International Journal of Latest Research in Engineering and Technology (IJLRET)
ISSN: 2454-5031
www.ijlret.com || Volume 03 - Issue 12 || December 2017 || PP. 65-71

Determining Cervical Cancer Possibility by Using Machine Learning Methods

Muhammed Fahri Unlersen[1], Kadir Sabanci[2], Muciz Özcan[1]
[1]*Department of Electrical and Electronics Engineering, Necmettin Erbakan University, Turkey*
[2]*Departmentof Electrical and Electronics Engineering,KaramanogluMehmetbey University, Turkey*

Abstract: World Health Organization (WHO) calls cancer as a generic term for a large group of diseases that can affect any part of the body. Sometimes it could be the cause of loss of patients. Cancer mortality can be reduced if cases are detected and treated early. In this fact, it is important to determine someone has highly cancer risk by using a survey. In this study, a classification of patients due to their answers to a survey has been done to determine someone who has highly cervical cancer risk. The dataset has 858 records with 33 attributes and a biopsy result. In the dataset, number of the patients with cervical cancer diagnosis is 55 and the number of healthy patient is 803. This dataset has been divided into two groups randomly as train and test. The train group is 66% of the main dataset so there are 565 records in the train dataset. The rest of the dataset has been assigned as test dataset. The test dataset has 292 records. The classification has been done by using various methods like Multilayer Perceptron, BayesNet and k-Nearest Neighbor. Correctly classified instances and percentage, true positive and false positive classified instances rate for each class and confusion matrix have been presented for each method. And all of the results are discussed.

Keywords: BayesNet, Cervical Cancer, Classification, Early Diagnosis, k-NN, Multilayer Perceptron.

I. INTRODUCTION

Cancer is a leading cause of death worldwide, accounting for 8.8 million deaths in 2015. Cancer arises from the transformation of normal cells into tumor cells in a multistage process that generally progresses from a pre-cancerous lesion to a malignant tumor. When identified early, cancer is more likely to respond to effective treatment and can result in a greater probability of surviving, less morbidity, and less expensive treatment. Significant improvements can be made in the lives of cancer patients by detecting cancer early and avoiding delays in care [1]. As in many other diseases, the existence of several screening and diagnosis methods creates a complex ecosystem from a Computer Aided Diagnosis (CAD) system point of view. Especially in developing countries, there are so restricted resources. Additionally, sometimes patients do not take care to routine screening. Therefore, the most important problems during diagnosis are determination of the finest screening plan and estimation of individual risk each patient. In most of these screening methods results have been highly correlated with the experience of the physician and its subjective decision [2].

To reduce of unnecessary screenings, a survey could be apply to patients to determine most risky group. So the patients could be take into a screening with a plan prepared according to the order of their cancer risk.

In literature it is possible to find many studies about evaluation of some medical screenings. Sen et al. (2013) in their paper have presented artificial neural network in pancreatic disease diagnosis based on a set of symptoms. An approach to detect the various stages of pancreatic cancer affected patients has been presented in this article. According to the manual detection procedure, using neural network detection has been more efficient [3]. Fernandes et al. presented a regularization-based TL approach to transfer the contribution type for each feature on linear models. In order to show its adequacy to different contexts, the proposed model-relatedness regularize was instantiated to several learning tasks related to cervical cancer screening. It was pointed out that positive results have been obtained [4]. In another study by Kalyankar et al., a survey of different type of cancer such as skin cancer, breast cancer, pancreatic cancer detection using different advanced technology such as Improved Genetic Algorithms, Artificial Neural Network, Hierarchical Clustering, Neuro-Fuzzy system, Raman Spectra has been presented. In this study, classification performance has varied between 80.5% and 95.8% depending on the cancer type and classification [5]. In 2016, Kanimozhi et al. have presented a study of different data mining techniques that can be employed in automated heart disease prediction systems. Various techniques and data mining classifiers are defined in this work for efficient and effective heart disease prediction. Each method has been provided from various studies in the literature which has used different databases. So, the success rates have varied between 45% to 99.1% depending on classification methods and attributes in the used databases [6]. The article that belongs Fatima et al. in 2017, has presented the comparative analysis of different machine learning algorithms for diagnosis of different diseases such as heart disease,

diabetes disease, liver disease, dengue disease and hepatitis disease. The classification performances of methods mentioned in this study has been given in a table. The most successful method is Naive Bayes with 97% of success rate and the worst performance is belongs to Neural Network with %70 success rate [7].

