Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2016) 24: 3685 – 3697
© TÜBİTAK
doi:10.3906/elk-1501-181

Statistical learning approaches in diagnosing patients with nontraumatic acute abdomen

Gökmen ZARARSIZ[1,*], **Hızır Yakup AKYILDIZ**[2], **Dinçer GÖKSÜLÜK**[3], **Selçuk KORKMAZ**[3], **Ahmet ÖZTÜRK**[1]

[1]Department of Biostatistics, Faculty of Medicine, Erciyes University, Kayseri, Turkey
[2]Department of General Surgery, Faculty of Medicine, Erciyes University, Kayseri, Turkey
[3]Department of Biostatistics, Faculty of Medicine, Hacettepe University, Ankara, Turkey

Received: 24.01.2015 • **Accepted/Published Online:** 02.06.2015 • **Final Version:** 20.06.2016

Abstract: A quick evaluation is required for patients with acute abdominal pain. It is crucial to differentiate between surgical and nonsurgical pathology. Practical and accurate tests are essential in this differentiation. Lately, D-dimer level has been found to be an important adjuvant in this diagnosis and obviously outperforms leukocyte count, which is widely used for diagnosis of certain cases. Here, we handle this problem from a statistical perspective and combine the information from leukocyte count with D-dimer level to increase the diagnostic accuracy of nontraumatic acute abdomen. For this purpose, various statistical learning algorithms are considered and model performances are assessed using several measures. Our results revealed that the naïve Bayes algorithm, robust quadratic discriminant analysis, bagged and boosted support vector machines, and single and bagged k-nearest neighbors provide an increase in diagnostic accuracies of up to 8.93% and 17.86% compared with D-dimer level and leukocyte count, respectively. Highest accuracy was obtained as 78.57% with the naïve Bayes algorithm. Analysis has been done via the R programming language based on the codes developed by the authors. A user-friendly web-tool is also developed to assist physicians in their decisions to differentially diagnose patients with acute abdomen. It is available at http://www.biosoft.hacettepe.edu.tr/DDNAA/.

Key words: Abdominal pain, D-dimer level, decision support system, nontraumatic acute abdomen, statistical learning

1. Introduction

Acute abdomen is a term that refers to the clinical syndromes characterized by the sudden onset of abdominal pain symptoms and tenderness. These pains are mostly caused by appendicitis, cholecystitis, perforated peptic ulcer, bowel obstruction, diverticulitis, pancreatitis, urinary colic, and nonspecific and nonsurgical abdominal pains. It is one of the most common symptoms in emergency departments and requires rapid evaluation. Nearly 5% of the total patients presenting to emergency departments have acute abdominal pain. Deciding whether the source of the pain is from a surgical or nonsurgical pathology is crucial in the prevention of morbidity and mortality [1–3].

A brief history and complete physical examination are obligatory for proper diagnosis. The location of the pain may be an indicator in the diagnosis. Prompt radiographs and laboratory tests are usually helpful in the differential diagnosis, but are substantially time-consuming. In abdominal surgical pathologies, early diagnosis and management are the most important factors on the outcome. Unfortunately, there are no precise predictors of which patients have surgical pathology. In an emergency department with a high load of patients, it is almost

[*] Correspondence: gokmen.zararsiz@hacettepe.edu.tr

impossible to have enough time for a rapid and correct differential diagnosis. Due to this lack of time, the increased use of unnecessary computed tomography (CT) scans with intravenous contrast is a major concern. CT has some limitations as well as side effects and because of some other reasons like renal insufficiency, contrast allergy, etc., it cannot be implemented efficiently. Ultrasonography may be another solution. However it is not highly sensitive for every condition, and it is recommended for use in right upper quadrant pains. Magnetic resonance imaging (MRI) is an accurate method in the diagnosis of acute abdominal pain. However, its high cost and lack of immediate availability limit its use in acute care settings. It is obvious that we need alternative test(s) that is/are easy to perform and has/have high discriminating accuracy of surgical cases from nonsurgical ones. In the emergency department, once surgical pathology is confirmed with a high accuracy quick test, the surgeon will be involved in the diagnostic process without considerable time loss [3–9].

Recently, Akyıldız et al. [10] showed the efficiency and usefulness of the D-dimer test, which is also used in the diagnosis of pulmonary embolism, venous thromboembolism, disseminated intravascular coagulopathy, and intraabdominal pathologies. The authors demonstrated its performance over leukocyte count in the differential diagnosis of acute abdomen patients. They reported that D-dimer level provides better predictive performance than leukocyte count. A different solution may be to combine these two tests in a suitable way to improve the diagnostic performance. One way is to use “and/or rules” to decrease the number of false positive or false negative test results [11]. After identifying the cut-off values and assigning the positive or negative results for each test, the “and rule” defines the combined test results as positive only if both tests give positive results. Combined test results are negative in other cases and this rule is used to decrease the number of false positives. Conversely, the “or rule” is used to decrease the number of false negatives and defines the combined test results as negative only if both tests provide negative results. Although decreasing the number of false positives leads to an increase in specificity, it will decrease the sensitivity of the diagnostic test. Similarly, decreasing the number of false negatives will improve the sensitivity; however, it will decrease the specificity of the test. Clearly, there is a trade-off here and this type of combination is useless for the case where a physician considers both positive and negative results. Thus, it is necessary to use different combinations of approaches increase the general test performance instead of sensitivity or specificity measures.

In the last decade, statistical learning approaches have been used for this purpose and very successful results were obtained in the diagnosis of various medical problems. Bardella et al. successfully combined a serum IgA antigliadin antibodies assay and cellobiose/mannitol sugar permeability tests in a multiple logistic regression model for the diagnosis and screening of celiac disease [12]. Bozkurt et al. applied several algorithms and found that distributed time delay networks and probabilistic neural network classifiers performed best in predicting diabetes [13]. Chen et al. efficiently separated colon cancer and normal tissues using a random forests algorithm coupled with near-infrared spectroscopy [14]. Hundreds of similar examples in other medical examples can be found by a PubMed search using any of the following keywords: “statistical learning”, “data mining”, and “machine learning”.

