J Turgut Ozal Med Cent
DOI: 10.5455/jtomc.2017.07.098 2017;24(3):319-23

Original Article

Development of web-based software for acute coronary syndrome and a medical data mining application

Emek Guldogan[1], Julide Yagmur[2]

[1]Inonu University Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Turkey
[2]Inonu University Faculty of Medicine, Department of Cardiology, Malatya Turkey

Abstract

Aim: Medical data mining is based on data mining methods and related intelligent methods (e.g., granular computing, neural networks and soft computing) used in medicine. In this research, it was aimed to develop a web-based software and to implement medical data mining on the records of the patients with acute coronary syndrome.

Materials and Methods: The data in this study included retrospective observations recorded in the database from the web-based software developed for Cardiology Department, Turgut Özal Medical Center, Inonu University. PHP (Personal Home Page) programming language and MySQL Database Management System were employed for the development of the web-based software system. Laplace Support Vector Machines (LSVM) was constructed to predict absence or presence of diabetes mellitus in patients with acute coronary syndrome.

Results: A web based software performing data entry, query, delete, update, etc. was developed. As a result of medical data mining application, the accuracy and area under ROC curve with 95% CI were obtained as; 0.9804 (0.9716 - 0.987) and 0.9332 (0.9096 - 0.9567), respectively.

Conclusion: The developed web-based software created a very important infrastructure for implementing medical data mining applications. It was determined that the LSVM model produced very good predictive results to estimate absence or presence of diabetes mellitus in patients with acute coronary syndrome.

Keywords: Diabetes Mellitus; Laplace Support Vector Machine; Medical Data Mining.

INTRODUCTION

Acute coronary syndrome (ACS) is a syndrome that occurs due to reduced blood flow to the coronary arteries, such as the inability of the heart muscle to function properly or to lose its function (1,2). Diabetes mellitus (DM) is a disease that can develop with the elevation of blood glucose levels and can often result in the co-occurrence of genetic and environmental factors. According to the World Health Organization, it is estimated that DM will be the 7th leading cause of death in 2030 (3).

Data mining can be defined as a process that reveals relationships and patterns in massive data sets using various statistical and machine learning methods and is operated to make consistent predictions from these (4,5). Data mining is an interdisciplinary discipline that covers statistics, artificial intelligence, management information systems, pattern recognition, mathematical modeling, and database activities (6). Data mining methods are generally divided into two categories: supervised learning and unsupervised learning. In supervised learning, output values are predicted according to input data. In supervised learning, predicted output values are predetermined (7).

Classification of individuals as patients and controls based on risk factors for a disease and demographic data is an example of supervised learning. In unsupervised learning, it is only targeted to group input data based on input values in cases where there are no output values and they are not determined (7).

In recent years, the discovery of medical information, the artificial intelligence and the medical data mining have attracted considerable interest in the health sciences, and there are many publications on these issues. In this study, it was aimed to develop a web-based software for storing the records of patients with acute coronary syndrome in the database, to perform the medical data mining application in these data, to examine the factors that may be related to DM in individuals with acute coronary syndrome, and to rank them according to their importance levels.

Received: 10.07.2017 **Accepted:** 10.07.2017
Corresponding Author: Emek Guldogan, Inonu University Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Turkey. E-mail: emek.guldogan@inonu.edu.tr

MATERIALS and METHOD

Data

The present study was approved by the Ethical Committee of the Malatya Clinical Investigation Board with the protocol numbered 2016/161. In this study, a web-based software, which is written with the PHP programming language and can perform processes such as data entry, query, delete, and update, has been developed for the Department of Cardiology. In this context, descriptive information for the data in the following table are given in Table 1.

Sample Size

The estimated glucose level difference between the two groups was 20 and the assumed common standard deviation was 55. Type I error (alpha) was 0.05 and type II error (beta) was 0.10. When a power analysis was performed by taking these into account, at least 320 individuals in total (at least 160 individuals in each group) should be required (8). Moreover, when the number of independent/predictor variables/properties in a multivariate statistical model is 6 or greater, the equality of n > 104 + k (k: independent/estimator variable/property number) can be used in determining sample size (9,10). The data of 1378 individuals were included in this study.