In our study a dataset created by data collected from a survey has been classified. Same classification could be done by an expert but to improve objectivity of results, the evaluation of the screening outputs need to be done by machine learning methods. So the classification has been done by machine learning methods like Multilayer Perceptron (MLP), BayesNet and k-Nearest Neighbor (kNN).

II. MATERIAL AND METHODS

The dataset has been obtained from the dataset archive belongs to the University of California, Irvine. The dataset has been collected at *Hospital Universitario de Caracas* in Caracas, Venezuela. The dataset comprises demographic information, habits and historical medical records of 858 patients. The attributes in the dataset have been presented in the Table 1 [4].

Table 1. Attribute Information

Feature	***Type***	***Feature***	***Type***
Age	int	STDs:pelvic inflammatory disease	bool
# of partners	int	STDs:genital herpes	bool
Age of 1st intercourse	int	STDs:molluscumcontagiosum	bool
# of pregnancies	int	STDs:AIDS	bool
Smokes	bool	STDs:HIV	bool
Smokes years	int	STDs:Hepatitis B	bool
Smokes packs/year	int	STDs:HPV	bool
Hormonal Contraceptives	bool	STDs: Number of diagnosis	int
Hormonal Contraceptives years	int	STDs: Time since first diagnosis	int
IUD	bool	STDs: Time since last diagnosis	int
IUD years	int	Dx:Cancer	bool
STDs	bool	Dx:CIN	bool
STDs number	int	Dx:HPV	bool
STDs:condylomatosis	bool	Dx	bool
STDs:cervicalcondylomatosis	bool	Hinselmann: target variable	bool
STDs:vaginalcondylomatosis	bool	Schiller: target variable	bool
STDs:vulvo-perinealcondylomatosis	bool	Cytology: target variable	bool
STDs:syphilis	bool	**Biopsy: class ortarget variable**	**bool**

There are 35 attributes in the dataset. The dataset has been divided into two groups as train and test randomly. The train dataset is 66% of the main dataset. So there are 566 records in the train dataset. The rest of the main dataset has been assigned as the test dataset. So the test dataset has 292 records. The train dataset has been used during training of proposed model. The obtained success rates was calculated by executing the trained model on the test dataset.

In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. An example would be assigning a given

email into "spam" or "non-spam" classes or assigning a diagnosis to a given patient as described by observed characteristics of the patient (gender, blood pressure, presence or absence of certain symptoms, etc.). Classification has been being used many applications. In some of these applications, it is employed as a data mining procedure, while in others it is employed as more detailed statistical modeling [8].

Some of classification categories have been listed as follow.

- Computer vision where medical imaging, optical character recognition and video tracking could be grouped under,
- Speech recognition,
- Handwriting recognition,
- Biometric identification,
- Document classification,
- Internet search engines,
- Pattern recognition etc .

There are various performance indicators. In this study because of having only two class, Percentage of Correctly Classified Instances (PCCI) has been used as performance indicators. In the following expressions, 'Positive' means that biopsy test highly recommended and vice versa. The results could be divide into four group. They are;

- Correctly Classified Class 0 Instances also called as True Negative Class 0 (TNC0)
- Falsely Classified Class 0 Instances also called as False Positive Class 0 (FPC0)
- Falsely Classified Class 1 Instances also called as False Negative Class 1 (FNC1)
- Correctly Classified Class 1 Instances also called as True Positive Class 1 (TPC1)

These groups could be presented as in Table 2.

Table 2. Result Groups (Confusion Matrix)

	Classified as	
	Class0	***Class1***
Class0	Class 0 instances that classified as Class 0 (TNC0)	Class 0 instances that classified as Class 1 (FPC0)
Class1	Class 1 instances that classified as Class 0 (FNC1)	Class 1 instances that classified as Class 1 (TPC1)

Number of Total True Classified Instance (TTCI) has been calculated by Equation 1.

$$TTCI = TNC0 + TPC1 \quad (1)$$

In this study, classification algorithms has been executed on WEKA. WEKA developed by Waikato University in New Zealand, is an open-source data mining software with a functional graphical interface which incorporates machine learning algorithms [9]. WEKA includes various data pre-processing, classification, regression, clustering, association rules, and visualization tools. The algorithms can be applied on the data cluster either directly or by calling via Java code [10, 11]. They are also suitable for developing new machine learning algorithms.In this study tree datamining algorithm have been proposed.