In this study, we applied various statistical learning algorithms to combine both leukocyte count and D-dimer level measures for the purpose of improving the diagnostic accuracy of nontraumatic acute abdomen. We also developed a decision support tool to assist physicians in this differential diagnosis.

2. Methodology

2.1. Data collection

We used the data set from Akyıldız et al. [10] that contains data from patients admitted to the Erciyes University Medical Faculty’s General Surgery Department with the complaint of abdominal pain. Data include the leukocyte counts and D-dimer levels of 225 patients (115 females, 110 males) belonging to two groups. The first group had 115 (51.1%) patients who needed immediate laparotomy and the second group had 110 (49.9%) patients who did not need immediate laparotomy. Conventional treatment is assessed in this grouping and the patients operated on based on their postoperative pathologies are assigned to the first group, while the patients with a negative laparotomy are assigned to the second group. A scatter plot of the data is given in Figure 1. As seen from the plot, data are nested within each other and no simple rule is present to successfully discriminate the groups.

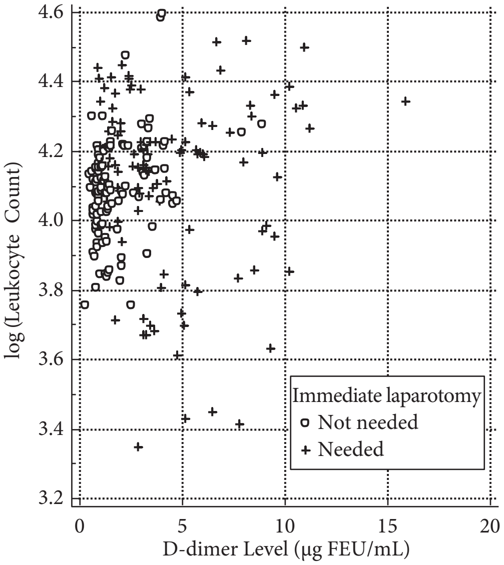


Figure 1. A scatter plot indicating the D-dimer levels and leukocyte counts of patients who need or do not need immediate laparotomy.

2.2. Statistical learning approaches

To discriminate the groups in a more advanced way, we benefited from the capabilities of various statistical learning algorithms. In this context, we used a number of discriminant classifiers, decision tree models, kernel-based classifiers, ensemble classification models, and some other models including logistic regression, naïve Bayes, neural networks, and k-nearest neighbors. We also considered leukocyte count and D-dimer tests separately to see the accuracy increase in the single diagnostic performances. In this section, we give a brief overview of these statistical learning models.

Discriminant classifiers aim to find class posterior probabilities for optimal classification. For this purpose, they use Bayes' theorem as follow:

$$\Pr\left(C = k \,|X = x\,\right) = \frac{f_k(x)\pi_k}{\sum\limits_{c=1}^{k} f_l(x)\pi_c} \tag{1}$$

Here, $f_k(x)$ is the class-conditional density function and π_k is the prior probability for class k. As a density function, linear discriminant analysis (LDA) uses multivariate Gaussian distribution such that, $f_k\left(x\right) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)}$. Here, p is the number of variables, μ_k is the sample mean vector,

Σ_k is the sample covariance matrix for class k, and LDA uses a common covariance matrix for each class. The discriminating function can be obtained by introducing the Gaussian density function into Eq. (1) and new test data will be assigned to the class based on this function or the posterior probabilities [15].

Other discriminant classifiers used in this study are extensions of LDA. In quadratic discriminant analysis, covariance matrices are assumed to be unequal for each class. Robust linear and robust quadratic discriminant analysis (RLDA, RQDA) use minimum covariance determinant robust estimators to estimate μ_k and Σ_k instead of using sample group means and covariance matrices. In mixture discriminant analysis, each class is modeled by a mixture of two or more Gaussian functions with different centroids. Flexible discriminant analysis (FDA) recasts the LDA problem as a nonparametric form of a linear regression problem. FDA uses scoring functions θ to assign discrimination scores to the classes and transformed class labels are predicted from linear regression on predictors X, such that $X^T\beta$. The scores are chosen to minimize the average squared residuals given in Eq. (2):

$$ASR = \frac{1}{N}\sum_{k=1}^{K}\left[\sum_{i=1}^{N}\left(\theta_k\left(y_i\right) - X^T\beta_k\right)^2\right] \tag{2}$$

where $\theta_k(y)$ is the discrimination score for class k and $X^T\beta_k$ is the regression mapping. FDA has the power of replacing regression fits with nonparametric alternatives to get more flexible classifiers than LDA [16–18].

Decision tree classifiers partition the feature space into a set of rectangles, apply simple models in each one, and display the results in a flowchart-like structure. In this structure, internal nodes correspond to features (diagnostic test), branches represent test results, and leaf nodes represent each class label. Classification rules can be obtained by these decision trees following the paths from root to leaf. Let $\hat{p}_{mk}$ be the proportion of class k in node m as given in Eq. (3):

$$\hat{p}_{mk} = \frac{1}{N_m}\sum_{x_i \in R_m} I(y_i = k) \tag{3}$$

where N_m is the number of observations in node m and R_m represents the region of the node. The observations in node m are classified into the class with maximum probability satisfying the condition $k\left(m\right) = \text{argmax}_k\hat{p}_{mk}$. The splitting and stopping rule is determined by the impurity of the corresponding node. The measure of impurity can be specified with several measures including the misclassification error, Gini index, and cross entropy or deviance. For binary classification problems, these measures simplify to $1 - \max\left(p, 1-p\right)$, $2p\left(1-p\right)$, and $p\log p - (1-p)\log?(1-p)$, respectively [15]. The CART and J48, also called C4.5, algorithms are among the commonly used algorithms to build the decision tree. Both algorithms grow the full tree and then prune it back to control overfitting. A considerable difference between these two algorithms is that CART uses a holdout method to build the tree and allows only binary splitting rules, while C4.5 uses the entire data set to build the final tree and uses multiple splitting rules. C5.0 is an improved version of the J48 algorithm with several advantages: it is faster, it provides smaller decision trees, it is more memory-efficient, it supports boosting to improve the performance, it allows the user to weight cases, and it winnows the useless features automatically. Conditional inference trees (CTrees) avoid the variable selection bias and use significant testing instead of maximizing the information gain or Gini coefficient [15,17,19].

