IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 6, Ver. III (Nov.-Dec. 2016), PP 71-77
www.iosrjournals.org

Lung Cancer Detection and Classification with Classification Algorithms

Zehra Karhan[1], Taner Tunç[2]
Ondokuz Mayıs University Samsun/Turkey

Abstract *: In this study, lung cancer was detected by looking at blood value obtained from Ondokuz Mayıs University Department of Chest Diseases. During the detection phase; classification is performed and the results were evaluated with the performance comparison between machine learning algorithms. The patient's information was used in the present data; Age, hb, wbc, neu, lymph, plt, mpv, plr, nlr. The necessary normalizations were made on our data before classification. In classification stage, k-nearest neighbors, support vector machines, naive bayes, artificial neural networks, and logistic regression methods are used, which are frequently used classification algorithms. It has been given comparatively in the detection of lung cancer accuracy rates, F-1 measure, precision, sensitivity, specificity among classifiers. Classification accuracy; support vector machines, neural networks, k-nearest neighbor, logistic regression, naive bayes algorithms gave the best result depending on the data, respectively.*

Keywords: *support vector machines, neural networks, k-nearest neighbor, logistic regression, naive bayes.*

I. Introduction

Lung cancer is a cancer type with a high incidence. It is one of the cancer types that causes death the most. Lung cancer constitutes 12.8% of all cancer types worldwide; also it constitutes 17.8% of the cancer deaths. The incidence (the number of a specific disease is within a certain time period) of lung cancer increases by 0.5% every year globally worldwide. The proportion of this cancer in males represents 38.6%, and in females it represents 5.2% in all cancers. However, it is suggested that 15% of lung cancer patients live 5 years or more after the diagnosis, together with this early diagnosis and used drugs when they are diagnosed early [1,2].

In lung cancer, cells multiply uncontrollably, like every cancer type. It also leads to the change of various hematological values with this reproduction [3]. It makes it possible at computer-aided diagnosis of lung cancer by evaluating these parameters.In this study, it became possible to diagnose cancer by looking at some information in blood values. This information; Hemoglobin (hb), mean red blood cell value (wbc), mean white blood cell value (neu), some kind of white blood cell (lenfo), platelet (plt), mean platelet value (mpv), positive probability ratio (plr), negative probability ratio (nlr) and age of the patient.Machine learning algorithms have been exploited by using blood values to diagnose lung cancer. The performances among the algorithms are given comparatively; k-nearest neighbors algorithm, naive bayes, support vector machines, artificial neural networks and logistic regression.

II. Material & Methodology

The lung data obtained from patients of hemoptysis to Ondokuz Mayıs University Department of Chest Diseases between November 2003 and September 2006. Posteroanterior chest radiography, complete blood count was performed for each patient. It includes 178 observations, where each sample has 10 properties which are detailed in Table I.

Table I. Lung Dataset Explanations

Lung Cancer Data Description		
Attribute No	***Attribute Name***	***Type***
1	HB	Numeric
2	Age (years)	Numeric
3	WBC	Numeric
4	NEU	Numeric
5	LENFO	Numeric
6	PLT	Numeric
7	MPV	Numeric
8	PLR	Numeric
9	NLR	Numeric
10	Severity (0 or 1)	Binary

Lung Dataset: the the severity value '1' means a cancer for lung cancer (patient) and '0' is a healthy for lung cancer. There are 77 (48.7%) cases in class '1' and 81(51.3%) cases in class '0'. The HB value is 1 to 4; mean hemoglobin. The WBC value is 1 to 5; means red blood cells. The NEU value is 1 to 4; means white blood cells. The LENFO value is 1 to 5; means a type of white blood cells . The PLT value is 1 to 5; means platelets. The MPV value is 1 to 5; means mean platelet volume. The PLR value is 1 to 5; means the positive likelihood ratio. The NLR value is 1 to 5; means the negative likelihood ratio. This study reports a comparative study of classification algorithms, KNN, SVM, NN, Naive Bayes and Logistic Regression. Evaluation; it has been performed on the lung dataset to see the difference among the five classification methods. General scheme of the study is shown in Figure 1.