Medical Data Mining

The extreme/outlier values were detected using the local outlier factor (LOF) method (11), and then the detected extreme/outlier values were removed from the data set.

Table 1. Descriptive information for the variables evaluated in this study

	Variables	Variable Type	Variable Description	Variable Role
Demographic	Diabetes Mellitus (DM)	Categorical	Yes/No	Dependent/Target
	Age	Numerical	Natural number	Independent/Predictive
	Gender	Categorical	Woman/Man	Independent/Predictive
	Body Mass Index (BMI)	Numerical	Positive real number	Independent/Predictive
Disease History - Disease story	Hypertension (HT)	Categorical	Present/Absent	Independent/Predictive
	Smoking status	Categorical	Present/Absent	Independent/Predictive
	Renal insufficiency history	Categorical	Present/Absent	Independent/Predictive
	Myocardial Infarction (MI) history	Categorical	Present/Absent	Independent/Predictive
	ACS family history	Categorical	Present/Absent	Independent/Predictive
	Malignancy history	Categorical	Present/Absent	Independent/Predictive
	Hyperlipidemia history	Categorical	Present/Absent	Independent/Predictive
	Peripheral artery history (PAH)	Categorical	Present/Absent	Independent/Predictive
	Coronary Artery By-pass Graft (CABG) history	Categorical	Present/Absent	Independent/Predictive
	Stroke history	Categorical	Present/Absent	Independent/Predictive
	Heart failure history	Categorical	Present/Absent	Independent/Predictive
Medicine History	Acetylsalicylic acid (ASA)	Categorical	Present/Absent	Independent/Predictive
	Clopidogrel	Categorical	Present/Absent	Independent/Predictive
	Beta Blocer	Categorical	Present/Absent	Independent/Predictive
	Calcium Channel Blocker	Categorical	Present/Absent	Independent/Predictive
	Statin	Categorical	Present/Absent	Independent/Predictive
	Digoxin	Categorical	Present/Absent	Independent/Predictive
	Angiotensin Converting Enzyme (ACE) inhibitor	Categorical	Present/Absent	ndependent/Predictive
	Angiotensin Receptor Blocker (ARB) other	Categorical	Present/Absent	Independent/Predictive
Laboratory	Creatinine	Numerical	Positive real number	Independent/Predictive
	Blood Urea Nitrogen (BUN)	Numerical	Positive real number	Independent/Predictive
	Cholesterol	Numerical	Positive real number	Independent/Predictive
	Triglycerides	Numerical	Positive real number	Independent/Predictive
	Low-density lipoprotein (LDL)	Numerical	Positive real number	Independent/Predictive
	High-density lipoprotein (HDL)	Numerical	Positive real number	Independent/Predictive
	Systolic Blood Pressure (SBP)	Numerical	Positive real number	Independent/Predictive
	Diastolic Blood Pressure (DBP)	Numerical	Positive real number	Independent/Predictive
	Diuretic	Categorical	Present/Absent	Independent/Predictive
	Glucose	Numerical	Positive real number	Independent/Predictive

A standardization method was applied to the quantitative variables in the data. The SVM model was created using the Laplacian kernel function. The predictive performances of the Laplacian SVM were evaluated using the 10-fold cross validation method. The parameter ranges for C and Sigma, which are the optimization parameters of the Laplacian kernel function, are respetively (2-2-25) and (0.02-0.20), while the number of combinations was determined as 24. Here, the C (cost) parameter controls the balance between uniformity of the separating hyperplane and misclassified training data (12). Sigma is the other parameter of the Laplacian kernel function.