Kernel-based classifiers are preferred approaches when the data classes are not linearly separable. The aim here is to use kernel functions and map the data into a high-dimensional feature space to make linear models work in nonlinear settings. Support vector machines (SVMs) are one of popular statistical learning tools due to various advantages: having a strong mathematical background based on statistical learning theory, an accurate performance in problems from various fields, and the capacity to handle high-dimensional data. The SVM aims to find the optimal separating hyperplane, which maximizes the margin between classes. The closest data points to this hyperplane are called support vectors and the margin is the distance between these support vectors. Quadratic programming and Lagrange multipliers (Ψ_i) are used to find the optimal hyperplane. For nonlinear classification problems, the SVM uses kernel functions $\Phi(.)$ such as the radial-basis function (RBF) and polynomial functions. The classification function of a SVM is given in Eq. (4):

$$f\left(x\right) = sgn\left(\sum_{i=1}^{n} \Psi_i y_i \Phi\left(x_i\right)\Phi\left(x\right) + b\right) \tag{4}$$

In our problem, we considered the three cases of SVM with linear, RBF, and polynomial kernels as SVMlin, SVMrbf, and SVMpoly, respectively. Least squares support vector machines (lsSVM) are a modified form of SVM. The difference is that lsSVM solves a linear system instead of quadratic programming in the parameter optimization process. We included lsSVM with linear and RBF kernel functions in this study as lsSVMlin and lsSVMrbf, respectively. Conversely to the SVM algorithm, partial least squares (PLS) projects the data down to a few principal factors, explaining the maximum variance of both independent variables and the response simultaneously. After the projection, PLS classifies the data using linear classifiers [15,20,21].

Like discriminant classifiers, the naïve Bayes (NB) classifier also uses Bayes' theorem to estimate the posterior probabilities for each sample to determine which class to assign. However, this algorithm considers each feature independently to contribute to classification. The joint probability density function that is used in Bayes' formula for a given class k is found as in Eq. (5).

$$f_k\left(X\right) = \prod_{i=1}^{p} f_{kj}(X_j) \tag{5}$$

Posterior probabilities for each class are obtained by introducing Eq. (5) into Bayes' formula and subjects are assigned into one of the classes with respect to posterior class probabilities. Logistic regression (LR) is also a probabilistic model that uses a logistic function to discover the relationship between dependent and independent variables. A multiple logistic regression model can be written as in Eq. (6).

$$\Pr\left(C = k\left|X = x\right.\right) = \frac{e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p)}}{1+e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p)}} \tag{6}$$

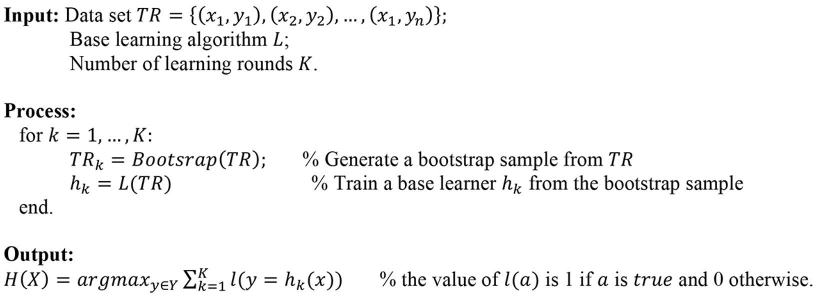
The maximum-likelihood approach is usually preferred to estimate model parameters with both NB and LR classifiers. The neural network (NN) classifier is an algorithm inspired by the central nervous system and brain and similarly the algorithm structure consists of interconnected neurons. NN takes the input data as input neurons and uses some functions to weight and transform the data. Activation is passed from one to another neuron until an output neuron is activated. The NN is specified by weights w_i and bias b:

$$y = \sum_{i=1}^{M} w_i x_i + b \tag{7}$$

The k-nearest neighbor (KNN) classifier is a lazy learner where the input contains the $k$closest training data in the feature space and the output consists of the class labels. In the classification process, training data are classified into the class that is most common in its k-nearest neighbors [15]. One of the most commonly used distance measures is Euclidean distance, such that $d_{ij} = \|x_i - x_j\|$, i.e. the norm of the vector. It is recommended to standardize each of the features to have zero mean and variance 1 before obtaining Euclidean distances since each feature might be measured in different units. The decision function of the KNN classifier is as follows:

$$y\left(d_i\right) = argmax_k \sum_{x_j \in kNN} y(x_j, c_k) \tag{8}$$

Ensemble methods apply multiple models instead of using a single model in order to improve the diagnostic accuracy. Bagging and boosting are among the most common types of ensembles. Bagging, also known as bootstrap aggregating, generates multiple bootstrap data sets from the training data, trains the data using a classifier, and combines the results of each model in a convenient way such as majority voting technique. The analogy of the bagging algorithm is given in Figure 2. The random forest (RF) algorithm is an example of a bagging ensemble that combines single decision tree models to achieve higher diagnostic accuracy. Similarly to bagging, boosting also resamples the data, creates an ensemble of single classifiers, and aggregates the results using majority voting. The difference is that boosting sequentially produces multiple models by giving higher weights to misclassified cases. Similarly to random forests, decision tree classifiers can be ensembled with the boosting approach as well. A boosted tree (boostTree) is used in such cases. Accordingly, bagged logistic regression (bagLR), bagged support vector machines (bagSVM), and bagged k-nearest neighbors (bagKNN) are bagging ensembles of LR, SVM, and KNN classifiers; boosted logistic regression (boostLR) and boosted support vector machines (boostSVM) are boosting ensembles of LR and SVM, respectively [15,17,22,23]. Further details about these algorithms can be found in the referenced papers.



```
Input: Data set $TR = \{(x_1, y_1), (x_2, y_2), \dots, (x_1, y_n)\}$;
       Base learning algorithm $L$;
       Number of learning rounds $K$.

Process:
  for $k = 1, \dots, K$:
      $TR_k = Bootsrap(TR)$;     % Generate a bootstrap sample from $TR$
      $h_k = L(TR)$               % Train a base learner $h_k$ from the bootstrap sample
  end.

Output:
$H(X) = argmax_{y \in Y} \sum_{k=1}^{K} l(y = h_k(x))$      % the value of $l(a)$ is 1 if $a$ is $true$ and 0 otherwise.
```

Figure 2. Pseudocode of bagging algorithm.

