)

$$w_1x_1 + w_2x_2 + b = 0 \quad (4)$$

$$w1X1 + w2X2 + b=0 \quad (3)$$

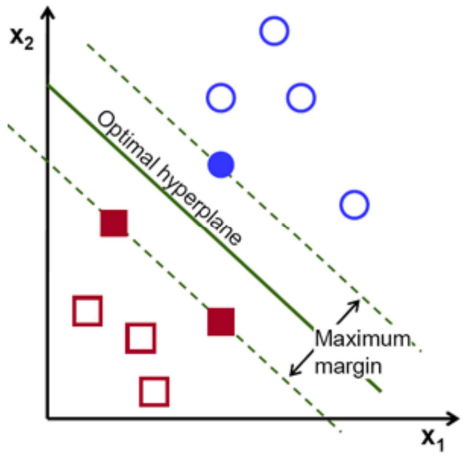

Figure 6. *Discrimination in two classes.*

In the above equation, w and b represent the gradient and the distance of the line from the origin point, respectively. The central divider line can be described by the following equation (5):

$$w^T x+b=0 \quad (5)$$

Instead of using this line in SVM, two parallel lines to create a more secure boundary with equation are described (6) and (7):

$$w^T x+b=1 \quad (6)$$

$$w^T x+b=-1 \quad (7)$$

Considering the data, we will have +1 and -1 in two classes: $\{x_i, y_i\}$; i=1, 2,…, n
$x_i \in R$, $y_i \in \{-1, 1\}$

$$\text{If } y_i=1 \rightarrow w^T x_i+b>1 \quad (8)$$

$$\text{If } y_i=-1 \rightarrow w^T x_i+b<-1 \quad (9)$$

According to "figure 6" the distance of two parallel lines from the midline is called d_1 and d_2. The best way to separate classes is to have two equal distances $d_1=d_2$.

The best case for separating classes will occur when $d_1=d_2$.

By doing calculations we have (10):

$$d=d_1 + d_2 = \frac{2}{w} \quad (10)$$

By decreasing the value of w, the separator distance between the lines will increase, so the objective function for minimization can be expressed as follows (11):

$$L=\frac{1}{2}w^T w - \sum_i^n \alpha_i y_i (w^T x_i + b) -1 \quad (11)$$

Which α is known as the Lagrange coefficient.

After solving the above equations, the values of b and w are presented with equations (12) and (13), respectively:

$$w=\sum_i \alpha_i y_i x_i \quad (12)$$

$$b_i = y_i - w^T x_i \; b = mean(b_i) \quad (13)$$

As mentioned the SVM method is actually a binary classification, while most of the topics like the dataset in this article are related to the multi-class classifiers. In such cases, a multi-class problem can be reduced to several binary cases, and a multi-class problem can be solved by comparing pairs of classes and combining their outputs. [13, 14]

By using this method, the correct classification rate of 73.21% in 190 ms has been calculated.

3. Improve the CCR with Intelligent Algorithms

3.1. Fisher's Discriminant Ratio

Assuming a Gaussian distribution for natural class sample data, FDR can be used to determine the degree of class discrimination from each other and the effect of each feature on the degree of this discrimination and the possibility of

combining features for better discrimination.

In order to better discriminate the classes, the common area between the classes should be as small as possible. The greater the difference between the mean classes and the smaller the variance of the classes, the smaller the common area and the better the discrimination of the classes. Unknown coefficients are considered for data features and the values of these coefficients are determined using optimization algorithms in a way that maximizes FDR. For a four classes problem, FDR can be defined as follows [15]:

$$FDR = \frac{\begin{matrix}(a'*(mu1-mu)*(mu1-mu)'*a)+\cdots \\ (a'*(mu2-mu)*(mu2-mu)'*a)+\cdots \\ (a'*(mu3-mu)*(mu3-mu)'*a)+\cdots \\ (a'*(mu4-mu)*(mu4-mu)'*a)\end{matrix}}{a'*\sigma_1*a+a'*\sigma_2*a+a'*\sigma_3*a+a'*\sigma_4*a} \quad (14)$$

That a is a matrix of unknown coefficients of class features.