k-Nearest Neighbor Algorithm: The k-NN is a supervised learning algorithm that solves classification problems. The important point is the determination of the features of each category in advance [12]. According to the k-NN algorithm used in the classification, based on the attributes drawn from the classification stage, the distance of the new individual that is wanted to be classified to all previous individuals is considered and the nearest k class is used. As a result of this process, test data belongs to the k-nearest neighbor category that has more members in a certain class. The most important optimization problems in the k-NN method are the identification of the number of neighbors and the method of distance calculation algorithm. In the study, the identification of the optimum k number is performed with experiments, and the Euclidean Distance Calculations method is used as a distance calculation method.

Euclidean calculation method has been presented in Equation 2 [13].

$$d(p_i, p_j) = \left(\sum_{s=1}^{n} (p_{is} + p_{js})^2\right)^{1/2} \quad (2)$$

wherep_i and p_j are two different points, and the distance calculation process between them is needed. n is the number of dimensions of the space where the p_i and p_j has been.

Multilayer Perceptron: It is a feed forward type artificial neural network model which maps input sets onto appropriate output sets. A multilayer perceptron (MLP) is composed of multiple layers of nodes where each layer is connected to the next. Each node is a processing element or a neuron that has a nonlinear activation function except the input nodes. It uses a supervised learning technique named back propagation and it is used for training the network. The alteration of the standard linear perceptron, MLP is capable of distinguishing data which are not linearly separable [14].

Bayes Net: It is a probabilistic graphical model and a statistical model representing a group of random variables in addition to their conditional dependencies through a directed acyclic graph. For instance, a Bayesian network can represent the probabilistic relations between diseases and symptoms. When the symptoms are given, the network can calculate the probabilities of the existence of various diseases [15].

III. RESULTS AND DISCUSSION

In this study several methods have been investigated and three methods that have the best performances has been presented. The dataset divided into two groups as training and test. All of the methods mentioned in this study have been trained with same train dataset. Similarly, all of the methods have been tested by same test data set. But none of the records in the test dataset presents in the train dataset.

In kNN tests, number of neighbor has been changed between from 1 to 90. For each neighbor number, number of instances that have no cervical cancer diagnosis that have been classified as Negative (TNC0), number of instances that have cervical cancer diagnosis that have been classified as Negative (FNC1), number of instances that have no cervical cancer diagnosis that have been classified as Positive (FPC0), number of instances that have cervical cancer diagnosis that have been classified as Positive (TPC1) have been obtained and listed.

In Table 3, the most valuable 20 of 90 tests have been presented. Results of whole kNN tests have been presented in Figure 1.

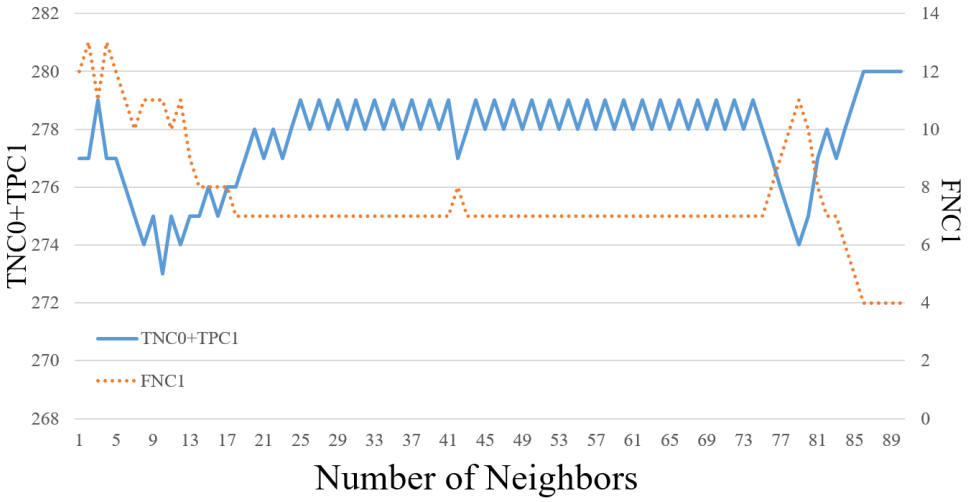

Fig 1. Results of whole kNN tests

Table 3. Results of k-NN

kNN	*TNC0*	*FNC1*	*FPC0*	*TPC1*	*TTCI*
1	*271*	*12*	*3*	*6*	*277*
2	*272*	*13*	*2*	*5*	*277*
3	*272*	*11*	*2*	*7*	*279*
4	*272*	*13*	*2*	*5*	*277*
5	*271*	*12*	*3*	*6*	*277*
6	*269*	*11*	*5*	*7*	*276*
7	*269*	*11*	*5*	*7*	*276*
8	*267*	*10*	*7*	*8*	*275*