2.3. Model building and performance assessment

A logarithmic transformation (base 10) is applied to leukocyte counts. Next the data are centered and scaled using z-score transformation. Before applying the algorithms, data are randomly split into two parts as 75% and 25%, respectively. The first part is called the training set, which includes 169 patients and is used for model building and parameter optimization. The second part is called the validation set; it includes 56 patients and is used for performance assessment. In the training set, 10-fold cross-validation is used and repeated 10 times to find the optimal parameters of each algorithm with a grid search and to generalize the results. In detail, first the training data are partitioned into ten equal parts (17 samples exist in each part), next the first nine folds (153 samples) are used for model building and the last fold is used to test the model, and finally this process is repeated ten times with each fold used exactly once as test data. The validation set is considered as an unknown separate data set. Trained statistical learning models are applied to this data set and the performance of each model is assessed here. All model building processes are applied in the caret package of R software (http://www.R-project.org/) version 3.1.1 [24].

Bootstrap and boosting numbers are set at 100 for ensemble models. The RBF kernel function is used in bagSVM models. In RF modeling, 500 trees are used in model building. The number of neighbors is defined as 5 and the Euclidean distance metric is used in KNN modeling. Probability threshold value is set at 0.5 in LR models. Complexity parameter is obtained as 0.398 in optimal CART modeling. Sigma and complexity parameters are optimized to 1.38 and 0.25 for SVM models, respectively. The J48 confidence parameter is identified as 0.25. In FDA modeling, product degree and number of terms are identified as 1 and 3, respectively. Numbers of hidden layers and weight decay values for the NN were 15 and 0.316%, respectively.

To assess the performance of each model, several statistical diagnostic measures are calculated including accuracy rate (true classification rate), sensitivity, specificity, positive predictive value, negative predictive value, detection rate, balanced accuracy rate, F-score, Matthews correlation coefficient, and Kappa statistic. Details of the calculation of these statistics are given in Tables 1 and 2. For a better diagnostic test performance, each measure should be maximized.

Table 1. A 2 × 2 classification contingency table (confusion matrix).

Diagnostic test result	Actual result (gold standard)		Total
	Positive	Negative	
Positive	TP	FP	TP+FP
Negative	FN	TN	FN+TN
Total	TP+FN	FP+TN	n

Table 2. Calculation of diagnostic measures used in this study.

Diagnostic measure	Calculation
Accuracy rate	$ACC = (TP + TN)/n$
Sensitivity	$SEN = TP/(TP + FN)$
Specificity	$SPE = TN/(FP + TN)$
Positive predictive value	$PPV = TP/(TP + FP)$
Negative predictive value	$NPV = TN/(TN + FN)$
Detection rate	$DR = TP/n$
Balanced accuracy rate	$bACC = (SEN + SPE)/2$
F-score	$F1S = 2TP/(2TP + FP + FN)$
Matthews correlation coefficient	$MCC = \frac{TPxTN - FPxFN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
Kappa statistic	$\kappa = (ACC - p_e)/(1 - p_e)$ where, $p_e = ((TP + FN)\,(TP + FP) + (FP + TN)(FN + TN))/n^2$

After calculating the diagnostic performances of each algorithm, a hierarchical clustering method is also applied to cluster the used algorithms based on their diagnostic performances. Here, the calculated diagnostic measures for each algorithm, as shown in Table 2, were taken as input to the hierarchical clustering method. The Euclidean distance metric and Ward method are used in this clustering process. Furthermore, area under receiver operating characteristic (AUROC) curves were also calculated to identify the best performing algorithms in the differential diagnosis of acute abdomen patients.

3. Results

Classification results and the computational costs for the related models (MacOS, 2.7 GHz quad-core CPU with 16 GB memory) are given in Table 3. D-dimer level outperformed leukocyte count with a 20.00% increase in sensitivity and 8.93% increase in accuracy. The performance of decision tree algorithms was not obviously better than a single D-dimer level diagnostic test. Only the CART algorithm made a slight improvement of 3.23% in specificity and 1.79% in accuracy compared to D-dimer level. Discriminant and kernel-based classifiers performed sufficiently well compared to single tests. Using robust estimates made a low increase in QDA and a decrease in the LDA classifier. Quadratic analysis (QDA and RQDA) gave the highest accuracies in discriminant analysis and least square SVMs (lsSVMlin and lsSVMrbf) gave the highest accuracies in kernel-based classifiers. SVMpoly's performance was poor because of its very low sensitivity results. LR's performance was similar to D-dimer level, but LR's sensitivity was lower and its specificity was higher. The NN made a tolerable improvement with 73.21% accuracy, 72.00% sensitivity, and 74.19% specificity. Performances of KNN and NB were satisfactory with 77.50% and 78.57% accuracy rates. Diagnostic accuracy of ensemble classifiers varies depending on the method used. It is seen that SVM ensembles perform quite well and improve the performance of single SVM learners. BoostLR's results were the same, while bagLR made a very slightly increase in the performance of LR. Bagging also worked for the KNN algorithm and made a 1.19% improvement in sensitivity and 0.59% improvement in accuracy. Conversely, bagging did not work, but boosting improved the diagnostic accuracy for decision tree classifiers. A 73.93% accuracy rate was obtained for the boostTree algorithm, which was superior to CART's 71.43% and RF's 67.86% accuracy results. Results were similar for other general performance measures.