Web-Based Software Development

The PHP (Personal Home Page) programming language was used in this software. PHP is a scripting language that runs on the server and is embedded into HTML codes. There is no compiler requirement for PHP codes to run. The desired text editor can be preferred for writing codes. HTML codes we prepare on the web pages give fixed outputs as long as they are not compiled specially. Therefore, there are things that cannot be done with a plain HTML code. Using HTML codes, we cannot read or write a text file on a web server and cannot connect to any database management system. There is a need for scripts to be placed in HTML codes for such processes (13,14).

The MySQL Database Management System (DBMS) was used in this web-based software. It has a multi-channel and -user, high-speed, and reliable structure. The MySQL DBMS can be accessed by programming languages such as PHP, Python, and Java. APACHE server program and PHP are frequently used together in web-database applications. The MySQL DBMS provides a flexible structure with a variety of table formatting options and processing variants. The MySQL DBMS is a good choice in projects where speed and ease stand out. However, if the number of tables is too many and complex, the advanced features on the traditional DBMS servers would be demanded (14-16).

The login screen of the web-based cardiology data entry system was designed as in Figure 1. This screen is the interface in which user of the hospital, whose data are, makes authentication.

The interface in which the demographic data related to the patient are recorded is shown in Figure 2. Necessary information about the date of application to the hospital, name, surname, Turkish identity number, telephone number, gender, year of birth, socio-economic level, literacy status, occupation, height, weight, medical history and drug use were recorded using this screen.

The display in which the patient's data such as blood pressure, heart rate, and laboratory parameters are recorded is shown in Figure 3.