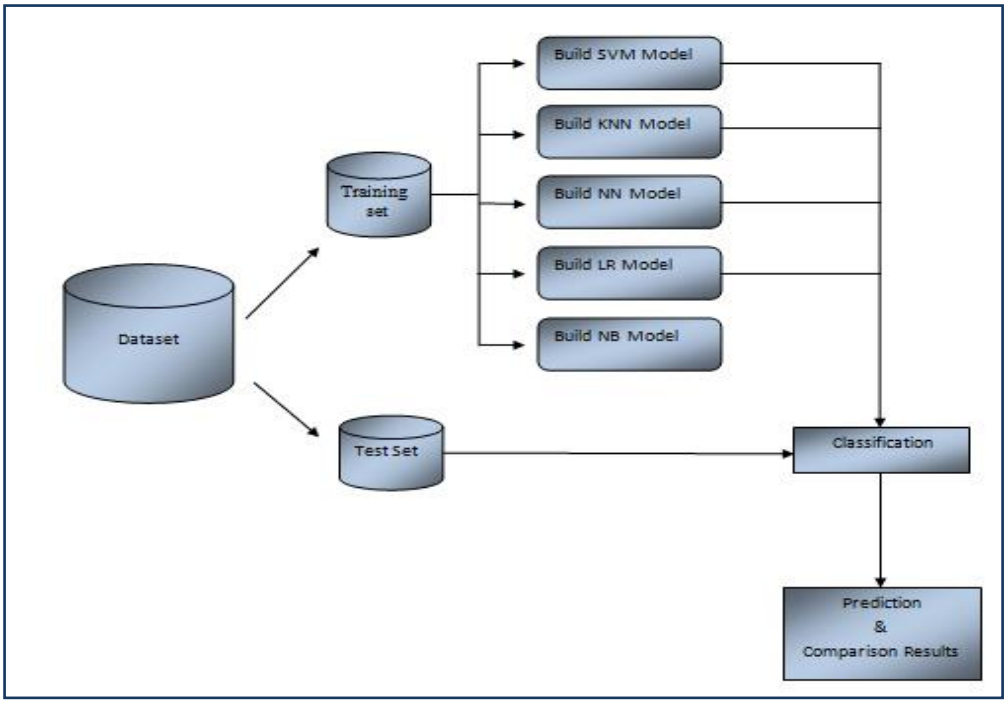

Figure 1. An Overview Of Study

III. Classification Methods

3.1. Support Vector Machine

Support Vector Machine (SVM) method was introduced by Vapnik et al [4,5]. Create a model by using training data and according to this model our test data is included in a certain class. To separate these classes, hyperplane is determined in the model building process. SVM is one of the strongest algorithm for machine learning and especially in multi-class problems [6,7]. SVM constructs hyperplanes linearly and have to find the optimal hyperplane by maximizing the wideness between support vector points and minimize the risk of misclassification examples of the test dataset [8,9]. Fig. 2 shows the structure of SVM. The dataset is separated into two different classes which are shown as orange and green images. In addition, Fig.2 presents hyperplanes, margin and support vectors used to distinguish the classes.

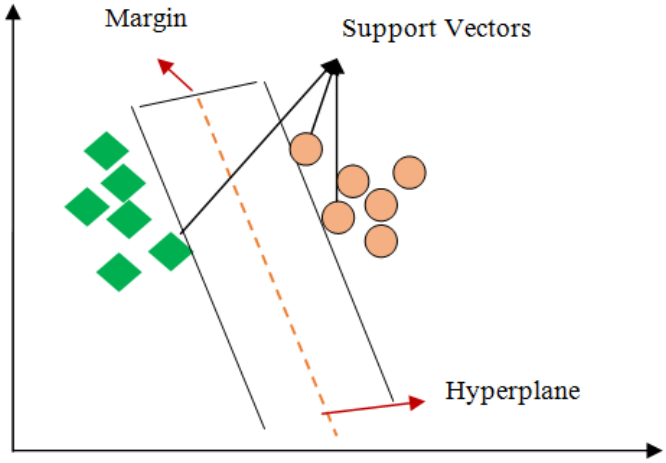

Figure 2. The general structure of SVM

3.2. K-Nearest Neighbor

The K-Nearest Neighbor (KNN) algorithm is one of the simplest classification approaches in machine learning methods. It has been used in many applications in the field of data mining [10,11]. By this algorithm, the classifying object assigns to the label of its k-nearest neighbor or neighbors according to the feature vector from reference space (training data). It is classified according to the distance and the determined number of k, so this algorithm is named as k-nearest neighbor algorithm [12,13].