A dendrogram of clustering results is given in Figure 3a and AUROC curves of each classifier are given in Figure 3b. Here, we defined six clusters by examining the dendrogram. One can see that the best performing diagnostic tests, NB, RQDA, bagSVM, boostSVM, bagKNN, and KNN, are grouped in cluster V. Accordingly, cluster IV tests (QDA and lsSVMlin) performed second best, cluster II tests (MDA, boostTree, lsSVMrbf, and NN) performed third best, cluster III tests (CART, FDA, CTree, D-dimer, C5.0, J48, and SVMrbf) performed fourth best, and cluster I tests (PLS, RF, bagLR, boostLR, LR, LDA, RLDA, and SVMlin) performed fifth best in the differential diagnosis of nontraumatic acute abdomen. It is seen that all decision tree classifiers are grouped in cluster III. Leukocyte count and SVMpoly tests gave the worst performances and were grouped in cluster VI. As seen from Figure 3b, the best performing classifiers in cluster V have the highest AUROC values.

Algorithms in cluster V are found to be the best performing classifiers with 76.79%–78.57% accuracy and 0.536–0.562 kappa statistics. When compared to D-dimer level and leukocyte count, they made an increase in diagnostic accuracies of 7.15%–8.93% and 16.08%–17.86%, respectively. ROC curves indicating the predictive performance of the two best performers, NB and bagSVM, and also single D-dimer level and leukocyte count tests are given in Figure 4. AUROC curves and 95% binomial exact confidence intervals were 0.88 (0.81–0.92), 0.87 (0.80–0.91), 0.82 (0.75–0.87), and 0.63 (0.55–0.71), respectively, and all pairwise comparisons except between NB and bagSVM were found as statistically significant based on the z test ($P < 0.05$).

Moreover, for the applicability of the best performing statistical learning approaches, we have developed the DDNAA web-tool to assist physicians' decisions in the differential diagnosis of nontraumatic acute abdomen. DDNAA can be accessed from http://www.biosoft.hacettepe.edu.tr/DDNAA/. A snapshot of the DDNAA is

Table 3. Performance assessment of various statistical learning algorithms in diagnosing patients with nontraumatic acute abdomen.

Classification model	ACC (%)†	SEN (%)	SPE (%)	PPV (%)	NPV (%)	DR (%)	bACC (%)	F1S (%)	MCC (%)	κ	Time (s)
Single classifiers											
log(Leukocyte count)	60.71	60.00	61.29	55.56	65.52	26.79	60.36	57.69	21.18	0.211	-
D-dimer	69.64*	**80.00**	61.29	62.50	79.17	35.71	70.65	70.18	41.48	0.402**	-
Discriminant classifiers											
Linear discriminant analysis	73.21**	64.00	80.65	72.73	73.53	28.57	72.32	68.09	45.44	0.452***	0.993
Robust linear discriminant analysis	71.43*	72.00	70.97	66.67	75.86	32.14	71.48	69.23	42.75	0.426***	4.909
Quadratic discriminant analysis	76.79***	56.00	**93.55**	**87.50**	72.50	25.00	74.77	68.29	54.52	0.513***	0.966
Robust quadratic discriminant analysis	76.79***	**80.00**	74.19	71.43	**82.14**	35.71	77.10	75.47	53.88	0.536***	3.230
Mixture discriminant analysis	73.21**	**80.00**	67.74	66.67	80.77	35.71	73.87	72.73	47.59	0.468***	31.421
Flexible discriminant analysis	69.64*	76.00	64.52	63.33	76.92	33.93	70.26	69.09	40.39	0.398**	11.392
Decision tree classifiers											
Classification and regression trees	71.43*	**80.00**	64.52	64.52	80.00	**44.64**	72.26	71.43	44.52	0.435***	4.643
C5.0	69.64*	**80.00**	61.29	62.50	79.17	35.71	70.65	70.18	41.48	0.402**	17.436
J48	69.64*	**88.00**	54.84	61.11	**85.00**	39.29	71.42	72.13	44.44	0.411**	9.214
Conditional inference tree	69.64*	**80.00**	61.29	62.50	79.17	35.71	70.65	70.18	41.48	0.402**	18.880
Kernel-based classifiers											
Support vector machines with linear kernel	71.43*	68.00	74.19	68.00	74.19	30.36	71.10	68.00	42.19	0.422***	1.562
Support vector machines with polynomial kernel	64.40*	40.00	**89.29**	79.07	59.52	20.12	64.65	53.12	33.62	0.292**	33.693
Support vector machines with radial basis function kernel	71.43*	**84.00**	61.29	63.64	**82.61**	37.50	72.65	72.41	45.77	0.439***	12.043
Partial least squares	69.64*	60.00	77.42	68.18	70.59	26.79	68.71	63.83	38.09	0.379**	1.347
Least squares support vector machines with linear kernel	75.00**	56.00	**90.32**	**82.35**	71.79	25.00	73.16	66.67	50.08	0.478***	62.325
Least squares support vector machines with radial basis function kernel	75.00**	72.00	77.42	72.00	77.42	32.14	74.71	72.00	49.42	0.494***	90.499
Ensemble classifiers											
Random forests	67.86*	60.00	74.19	65.22	69.70	26.79	67.10	62.50	34.56	0.345**	7.711
Bagged logistic regression	70.40*	64.71	76.19	73.33	68.09	32.54	70.45	68.75	41.16	0.409**	<0.100
Bagged support vector machines	**78.12*****	**80.00**	76.19	77.27	79.01	**40.24**	**78.10**	**78.61**	**56.24**	**0.562*****	96.344
Bagged k-nearest neighbors	78.09***	74.12	82.14	80.77	75.82	37.28	**78.13**	**77.30**	**56.43**	**0.562*****	32.520
Boosted trees	73.93**	72.94	75.00	74.70	73.26	36.69	73.97	73.81	47.95	0.479***	151.040
Boosted logistic regression	69.82*	63.53	76.19	72.97	67.37	31.95	69.86	67.92	40.03	0.397**	<0.100
Boosted support vector machines	**78.12*****	**80.00**	76.19	77.27	79.01	**40.24**	**78.10**	**78.61**	**56.24**	**0.562*****	99.355
Other classifiers											
Logistic regression	69.82*	63.53	76.19	72.97	67.37	31.95	69.86	67.92	40.03	0.397**	<0.100
Naïve Bayes	**78.57*****	68.00	87.10	**80.95**	77.14	30.36	77.55	73.91	**56.58**	0.560***	2.807
Neural networks	73.21**	72.00	74.19	69.23	76.67	32.14	73.10	70.59	46.05	0.460***	159.523
k-Nearest neighbors	77.50***	74.12	80.95	79.75	75.56	37.28	77.54	76.83	55.19	0.551***	4.065

*P <0.05, **P <0.01, ***P <0.001. ACC: Accuracy rate, SEN: sensitivity, SPE: specificity, PPV: positive predictive value, NPV: negative predictive value, DR: detection rate, bACC: balanced accuracy rate, F1S: F-score, MCC: Matthews correlation coefficient, κ: kappa statistic. †A one-sided hypothesis test is computed to evaluate whether the overall accuracy rate is greater than the rate of the largest class. Bold values indicate the top three winner algorithms for each performance measure.