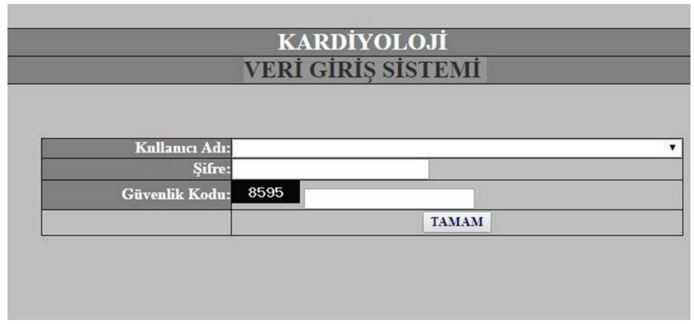

Figure 1. The login screen of the web-based cardiology data entry system

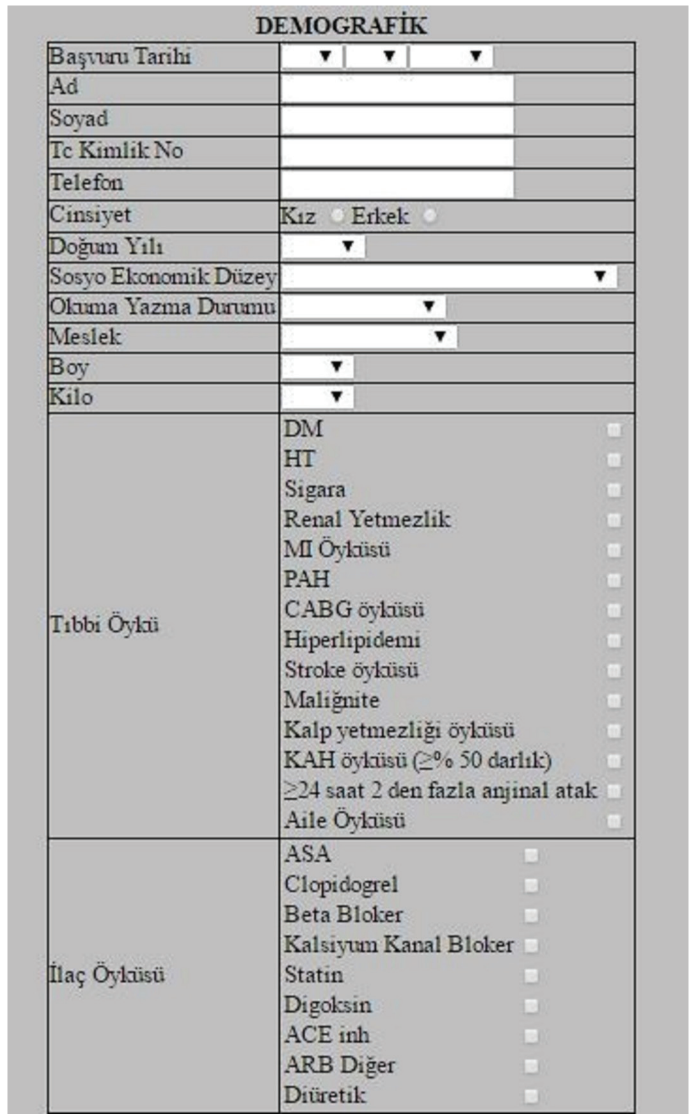

Figure 2. The interface in which the demographic data related to the patient are recorded

A relational database called cardiology was defined in the MySQL DBMS, and "patient_information_demographic", "patient_information_application", "institutions" and "users" tables were created.

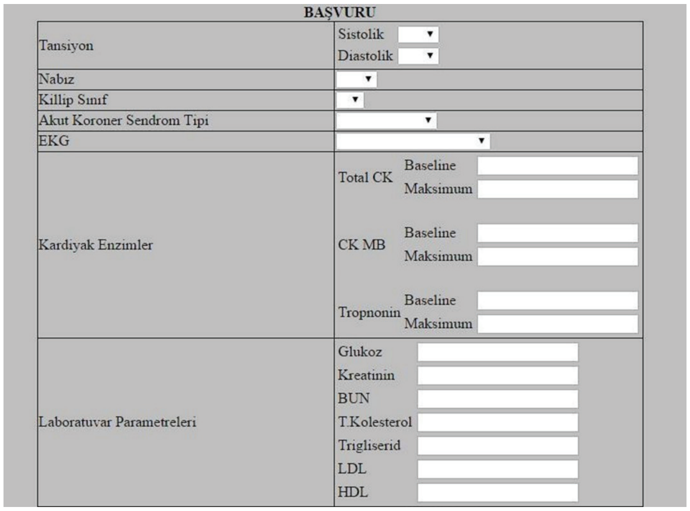

Figure 3. The display in which the patient's data such as blood pressure, heart rate, and laboratory parameters are recorded

RESULTS

As a result of extreme/outlier value analysis performed based on the local outlier factor (LOF) method, two observations were removed from the dataset generated by non-DM individuals. Accordingly, 1176 (85.3%) individuals were in the DM group and 202 (14.7%) individuals were in the non-DM group.

As a result of the medical data mining application, the performance metrics and 95% confidence interval values for predictions for the Laplacian SVM were found to be 0.9804 (0.9716 - 0.987) for accuracy and 0.9332 (0.9096 - 0.9567) for the area under the ROC curve (AUC).

When the SVM model was created using the Laplacian kernel function, the accuracy metric was used in determining the most appropriate optimization parameter. In this situation, when the cost parameter was 16, the accuracy value was obtained as 0.8737.

The significance levels of the variables used in the study for the SVM model created by the Laplacian kernel function are given in Table 2. The calculated significance levels were normalized in the range of (0-100).

Table 2. The significance levels of the variables used in the study for the SVM model created by the Laplacian kernel function

Variable Type	Variable Importance
Glucose	100.00
BUN	58.41
Creatinine	55.63
ASA	51.86
Hypertension	49.75
Diuretic	42.91
ACE Inhibitor	40.04
BMI	38.08
Triglycerides	36.29
Beta Blocer	36.04
ARB other	34.75
Statin	33.99
MI history	33.23
Cholesterol	31.01
Renal insufficiency	30.78
Hyperlipidemia history	30.60
SBP	30.53
Heart failure history	29.46
DPB	29.41
PAH	28.74
Malignancy history	27.92
Clopidogrel	27.28
Calcium Channel Blocker	27.17
CABG history	27.05
Stroke history	26.88
Family history	22.73
HDL	20.44
LDL	20.23
Gender	9.18