9	*267*	*11*	*7*	*7*	*274*
10	*268*	*11*	*6*	*7*	*275*
20	*268*	*7*	*6*	*11*	*279*
27	*267*	*9*	*7*	*9*	*276*
35	*267*	*10*	*7*	*8*	*275*
42	*267*	*11*	*7*	*7*	*274*
50	*267*	*10*	*7*	*8*	*275*
56	*267*	*8*	*7*	*10*	*277*
63	*266*	*7*	*8*	*11*	*277*
71	*266*	*6*	*8*	*12*	*278*
78	*266*	*5*	*8*	*13*	*279*
86	***266***	***4***	***8***	***14***	***280***

With MLP method, it has been investigated that presenting how many neuron in the hidden layer, provides best result. For this purpose, confusion matrix has been recorded while number of neurons in the hidden layer of MLP has been changed from 1 to 50. In Table 4, the most valuable 15 of 50 tests have been presented. Results of whole MLP tests have been presented in Figure 2.

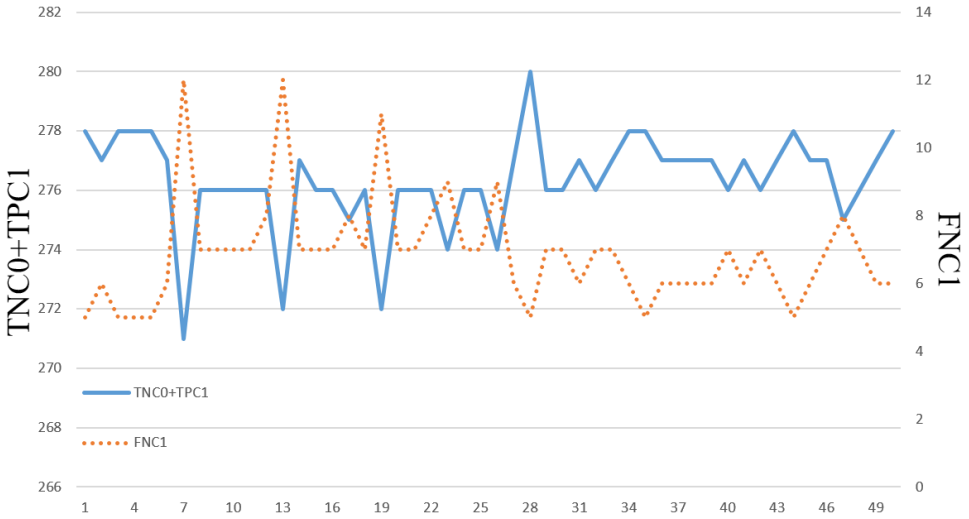

Fig. 2. Results of whole MLP tests

Table 4. Results of MLP

MLP	***TNC0***	***FNC1***	***FPC0***	***TPC1***	***TTCI***
1	*265*	*5*	*9*	*13*	*278*
3	*265*	*5*	*9*	*13*	*278*
5	*265*	*5*	*9*	*13*	*278*
7	*265*	*12*	*9*	*6*	*271*
9	*265*	*7*	*9*	*11*	*276*
10	*265*	*7*	*9*	*11*	*276*
15	*265*	*7*	*9*	*11*	*276*

20	*265*	*7*	*9*	*11*	*276*
25	*265*	*7*	*9*	*11*	*276*
28	***267***	***5***	***7***	***13***	***280***
30	*265*	*7*	*9*	*11*	*276*
35	*265*	*5*	*9*	*13*	*278*
40	*265*	*7*	*9*	*11*	*276*
45	*265*	*6*	*9*	*12*	*277*
50	*266*	*6*	*8*	*12*	*278*

Same datasets have been investigated by Bayes Net method. The results obtained by using Bayes Net has been presented in Table 5.