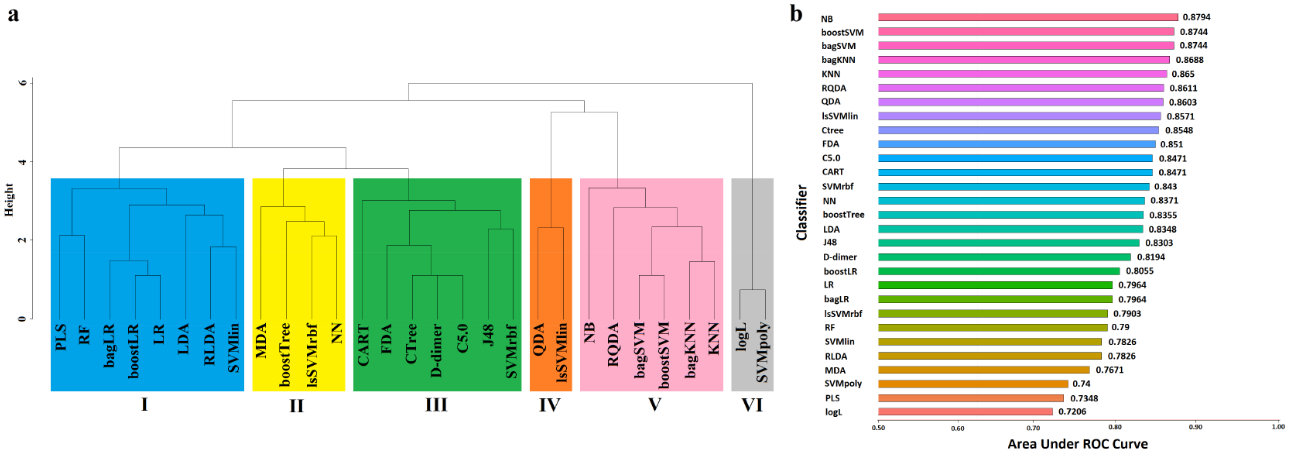


Figure 3. (a) Clustering results for statistical learning algorithms based on their diagnostic performance measures. (b) Area under ROC curves for each classifier. PLS: partial least squares, RF: random forests, bagLR: bagged logistic regression, boostLR: boosted logistic regression, LR: logistic regression, LDA: linear discriminant analysis, RLDA: robust linear discriminant analysis, SVMlin: support vector machines with linear kernel function, MDA: mixture discriminant analysis, boostTree: boosted tree, lsSVMrbf: least squares support vector machines with radial-based kernel function, NN: neural networks, CART: classification and regression trees, FDA: flexible discriminant analysis, C-Tree: conditional trees, SVMrbf: support vector machines with radial-based kernel function, QDA: quadratic discriminant analysis, lsSVMlin: least squares support vector machines with linear kernel function, NB: naïve Bayes, RQDA: robust quadratic discriminant analysis, bagSVM: bagged support vector machines, boostSVM: boosted support vector machines, bagKNN: bagged k-nearest neighbors, KNN: k-nearest neighbors, logL: $\log_{10}$(leukocyte count), SVMpoly: support vector machines with polynomial kernel function.

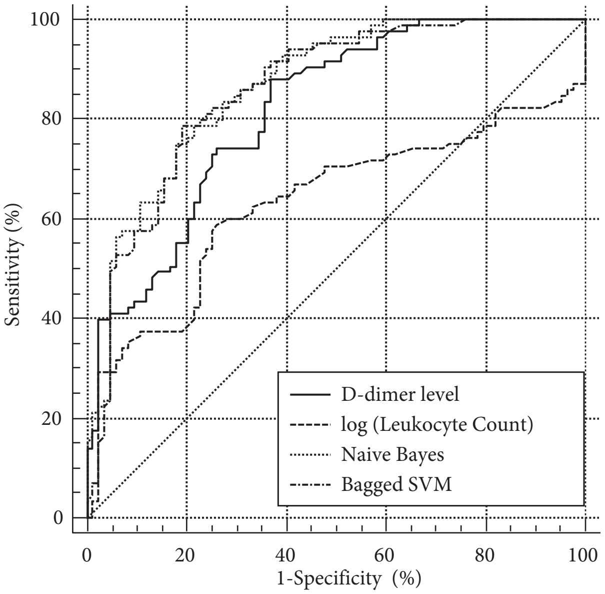
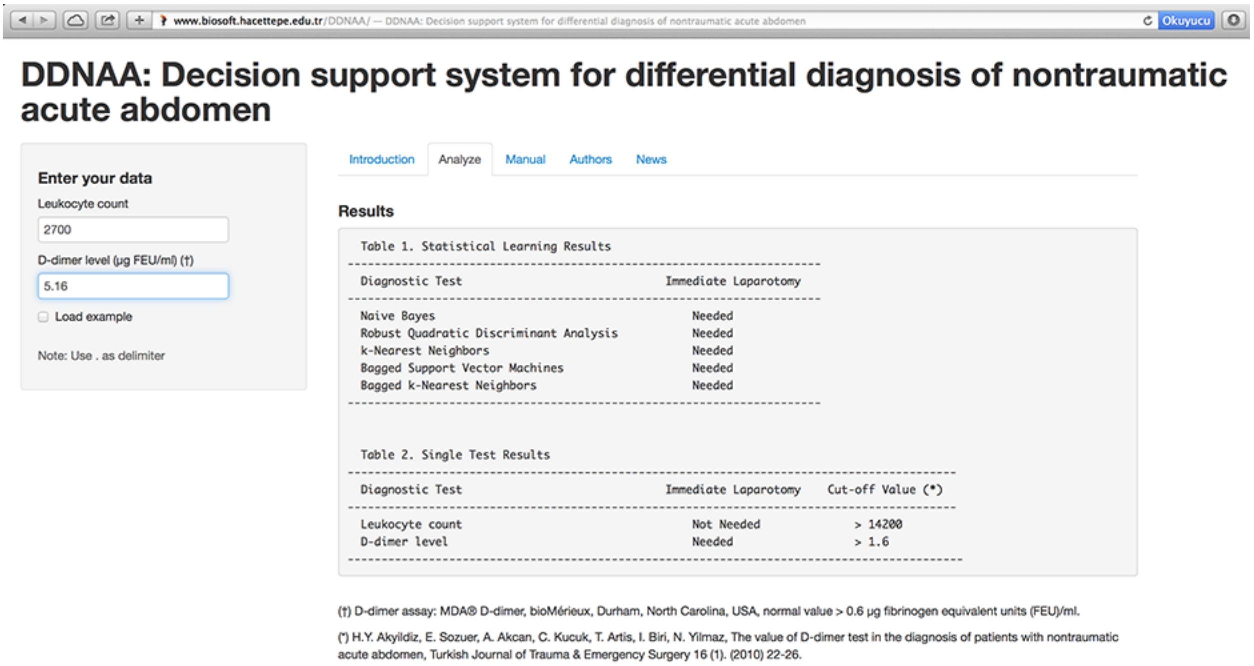