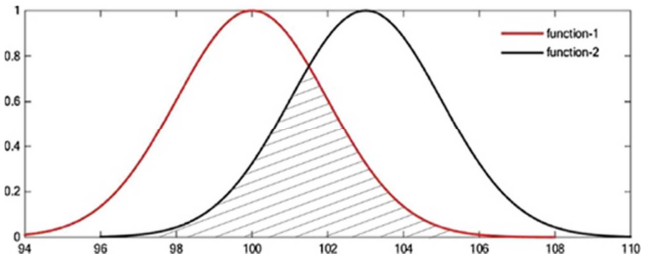

Figure 7. *An example of how Gaussian distributes a two-class data.*

3.2. Genetic Algorithm

A genetic algorithm is an optimization algorithm whose strategy is inspired by observation principles in natural evolution. The search technique is to find an approximate solution for optimizing models, math, and search problems. A genetic algorithm is a special type of evolutionary algorithm that uses evolutionary biology techniques such as inheritance to find the optimal formula for prediction. The problem to be solved has inputs that are transformed into solutions in a process modeled on genetic evolution. Solutions are then evaluated as candidates by the fitness function, and the algorithm terminates if the exit condition is met. In general, it is an iteration-based algorithm that most of its parts are selected as random processes that these algorithms consist of parts of the fitness function, display, selection, and change. This algorithm results in the least optimal solution by solving the problem. First, FDR is considered as a target function, and since it will find the minimum genetic algorithm, the FDR formula should be considered in reverse.

The steps of the genetic algorithm are as follows: [16]

Start

Selecting the initial population

Iteration

Evaluation of the fitness function concerning the population

Selecting better answers to reproduce the population

Iteration of this process until the optimal answer has resulted

The end

By defining the fitness function of the genetic algorithm as follows, minimizing the function leads to maximizing the FDR.

$$\text{Fitness function}=\frac{1}{FDR} \quad (15)$$

After applying the genetic algorithm and maximizing FDR, the coefficients of the feature are calculated as follows:

Table 3. *Coefficients of the features wich calculated by genetic algorithm.*

1-12	13-24	25-36	37-48
9.999140717367787	0.3481031283088445	-9.994296822113398	-0.11852034431932879
3.9559594038095742	0.26751539225995913	1.4698500942966852	0.35934206629611865
9.971390027657685	-0.15304297128765043	2.8393730487973823	-0.2365921265249078
3.079670991801512	0.40939614723323103	-2.86378336795255	0.0732026746900285
7.034837215629633	-1.305456036493018	-0.814270232520828	-0.1947026318936924
9.97327778599299	1.2249367430555402	-9.99830754570172	1.010573157083975
1.9998586537269993	-0.04368028063495277	-2.436979333758707	0.02079204119282707
0.8218887033953308	0.017851019748521324	1.8910441294060956	0.02671983936849287
0.4461522233132129	0.11514922054395171	0.33052427881813884	-0.0212676808985659
1.1569201620682001	0.04871029155507678	-0.8146703320637538	0.118711959317368
-2.1039600210358813	-0.052418699260442736	-0.6230488656061457	-0.15094500153791923
-0.10844184537505797	0.12693302358008296	3.4377796147943336	-0.1845123286245638

After calculating the coefficients, again to test classification methods with new datas. The CCR results are presented in Table 3.

Table 4. *The correct classification rate after applying the coefficients of the genetic algorithm.*

Algorithm	Correct classification rate after applying the coefficients
K_NN	100%
W K-NN	100%
Bayesian	97.61%
RBF	100%
SVM	99.40%
Per	99.78%

These results show that the genetic algorithm is effective on all methods and the three methods of K-NN, weighted K-NN, and RBF neural network reached 100% accuracy.

4. Conclusion

Since automatic diagnosis is much faster and more accurate than non-automatic methods, and on the other hand, early diagnosis of these diseases will reduce the mortality rate. In the automatic diagnosis method, the workload is reduced, and the accuracy is significantly increased.

In this article, skin diseases classify using K-NN, weighted K-NN, Bayesian, neural networks, and SVM methods. The best results were RBF neural network, perceptron neural network, weighted K-NN, Bayesian method, K-NN, and S. V., respectively. Then, using a genetic algorithm and applying coefficients for the features, the correct classification rate was increased, which RBF neural network method and K-NN, weighted K-NN, perceptron neural network, SVM, and Bayesian had the correct classification rate, respectively. The genetic algorithm has the highest increase in the correct classification rate in the SVM and K-NN methods. In the classification of classes, the RBF neural network's correct classification rate reached 100%. After applying the genetic algorithm's coefficients, RBF neural network, weighted K-NN, and K-NN reached the correct classification rate of 100%.

In article 15 of the Hindawi journal using the SVM method and image features of skin diseases of herpes, dermatitis, psoriasis is classified into three classes. The classification accuracy of each class is separately calculated 85%, 90%, 95%, respectively. On average, the correct accuracy of classification can be considered 90% [17]. In this article, using the SVM method, the correct classification rate was first calculated 73%, and after applying genetic coefficients, it was calculated 99%.