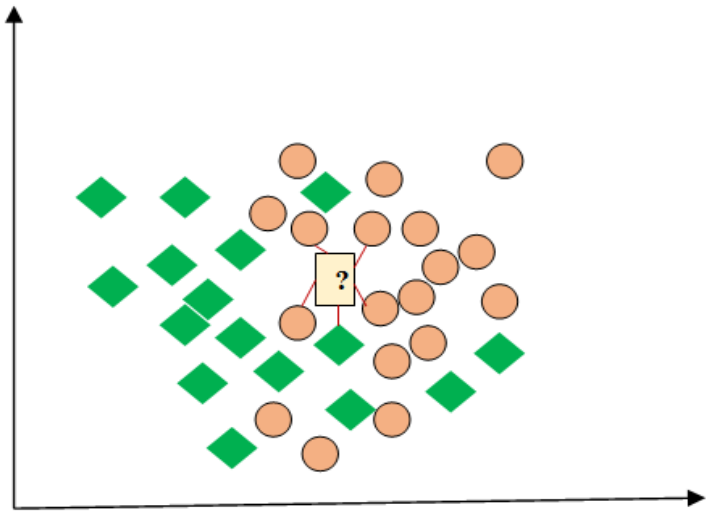

Figure 3. KNN Classification (for k=5 neighborhood)

If placed on the two-dimensional coordinate system in Fig.3 is referred to in the five example of the five nearest neighbor. If our k value is 5, then the object is simply classified by taking and testing 5 nearest neighbor. 4 of the 5 samples are circle, then square will be assigned to class circle.

3.3. Neural Network

Neural Network were developed by inspired from the information processing and elementary principle of the operation technique of human brain (neural system) [14]. With NN, simple biological neural system is imitated. It's processing, the signals are taken from other neurons and it continues by combining them. The neurons combine the all stage by each other and with this neural network will be constituted. The structure of the artificial neural network consists of at least 3 layers: the input layer, an output layer and at least one hidden layer [15,16]. Figure 4 shows the structure of an artificial neural network consisting of 4 input and 2 output layers.

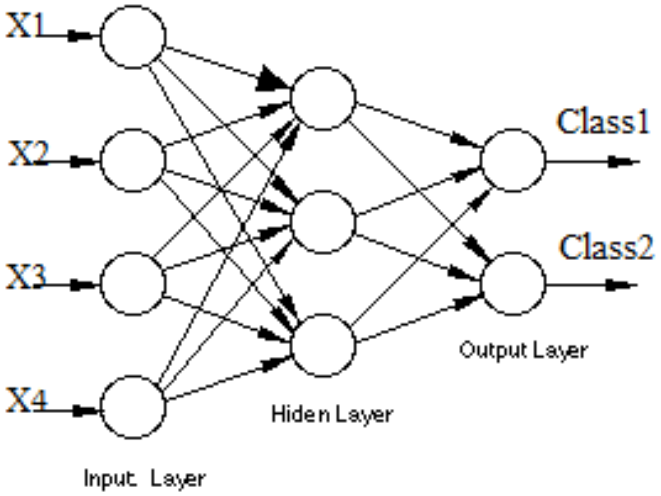

Figure 4. The structure of NN

3.4. Naive Bayes

Naive Bayes (NB) algorithm is used commonly in data mining and machine learning problems, especially in statistical problems [17]. The classification process is made taking advantage of statistical methods. We used following equation to calculate the probabilities Eq. 1

$$\frac{P(C1|X)}{P(C2|X)} > 1 \ , \qquad \frac{P(C1|X)}{P(C2|X)} = \frac{P(X|C1).P(C1)}{P(X|C2).P(C2)} \tag{1}$$

Firstly; A probability value is calculated , in order to decide which class they belong to. Depending on these calculated values of probability; the label of the class belonging to the highest probability value is given. For example, Figure 5 shows your new incoming S and each C1, C2, C3 labels also represent the class. Since our probability value on these labels is the highest C1, we say that new incoming S belongs to the C2 class.