DISCUSSION

Acute coronary syndrome occurs in the form of an advanced clinical condition of CAD. Therefore, identification and control of the risk factors associated with acute coronary syndrome are very important for prevention of cardiovascular diseases (primary prevention) and prevention of recurrence of diagnosed diseases (secondary prevention). In this context, in this study, in a sample of type 2 DM patients with acute coronary syndrome, the prediction performance of the SVM model created by the Laplacian kernel function was evaluated for classification of DM, which is considered to affect the development of CAD, and also significance levels of DM-related factors were obtained (17). The incidence of type 2 DM is increasing worldwide. Type 2 DM is a consequence of the interaction between genetic predisposition and behavioral and environmental risk factors. The genetic basis of type 2 DM has not yet been precisely defined, but there is strong evidence that changeable risk factors such as obesity and physical inactivity are among the main causes (18). In this study, many factors that may be associated with DM in individuals with acute coronary syndrome were examined by the Laplacian SVM, and they were ranked according to their significance levels.

In addition, in this study, it was aimed to develop a web-based software for storing the records of patients with acute coronary syndrome in the database and to perform the medical data mining application in these data. In this context, a web-based software has been developed with the PHP programming language for storing the records of patients with acute coronary syndrome in the database. Another aim of this study was to examine the prediction performance of the Laplacian SVM for classification of DM in individuals with acute coronary syndrome. The accuracy value of the Laplacian SVM for classification of DM was calculated to be quite high (0.9804). Moreover, the area under the ROC curve obtained from the Laplacian SVM was measured to be extremely high (0.9804).

This result shows that the Laplacian SVM very well classified the absence or presence of DM in patients with acute coronary syndrome. In this study, when the significance levels of the variables obtained from the Laplacian SVM created for the purpose of classification of DM in individuals with acute coronary syndrome are examined, the most important four factors were glucose (100.00%), blood urea nitrogen (BUN) (58.41%), creatinine (55.63%) and ASA (51.86%), respectively. The variables, whose significance levels were determined within the scope of this study, are compatible with the risk factors for type 2 DM indicated in the literature (19-25).

CONCLUSIONS

Consequently, a web-based software has been developed for storing the records of patients with acute coronary syndrome in the database. When the results of medical data mining conducted on the data in this database were taken into account in terms of these performance metrics, it was determined that the Laplacian SVM very successfully classified the absence or presence of DM in patients with acute coronary syndrome. In future studies, the Bayesian-based approaches such as the Naive Bayes, Gaussian Naive Bayes, Bayesian Belief Networks, and Bayesian Networks is recommended to be used in predicting the absence or presence of DM in patients with acute coronary syndrome.