Table 5. *The correct classification before and after applying the coefficients of the genetic algorithm.*

Algorithm	Correct classification rate before applying the genetic algorithm coefficients	Correct classification rate after applying the genetic algorithm coefficients
K_NN	87.5%	100%
W K-NN	92.26%	100%
Bayesian	91.07	97.61%
RBF	100%	100%
SVM	73.21%	99.40%
Per	99.72%	99.78%

References

[1] F. Bray, J. Ferlay, I. Soerjomatarma, R. L. Siegel, L. A. Torre, A. Jemal, “Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries,” 2018. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp. 68–73.

[2] S. S. Abu Naser, A. N. Akkila, “A Proposed Expert System for Skin Diseases Diagnosis”, 2008. K. Elissa, “Title of paper if known,” unpublished.

[3] J. Ferlay, M. Colombet, I. Soerjomataram, C. Mathers, D. M. Parkin, M. Piñeros, A. Znaor, F. Bray, “Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods: International Journal of Cancer”, 2018. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987.

[4] S. Zhang, X. Li, M. Zong, “Learning k for kNN Classification”, ACM Transactions on Intelligent Systems and Technology, Vol. 8, No. 3, 2017.

[5] J. Gou, L. Du, Y. Zhang, T. Xiong, “A New Distance-weighted *k*-nearest Neighbor Classi_er”, Journal of Information & Computational Science, 2012.

[6] D. M. Farid, M. Z. Rahman, “Anomaly Network Intrusion Diagnosis Based on Improved Self Adaptive Bayesian Algorithm”, JOURNAL OF COMPUTERS, VOL. 5, NO. 1, doi: 10.4304/jcp.5.1.23-31, 2010. Fahey, J. W., Zakmann, A. T., and Talalay, P. (2001). The chemical diversity and distribution of glucosinolates and Isothiocyanates among plants. Corrigendum Phytochemistry, 59: 200-237.

[7] S. Sarabi, M. Asadnejad, S. A. Tabatabaei Hosseini, S. Rajebi, “USING ARTIFICIAL INTELLIGENCE FOR DIAGNOSIS OF LYMPHATIC DISEASE AND INVESTIGATION ON VARIOUS METHODS OF ITS CLASSIFICATIONS: IJTPE”, Vol. 12, No. 2, 2020.

[8] M. M. Saritas, A. Yasar, “Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification”, International Journal of Intelligent Systems and Applications in Engineering, 2019.

[9] Z. Jafari, A. Mahdavi Yousefi, S. Rajebi, “INVESTIGATION ON DIFFERENT PATTERN CLASSIFICATION METHODS AND PROPOSING THE OPTIMUM METHOD WITH IMPLEMENTATION ON BLOOD TRANSFUSION DATASET”, Vol. 12, No. 2, 2020.

[10] Z. Jafari, S. Rajebi, S. Haghipour, “Using the Neural Network to Diagnose the Severity of Heart Disease in Patients Using General Specifications and ECG Signals Received from the Patients”, Vol. 5, No. 5, 882-892 (2020).

[11] S. Gupta, D. Kumar, A. Sharma, “DATA MINING CLASSIFICATION TECHNIQUES APPLIED FOR BREAST CANCER DIAGNOSIS AND PROGNOSIS”, Indian Journal of Computer Science and Engineering, Vol. 2, No. 2, 2011Kass, R. E. and A. E. Raftery (1995). Bayes Factors. Journal of the American Statistical Association 90, 773–794.

[12] S. Noman, S. M. Shamsuddin, A. E. Hassanien, “Hybrid Learning Enhancement of RBF Network with Particle Swarm Optimization”, Vol. 1, SCI 201, pp. 381–397, 2009.

[13] V. Yousefi, S. Kheiri, S. Rajebi, "EVALUATION OF K-NEAREST NEIGHBOR, BAYESIAN, PERCEPTRON, RBF AND SVM NEURAL NETWORKS IN DIAGNOSIS OF DERMATOLOGY DISEASE", Technical and Physical Problems of Engineering, Vol. 12, 2020.

[14] S. Mirzayi, S. Rajebi, "Diagnosis of Epilepsy Using Signal Time Domain Specifications and SVM Neural Network: Machine Learning Research".

[15] K. Al-Khaled, "Numerical study of Fisher's reaction–di'usion equation by the Sinc collocation method", Journal of Computational and Applied Mathematics, 137 (2001) 245–255.

[16] J. Protopopova, S. Kulik, "Educational Intelligent System Using Genetic Algorithm", National Research Nuclear University MEPHI, Moscow, Russia, 2020.

[17] L. S. Wei, Q. Gan, T. Ji, "Skin Disease Recognition Method Based on Image Color and Texture Features: Hindawi", 2018.