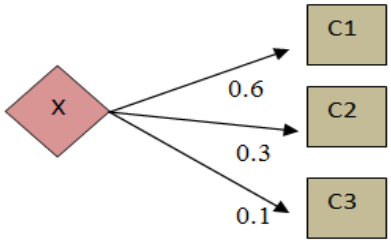

Figure 5. Example X data and Probability-Class Relations

3.5. Logistic Regression

Logistic Regression (LR), a popular mathematical modeling procedure used in the analysis of epidemiologic datasets, especially area of machine learning.

Logistic Regression method can be run in these steps [18]:

1. Calculate with the logistic function.
2. Learn the coefficients for a logistic regression model.
3. Finally, make predictions using a logistic regression model.

The logistic function is given below Eq.2.

$$f(x) = \frac{L}{1+e^{-k(x-x_0)}} \tag{2}$$

where;

e= Euler's number
x_0= Middle x-value of sigmoid function
L= The maximum value of curve
k= Abruptness of curve [19].
Input values (x) to estimate an output value (y); logistic regression equation is used. The logistic regression model is given in equation 3.

$$y = \frac{e^{b0+b1*x}}{1+e^{b0+b1x}} \quad (3)$$

Logistic regression parameters are estimated by maximizing logarithmic likelihood function using training data [20]. Figure 6 shows the example of a logistic regression for distinguish two classes (orange- yellow images).

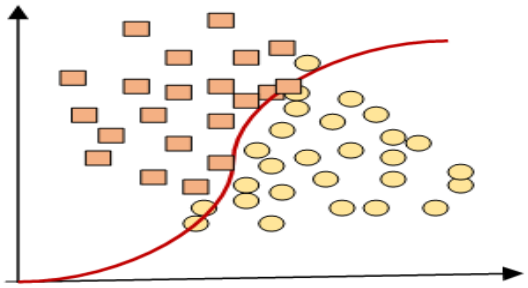

Figure 6. Logistic Regression Classification

IV. Application

The lung dataset consist of 178 observations. For the study, the data is divided into training data and test data. The training set is used to build the model of classifier and test set is used to confirm it. In this study, as trainning data and test data are used in 75% and 25% , respectively
Our dependent variable has two category so we begin by considering classification problems using only two classes. Formally, each instance I is mapped to one element of the set of positive and negative class labels. A classifier model is a mapping from instances to estimated classes[21].

	True Class	
Estimated Class	True Positives (TP)	False Positive (FP)
	False Negatives (FN)	True Negatives (TN)

Figure 6. Confusion Matrix

Among the observed true class, True positives considers the proportion of subjects who are "true positives (TP)" ,that is, correctly predicted as cases. Among the observed estimated class, Flase Positive considers the proportion of subjects who are "false positives (FP)", that is, are falsely predicted as cases. Classification performance measures are used which commonly used. Furthermore, statistical measurements also used that precision, F-measure, sensitivity (true positive rate) and specificity (true negative rate). We used following equation to precision Eq.4, F-measure Eq.5, measure the accuracy Eq. 6, specificity Eq. 7, sensitivity Eq. 8, Matthews Correlation Coefficient (MCC) Eq. 9, and Brier Score Eq. 10 [22].