REFERENCES

1. Amsterdam EA, Wenger NK, Brindis RG, Casey Jr DE, Ganiats TG, Holmes Jr DR, et al. 2014 AHA/ACC guideline for the management of patients with non-ST-elevation acute coronary syndromes: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation. 2014;130(25):2354-94.
2. Grech ED, Ramsdale DR. Acute coronary syndrome: unstable angina and non-ST segment elevation myocardial infarction. BMJ. 2003;326(7401):1259-61.
3. Organization WH. Global report on diabetes. World Health Organization, France, 2016.
4. Koyuncugil AS, Özgülbaş N. Veri Madenciliği: Tıp ve Sağlık Hizmetlerinde Kullanımı ve Uygulamaları. IJIT 2009;2(2):21-32.
5. Colak C, Karaman E, Turtay MG. Application of knowledge discovery process on the prediction of stroke. Comput Methods Programs Biomed 2015;119(3):181-5
6. Cios KJ, Moore GW. Uniqueness of medical data mining. Artif intell Med 2002;26(1-2):1-24.
7. Soni J, Ansari U, Sharma D, Soni S. Predictive data mining for medical diagnosis: An overview of heart disease prediction IJCA 2011;17(8):43-8.
8. Minitab I. MINITAB statistical software. Minitab Release. 2015;16.
9. Alpar R. Uygulamalı istatistik ve geçerlik-güvenirlik: spor, sağlık ve eğitim bilimlerinden örneklerle. 4. baskı. Detay Yayıncılık, Ankara, 2016;412-3.
10. Sümbüloğlu V, Sümbüloğlu K. Klinik saha araştırmalarında örnekleme yöntemleri ve örneklem büyüklüğü. 1.baskı. Hatiboğlu Yayınları, Ankara, 2005;134-5.
11. Breunig MM, Kriegel H-P, Ng RT, Sander J, editors. LOF: identifying density-based local outliers. ACM Sigmod Record 2000;29(2):93-104.
12. Cortes C, Vapnik V. Support-vector networks. Machine learning 1995;20(3):273-97.
13. Çaycı Ö. PHP ve MySQL. 2.baskı. Seçkin Yayıncılık, Ankara, 2003;368-9.
14. Veikkolainen T, Pesonen LJ, Evans DA. PALEOMAGIA: A PHP/MYSQL database of the Precambrian paleomagnetic data. Studia Geophysica et Geodaetica. 2014;1;58(3):425-41.
15. Aslan E, Durmaz F. Rehabilitasyon Amaçlı Bilgisayar Veri Tabani Yardımıyla Bölgesel Engelli Kişi Haritasının Oluşturulması. CBÜ Soma Meslek Yüksekokulu Teknik Bilimler Dergisi 2011;1(15):64-73.
16. Gong C, Xing J, Hu Y. Data communication of Android mobile terminal and PHP and MySQL based on JSON [J]. Industrial Instrumentation Automation 2013;1:021.
17. Targher G, Bertolini L, Poli F, Rodella S, Scala L, Tessari R, et al. Nonalcoholic fatty liver disease and risk of future cardiovascular events among type 2 diabetic patients. Diabetes 2005;54(12):3541-6.
18. Çolak MC, Çolak C, Kocatürk H, Sağıroğlu Ş, Barutçu İ. Predicting coronary artery disease using different artificial neural network models. Anadolu Kardiyol Derg 2008;8(4):249-54.
19. White WB, Cannon CP, Heller SR, Nissen SE, Bergenstal RM, Bakris GL, et al. Alogliptin after Acute coronary syndrome in patients with type 2 diabetes. N Engl J Med 2013;369(14):1327-35.
20. Gress TW, Nieto FJ, Shahar E, Wofford MR, Brancati FL. Hypertension and antihypertensive therapy as risk factors for type 2 diabetes mellitus. Atherosclerosis Risk in Communities Study. N Engl J Med 2000;342(13):905-12.
21. Bonora E, Kiechl S, Willeit J, Oberhollenzer F, Egger G, Meigs JB, et al. Population-based incidence rates and risk factors for type 2 diabetes in white individuals: the Bruneck study. Diabetes 2004;53(7):1782-9.
22. Kempf K, Herder C, Erlund I, Kolb H, Martin S, Carstensen M, et al. Effects of coffee consumption on subclinical inflammation and other risk factors for type 2 diabetes: a clinical trial. Am J Clin Nutr 2010;91(4):950-7.
23. Rich-Edwards JW, Colditz GA, Stampfer MJ, Willett WC, Gillman MW, Hennekens CH, et al. Birthweight and the risk for type 2 diabetes mellitus in adult women. Ann Intern Med 1999;130(4 Pt 1):278-84.
24. Baptiste-Roberts K, Barone BB, Gary TL, Golden SH, Wilson LM, Bass EB, et al. Risk factors for type 2 diabetes among women with gestational diabetes: a systematic review. Am J Med 2009;122(3):207-14.
25. Wannamethee SG, Shaper AG, Perry IJ, Smoking as a modifiable risk factor for type 2 diabetes in middle-aged men. Diabetes Care 2001;24(9):1590-5.