Table 5. Results of BayesNet

	TNC0	***FNC1***	***FPC0***	***TPC1***	***TTCI***
Bayes Net	*265*	*5*	*9*	*13*	*278*

The three methods that have best results in many data mining methods like SGD, SMO, Voted Perceptron, KStar, Multi Class Classifier Updateable, Decision Table, J48, Random Forest, Bayes Net, MLP and kNN, have been presented. The correctly classified instances are 280, 280 and 284 by kNN, MLP and Bayes Net respectively. The correctly classified instance percentage could be expressed for each method in the same order as 95.89%, 95.89% and 97.26% respectively.

IV. CONCLUSION

Although as seen in results section, the best true classification result has been obtained by Bayes Net, the problem makes the number of falsely classified instances important. While one of the objective is to make number of correctly classified instances maximum, one of the other important objective in cancer possibility estimation is to make number of instances classified as false positive minimum. Because number of instances classified as false positive means that number of patients that have cancer but not warned. In this study, this number is presented as False Negative (FNC1). For each method number of instances classified as false negative is 4, 5 and 6 with k-NN, MLP and Bayes Net respectively. That means that false classified instance rates are 1.37%, 1.71% and 2.05% respectively. Because of being very close results in number of correctly classified instance between Bayes Net and k-NN methods, the false negative results become more importance during determining of the best method. So in authors' opinion, for this problem best method is determined as k-NN method with 86 neighbor.

REFERENCES

[1] *Cancer Report*. February 2017 [cited WHO World Health Organization 25.09.2017]; Available from: http://www.who.int/mediacentre/factsheets/fs297/en/.

[2] Fernandes, K., J.S. Cardoso, and J. Fernandes, *Transfer Learning with Partial Observability Applied to Cervical Cancer Screening*, in *Pattern Recognition and Image Analysis: 8th Iberian Conference, IbPRIA 2017, Faro, Portugal, June 20-23, 2017, Proceedings*, L.A. Alexandre, J. Salvador Sánchez, and J.M.F. Rodrigues, Editors. 2017, Springer International Publishing: Cham. p. 243-250.

[3] Sen, T. and S. Das, *An Approach to Pancreatic Cancer Detection using Artificial Neural Network*, in *Proc. of the Second Intl. Conf. on Advances in Computer, Electronics and Electrical Engineering (CEEE2013)*. 2013. p. 56-60.

[4] Fernandes, K., J.S. Cardoso, and J. Fernandes, *Transfer Learning with Partial Observability Applied to Cervical Cancer Screening*, in *Iberian Conference on Pattern Recognition and Image Analysis*. 2017: Faro, Portugal. p. 61-69.

[5] Kalyankar, M.A. and N.R. Chopde, *Cancer Detection: Survey*. International Journal of Advanced Research in Computer Science and Software Engineering, 2013. **3**(11): p. 4.

[6] Kanimozhi, V.A. and T. Karthikeyan, *A Survey on Machine Learning Algorithms in Data Mining for Prediction of Heart Disease.* International Journal of Advanced Research in Computer and Communication Engineering, 2016. **5**(4): p. 6.

[7] Fatima, M. and M. Pasha, *Survey of Machine Learning Algorithms for Disease Diagnostic.* Journal of Intelligent Learning Systems and Applications, 2017. **9**(1): p. 16.

[8] Alpaydin, E., *Introduction to Machine Learning*. 2010: MIT Press.

[9] Witten, I.H., E. Frank, and M.A. Hall, *Data mining: practical machine learning tools and techniques*. 2011, London: Elsevier.

[10] Patterson, D., et al., *Performance Comparison of the Data Reduction System*, in *Proceedings of the SPIE Symposium on Defense and Security*. 2008: Orlando, FL.

[11] Hall, M., et al., *The WEKA Data Mining Software: An Update.* ACM SIGKDD Explorations Newsletter, 2009. **11**(1): p. 9.

[12] Wang, J., P. Neskovic, and L.N. Cooper, *Improving nearest neighbor rule with a simple adaptive distance measure.* Pattern Recognition Letters, 2007. **28**(2): p. 7.

[13] Zhou, Y., Y. Li, and S. Xia, *An improved KNN text classification algorithm based on clustering.* Journal of computers, 2009. **4**(3): p. 8.

[14] Gardner, M.W. and S.R. Dorling, *Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences.* Atmospheric Environment, 1998. **32**(14): p. 2627-2636.

[15] Frank, E., M.A. Hall, and I.H. Witten, *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. Fourth ed. 2016.