Precision= TP / TP+FP (4)

$$F\text{-}Measure = 2/((1/Precision)+(1/Recall)) \quad (5)$$

$$Accuracy(\%) = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

$$Specificity(\%) = \frac{FP}{FP+TN} \quad (7)$$

$$Sensitivity(\%) = \frac{TP}{TP+FN} \quad (8)$$

$$MCC = \frac{(TP*TN)-(FP*FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (9)$$

$$Brier\ Score = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2 \quad (10)$$

f_t =Estimated probability
o_t =Observed Probability

A receiver operating curve (ROC) fort the fitted model and computing the area under the curve as a measure of discriminatory performance. The use of ROC has become popular in recent years because of the availability of computer software to conveniently produce such a curve as well as compute the area under the curve.ROC is a plot of sensitivity by 1-specificity values derived from several classification tables corresponding to different cut-points used to classify subjects into one of two-groups or more, e.g., predicted cases and noncases of a disease [23,24]. Fig.7 shows ROC graphs for all methods.

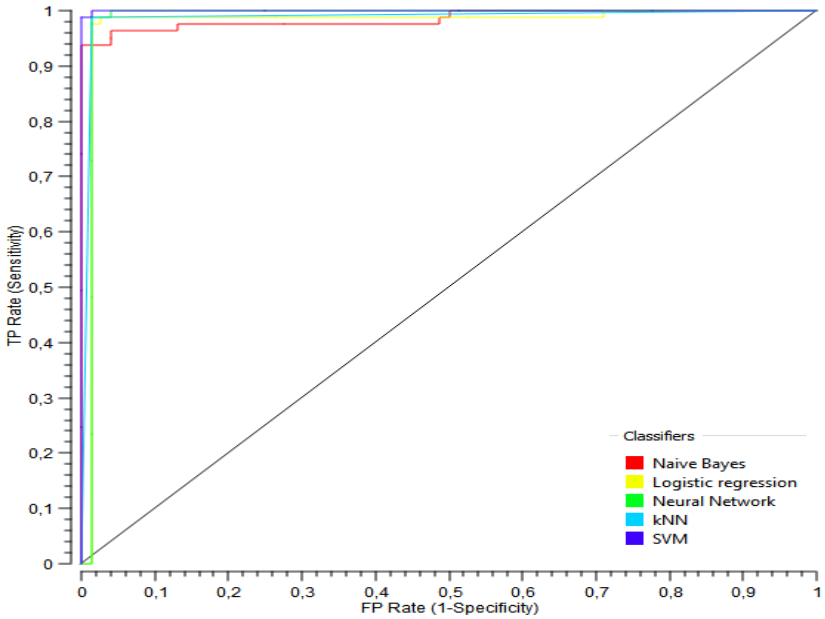

Figure 7. All Methods ROC Curves

V. Result

Table II shows the performance of KNN, SVM, NN and lojistic regression classifiers on lung cancer datasets. The highest classification accuracy is achieved with Support Vector Machine (SVM) by 99.3%.

Table II. The Performance of Classifiers

	Method	CA	Sens	Spec	AUC	F1	Prec	Brier	MCC
1	Naive Bayes	0.9618	0.9605	0.9630	0.9855	0.9605	0.9605	0.0722	0.9235
2	Logistic regression	0.9809	0.9737	0.9877	0.9778	0.9801	0.9867	0.0742	0.9618
3	Neural Network	0.9873	0.9868	0.9877	0.9868	0.9868	0.9868	0.0329	0.9745
4	kNN	0.9873	0.9868	0.9877	0.9873	0.9868	0.9868	0.0253	0.9745
5	SVM	0.9937	1.0000	0.9877	0.9997	0.9935	0.9870	0.0113	0.9873

VI. Conclusion

When you look at the performance chart; different results are produced for each classifier on the lung data set. Looking at the correct classification (CA) and other metrics; the best result is given by the support vector machine algorithms. SVM algorithm used high dimension to classify the observation so it's performance is the best. This technique can be applied on medical datasets to help physicians to make more accurate decisions about determination lung cancer. In addition, since the computer is supported, the person is independent (objective). Therefore, there is less mistakes. Finally, by adding extra preprocessing the correct rate can be enhanced.