Figure 4. A ROC curve demonstrating the predictive performances of naïve Bayes, bagged support vector machines, D-dimer level, and leukocyte count in differential diagnosis of nontraumatic acute abdomen.

given in Figure 5. Users can simply enter the patient's data and get multiple results from the best performing diagnostic tests mentioned in this study. They can also see the results for uncombined leukocyte count and D-dimer level tests.


www.biosoft.hacettepe.edu.tr/DDNAA/ — DDNAA: Decision support system for differential diagnosis of nontraumatic acute abdomen
Okuyucu
Introduction
Analyze
Manual
Authors
News

DDNAA: Decision support system for differential diagnosis of nontraumatic acute abdomen

Enter your data

Leukocyte count: 2700

D-dimer level (µg FEU/ml) (†): 5.16

□ Load example

Note: Use . as delimiter

Results

```
Table 1. Statistical Learning Results
----------------------------------------------------------------------
Diagnostic Test                              Immediate Laparotomy
----------------------------------------------------------------------
Naive Bayes                                      Needed
Robust Quadratic Discriminant Analysis           Needed
k-Nearest Neighbors                              Needed
Bagged Support Vector Machines                   Needed
Bagged k-Nearest Neighbors                       Needed
----------------------------------------------------------------------


Table 2. Single Test Results
-----------------------------------------------------------------------------------------
Diagnostic Test                              Immediate Laparotomy    Cut-off Value (*)
-----------------------------------------------------------------------------------------
Leukocyte count                                  Not Needed              > 14200
D-dimer level                                    Needed                  > 1.6
-----------------------------------------------------------------------------------------
```

(†) D-dimer assay: MDA® D-dimer, bioMérieux, Durham, North Carolina, USA, normal value > 0.6 µg fibrinogen equivalent units (FEU)/ml.

(*) H.Y. Akyildiz, E. Sozuer, A. Akcan, C. Kucuk, T. Artis, I. Biri, N. Yilmaz, The value of D-dimer test in the diagnosis of patients with nontraumatic acute abdomen, Turkish Journal of Trauma & Emergency Surgery 16 (1). (2010) 22-26.

Figure 5. A snapshot of the DDNAA web-tool.

4. Discussion and conclusions

The decision of whether the source of acute onset of abdominal pain symptoms is a surgical or nonsurgical pathology is vital in surgery and emergency departments and requires rapid evaluation. Prompt radiographs, laboratory tests, and CT scans are useful in the differential diagnosis. However, prompt radiographs and laboratory tests are time-consuming and CT scans have several limitations including side effects, contrast allergy, and renal insufficiency, which are frequent in city hospitals. A fast, easy-to-use, and accurate diagnostic test is strongly required in this diagnosis.

D-dimer level is a simple and noninvasive triage test that measures the concentration of fibrin degradation products. This test is a good marker for many conditions including venous thromboembolism, sudden arterial occlusion of lower extremities, acute aortic dissection, pulmonary embolism, acute ischemic stroke, symptomatic abdominal aortic aneurysm, pregnancy, eclampsia, severe trauma, liver disease, cancer, and recent surgery. Akyıldız et al. showed that it can also be a better marker than leukocyte count in differential diagnosis of nontraumatic acute abdomen [10]. Even though D-dimer level is a simple-to-use test and performs better than leukocyte count, its accuracy is not quite good enough. Even a 1% increase in diagnostic accuracy is vital when the prevention of diseases and mortality is considered. With the purpose of increasing the diagnostic accuracy, we investigated the use of statistical learning approaches in this problem and obtained good results. These statistical learning approaches are multivariate methods with different mathematical backgrounds and they gain information from each single diagnostic test.

Statistical learning approaches have received great interest from the research community in the diagnosis of various health problems. Silva et al. applied various statistical learning approaches to combine spectral domain OCT (SD-OCT) and standard automated perimetry (SAP) for glaucoma diagnosis. The authors obtained the highest predictive performance with the RF classifier and improved the sensitivity and specificity of single tests [25]. Shankle et al. applied these approaches to combine cognitive and functional skills to improve dementia screening. For this purpose, the authors applied statistical learning approaches and gained 13%–24% accuracy increase over the single Functional Activities Questionnaire (FAQ) and the Six-Item Blessed, Orientation, Memory and Concentration Exam (BOMC) tests [26].