References

[1]. Konsensusu, Antakya. Akciğer Kanseri. Diss. İstanbul Üniversitesi, 2010.
[2]. Göksel, Tuncay, et al. "Akciğer kanseri." Türkiye'de temel akciğer sağlığı sorunları ve çözüm önerileri. Ankara: Türk Toraks Derneği (2010): 55-70.
[3]. SERİLMEZ, Murat, et al. "Akciğer kanserinde hematolojik parametreler." Türk Onkoloji Dergisi 25.3 (2010): 87-92.
[4]. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), pp. 273–297.
[5]. B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, 5th Annual ACM Workshop on COLT, pp. 144-152, Pittsburgh, PA, 1992. ACM Press.
[6]. 1. K. Polat, S. Güneş, and A. Arslan, "A cascade learning system for classification of diabetes disease: Generalized Discriminant Analysis and Least Square Support Vector Machine", Experts System with Applications, Vol.34, 482-487, 2008.
[7]. A. J. Smola, and B. Schölkopf, "A tutorial on support vector regression", Statistics and Computing, Vol.14, pp.199–222, 2004.
[8]. K. Polat, S. Güneş, and A. Arslan, "A cascade learning system for classification of diabetes disease: Generalized Discriminant Analysis and Least Square Support Vector Machine", Experts System with Applications, Vol.34, 482-487, 2008.
[9]. AYHAN, Sevgi, and Şenol ERDOĞMUŞ. "Destek vektör makineleriyle sınıflandırma problemlerinin çözümü için çekirdek fonksiyonu seçimi."Eskişehir Osmangazi Üniversitesi İktisadi ve İdari Bilimler Dergisi 9.1 (2014).
[10]. M. W. Aslam, Z. Zhu, A.K. Nandi, "Feature generation using genetic programming with comparative partner selection for diabetes classification", Experts System with Applications, Vol 40, pp. 5402-5412, 2013.
[11]. R.Muralidharan, C. Chandrasekar, "Object Recognition using SVM-KNN based on Geometric Moment Invariant", International Journal of Computer Trends and Technology, Vol.1 (3), 2011.
[12]. V. Atmaca, "Örme Kumaşlardaki Üretim Hatalarının Görüntü İşleme Teknikleri İle Otomatik Tespiti ve Sınıflandırılması", İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, Y.Lisans Tezi, Haziran 2005.
[13]. P.-N. Tan, M. Steinbach, V. Kumar, "Introduction to Data Mining, International Edition", Pearson Education Inc., Boston, USA, 2006.
[14]. David Reby , Sovan Lek , Ioannis Dimopoulosc, Jean Joachim , Jacques Lauga ,Ste´phane Aulagnier, "Artificial neural Networks as a classification method in the behavioural sciences", Behavioural Processes 40 (1997) 35–43, 6 November 1996.
[15]. Fiona Nielsen, " Neural Networks – algorithms and applications", Neural Networks – algorithms and applications
[16]. 12/12-2001.
[17]. N. Bölücü, Z. Karhan, P. Duygulu Şahin, "Interactive Image Processing Based On Image Processing", 1st International Conference on Engineering Technology and Applied Sciences Afyon Kocatepe University, Turkey 21-22 April 2016.
[18]. Zhang H. , "The Optimality of Naive Bayes" , American Association for Artificial Intelligence, 2004.
[19]. http://machinelearningmastery.com/logistic-regression-tutorial-for-machine-learning/
[20]. https://en.wikipedia.org/wiki/Logistic_function
[21]. Yuksel Oner, Taner Tunc, Erol Egrioglu, Yildiz Atasoy, "Comparisons of Logistic Regression and Artificial Neural Networks in Lung Cancer Data",American Journal of Intelligent Systems 2013.
[22]. An introduction to ROC analysis Tom Fawcett Institute for the Study of Learning and Expertise, 2164 Staunton Court, Palo Alto, CA 94306, USA Available online 19 December 2005.
[23]. D. Delen, G. Walker, A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods", Artificial Intelligence in Medicine, Vol.34 (2), pp.113-127, 2005. Egan, J.P., 1975. Signal detection theory and ROC analysis, Series in Cognition and Perception. Academic Press, New York.
[24]. Fawcett, Tom. "An introduction to ROC analysis." Pattern recognition letters27.8 (2006): 861-874.