We found that D-dimer level is superior to leukocyte count in differential diagnosis of nontraumatic acute abdomen, as consistent with [10]. Additionally, combining the information from D-dimer level and leukocyte count diagnostic tests from a statistical perspective, statistical learning approaches made an increase in the accuracy of up to 8.93% as compared to D-dimer level. Besides accuracy rates, ROC analysis showed that the predictive performances of best performing classifiers NB and bagSVM were significantly higher than both D-dimer level and leukocyte counts. Thus, we think that using combined tests with statistical learning is more reliable and accurate for a better diagnosis. We also think that adding other diagnostic tests to these statistical learning models may give more accurate results.

With advantages such as being fast (less than 1s), easy to use, and containing statistical learning algorithms that have strong mathematical backgrounds and accurate performances, the DDNAA web-tool will assist physicians in their decision to differentially diagnose nontraumatic acute abdomen, and this decision support will lead to better treatments and decrease in morbidity and mortality.

An important point here is the presence of various commercial D-dimer assays such as AMAX, AutoDimer, D-Dimer Plus, IL Test, Miniquant, MDA, NycoCard, and VIDAS. Thus, results may show variability based on the assay used and other hospital settings. Researchers should note that D-dimer concentration in our data was obtained with the quantitative immunofiltration assay method (MDA®D-dimer, bioMérieux, Durham, NC, USA, normal value >0.6 μ g fibrinogen equivalent units (FEU)/mL). In the case of the presence of other assays, they should standardize the D-dimer levels based on the calibrators provided by manufacturers. We leave the diagnostic effect of other D-dimer assays along with leukocyte count as a further research topic.

References

[1] Arora B, Gupta A, Nandi S, Sarwal A, Goyal P, Gogna S, Karwasra RK. Comparative analysis of clinical, radiological and operative findings in acute abdomen. Int J Enhanc Res Med Dent Care 2015; 2: 3-6.

[2] Stoker J, van Randen A, Lameris W, Boermeester MA. Imaging patients with acute abdominal pain. Radiology 2009; 253: 31-46.

[3] Marincek B. Nontraumatic abdominal emergencies: acute abdominal pain: diagnostic strategies. Eur Radiol 2002; 12: 2136-2150.

[4] Fahel E, Amaral PCG, Filho EMA, Ettinger JETM, Souza ELQ, Fortes MF, Alcantara RSM, Regis AB, Neto MPG, Sousa MM et al. Non-traumatic acute abdomen: videolaparoscopic approach. JSLS-J Soc Laparoend 1999; 3: 187-192.

[5] Akyildiz HY, Akcan A, Oztürk A, Sozuer E, Kucuk C, Yucel A. D]dimer as a predictor of the need for laparotomy in patients with unclear non]traumatic acute abdomen. A preliminary study. Scand J Clin Lab Inv 2008; 68: 612-617.

[6] Maglinte DD, Gage SN, Harmon BH, Kelvin FM, Hage JP, Chua GT, Ng AC, Graffis RF, Chernish SM. Obstruction of the small intestine: accuracy and role of CT in diagnosis. Radiology 1993; 188: 61-64.

[7] Frager D, Medwid SW, Baer JW, Mollinelli B, Friedman M. CT of small-bowel obstruction: value in establishing the diagnosis and determining the degree and cause. Am J Roentgenol 1994; 162: 37-41.

[8] Flasar MH, Goldberg E. Acute abdominal pain. Med Clin N Am 2006; 90; 481-503.

[9] Cartwright SL, Knudson MP. Evaluation of acute abdominal pain in adults. Am Fam Physician 2008; 77: 971-978.

[10] Akyıldız HY, Sözüer E, Akcan A, Küçük C, Artis T, Biri İ, Yılmaz N. The value of D-dimer test in the diagnosis of patients with non-traumatic acute abdomen. Ulus Travma Acil Cer 2010; 16: 22-26.

[11] Marshall RJ. The predictive value of simple rules for combining two diagnostic tests. Biometrics 1989; 45: 1213-1222.

[12] Bardella MT, Molteni N, Cesana B, Baldassarri AR, Binanchi PA. IgA antigliadin antibodies, cellobiose/mannitol sugar test, and carotenemia in the diagnosis of and screening for celiac disease. Am J Gastroenterol 1991; 86: 309-311.

[13] Bozkurt MR, Yurtay N, Yılmaz Z, Sertkaya C. Comparison of different methods for determining diabetes. Turk J Electtr Eng Co 2014; 22: 1044-1055.

[14] Chen H, Lin Z, Wu H, Wang L, Wu T, Tan C. Diagnosis of colorectal cancer by near-infrared optical fiber spectroscopy and random forest. Spectrochim Acta Mol Biomol Spectrosc 2015; 135: 185-191.

[15] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference and Prediction. 2nd ed. New York, NY, USA: Springer, 2009.

[16] Todorov V, Filzmoser P. An object oriented framework for robust multivariate analysis. J Stat Softw 2009; 32: 1-47.

[17] Tan PN, Steinbach M, Kumar V. Introduction to Data Mining, 1st ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., 2005.

[18] Öztürk A, Özdamar K. Comparison of linear, quadratic and flexible discriminant analysis by using generated and real data. Erciyes Med J 2008; 30: 266-277.

[19] Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. J Comp Graph Stat 2006; 15: 651-674.

[20] Vapnik V. The Nature of Statistical Learning Theory. 2nd ed. New York, NY, USA: Springer-Verlag, 1995.

[21] Pochet N, Smet FD, Suykens JAK, De Moer BLR. Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. Bioinformatics 2004; 20: 3185-3195.

[22] Breiman L. Bagging predictors. Mach Learn 1996; 24: 123-140.

[23] Dietterich TG. Ensemble methods in machine learning. In: Proceedings of the 1st International Workshop on Multiple Classifier Systems; 21—23 June 2000; Cagliari, Italy. pp. 1-15.

[24] Kuhn M. Building predictive models in R using the caret package. J Stat Soft 2008; 28: 1-26.

[25] Silva FR, Vidotti VG, Cremasco F, Dias M, Gomi ES, Costa VP. Sensitivity and specificity of machine learning classifiers for glaucoma diagnosis using spectral domain OCT and standard automated perimetry. Arq Bras Oftalmol 2013; 76: 170-174.

[26] Shankle WR, Datta P, Dillencourt M, Pazzani M. Improving dementia screening tests with machine learning methods. Alzheimer's Res 1996; 2: 1-15.