Echocardiography Population Study in Russian Federation Using Machine Learning

Oleg Metsker[1], Georgy Kopanitsa[2]*, Alexey Yakovlev[1,2], Aleksandr Ilin[2], Sergey Kovalchuk[2]

[1]Almazov National Medical Research Centre, Saint Petersburg, Russia

[2]ITMO University, Saint Petersburg, Russia

Abstract

Background: In some situations, echocardiography is regulated by clinical recommendations, but such recommendations are usually based on expert opinion, and not on the studies of the effectiveness of a diagnostic method. High-quality echocardiographic diagnostics requires the use of expensive equipment and highly qualified physician specialist. However, it is of practical importance to use echocardiography as a screening method in patients without acute symptoms.

Objective: This article describes the study results of echocardiographic (ECHO) tests data on the example of cardiovascular patients. The data from more than 145,000 echocardiographic tests were analyzed. One of the objectives of the study is the possibility to identify patterns and relationships of patient characteristics for more accurate appointment procedures based on the history of the disease and the individual characteristics of the patient.

Method: The EMR from the medical information system was converted into frames. The data from echocardiography is partly structured, contains the tables and the records on natural language. On next step, the history, diagnoses, the results of examinations were extracted from EMR. Records on natural language were processed. In the next step, the data frames were analyzed to identify correlations, as well as other data mining methods were used (the sub-process in figure includes not only machine learning methods for solving the classification problem but also clustering, class specification, etc.). As a result, data frame contained 62 features, including target classes for finding significant pathologies. On the next step, gaps were removed (for example, mean, median, or digests). The next step was to train the model using machine learning methods (naive Bayesian classifier, k-nearest neighbors' algorithm, random forest method, decision trees). The last step was to analyze the features and clinical interpretation of decision trees and other data analysis results.

Results: In the course of research, the EMR of 145,966 echocardiographic cases of patients (about 80,000 patients) with different causes of treatment were analyzed. The most frequent reasons for echocardiography are: arterial hypertension, observation of patients with heart noise, atherosclerosis and stenosis, heart attacks, congenital and acquired defects of membranes and valves. A group of patients with fibrillation was considered as a particular case. One of the critical parameters of the classification is the characteristics of the left ventricle, which reflects the severity of changes in the cardiovascular system associated with hypertension. Further, anthropometric indicators (height, weight), sex and age allow distinguishing a group of patients with a high probability of finding a significant pathology.

Conclusions: Moreover, it was also possible to identify the classes and characteristics of patients for whom repeated diagnostic procedures are reasoned. Calculation of personal risks from empirical retrospective data helps to identify the disease in the early stages. To identify patients with high risk of disease complications allow physicians to make right decisions about timely treatment, which can significantly improve the quality of treatment, and help to avoid diseases complications, optimize costs and improve the quality of medical care.

Keywords

ECHO, Echocardiography, Machine learning, Russian cardiac population, Data mining

Correspondence to:

Dr. Georgy Kopanitsa
ITMO University, Birzhevaya-4
192001 Saint-Petersburg, Russia
Phone No.+79528088099
E-mail: georgy.kopanitsa@gmail.com

Citation: Metsker O, et al. (2020). Echocardiography Population Study in Russian Federation Using Machine Learning. EJBI. 6(2): 43-48

DOI: 10.24105/ejbi.2020.16.2.8

Received: July 08, 2020
Accepted: July 22, 2020
Published: July 29, 2020

1. Introduction

Echocardiography is one of the most informative non-invasive approaches to the diagnostics of cardiovascular diseases [1]. It allows estimating the size of a heart, its structure, measuring the flow rate using the Doppler method [2] and to calculating the levels and pressure gradients at different parts of heart [3]. This method permits obtaining critical diagnostic information in emergency patients [4]. In some situations, echocardiography is regulated by current clinical recommendations [5,6], but such recommendations are usually based on expert opinions, and not on the data driven studies of the real world evidences of the effectiveness of a diagnostic method [7]. High-quality echocardiographic diagnostics requires the use of expensive equipment and highly qualified specialists. However, sometimes it is important to use echocardiography as a screening method in patients without acute symptoms [8]. The expansion of indications for echocardiography may be due to the risks in case of non-detection of rare pathology or misdiagnosis in an emergency situation. This is considered estimated to be higher than the risks of unjustified costs [9].

The aggregation of significant arrays of echocardiographic data in hospital information systems (HIS) allows assessing the results and implications of the data driven models on the decisions making [10,11]. With the help of machine learning data driven methods, it is possible to reveal many cardiovascular diseases and complications [12,13]. It became possible to identify the groups of patients receiving the benefit from the primary and repeated examinations, as well as the cohort of patients and clinical situations where echocardiography is uninformative and not reasonable from an economic point of view [14].

Thus, the task of the study is to develop the on machine learning based methods to identify patterns and relationships of patient characteristics for an accurate procedure, considering the history of the disease and the individual characteristics of a patient.

4. Methods

The electronic medical records (EMR) of echocardiography tests (145,966 primary and secondary), accumulated from 2010 to 2018 in the National Medical Research Centre Almazov were analyzed in the study. The dataset was randomly split into train and test datasets with an 80-20 proportion.

The EMR was converted into frames. The data from echocardiography was partly structured. It contained the tables and the records in the natural language and structured data. Structured data were pre-processed and analyzed for acceptable values (Figure 1). The lines containing outliers and missing more than 30% of the values were removed from the data set. On next step we extracted the history, diagnoses, and the results of examinations. Records in the natural language were processed using the original method developed by the authors of the paper earlier for the features factorization [15].

In the next step, the data frames were analyzed to identify correlations (the sub-process in Figure 1 includes not only machine learning methods for solving the classification problem [16].

The resulting data frame contained 62 features.

4.1 Vital signs includes: Age, Gender, Height, Weight, Body surface area, BMI, Obesity, Myocardial mass index, Smoker, Systolic Blood pressure, Diastolic Blood pressure, End-diastolic volume, Simpson ejection fraction, End-systolic volume, Mass of the right ventricle

4.2 Laboratory tests: Alanine aminotransferase (ALT), Aspartate aminotransferase (AST), Hemoglobin, White blood cells (WBCs), Red blood cells (RBCs), Platelets; Kreatinin; Bilirubin, Hematocrit, Mean cell volume (MCV), Mean corpuscular hemoglobin (MCH), Red cell distribution width (RDW), Platelet distribution width (PD), Mean platelet volume (MPW), Procalcitonin (PCT), Lymphocytes (LYM), Neutrophils (NEUT), Cholesterol

4.3 Comorbidities: Aortic-moral-tricuspidal malformation, Hypertonic disease, Coronary heart disease, Congenital heart defect, cardiosclerosis, Mitral regurgitation (MR), Varicose veins, Stroke, Euthyroidism, Pyelonephritis, Cerebrovascular disease, Cholelithiasis, Peptic ulcer disease, Cholecystitis, Anemia, Diabetes mellitus , Hypothyroidism, Cardiomyopathy, Thyroiditis, Cholecystectomy, Rheumatoid arthritis, Pancreatitis, Hyperthyroidism, Nephropathy, Retinopathy, Nonthematic aortic (valve) stenosis, Congenital stenosis of aortic valve, Rheumatic aortic stenosis, Ventricular fibrillation and flutter

On the next step the lines with data gaps were removed using mean, median, and digests. On the next step, patients younger 16 years were filtered out as they did not meet the inclusion criteria. Further, the target class characteristics of the patient (Simpson's Ejection fraction 40% or less, impaired contractility of the left

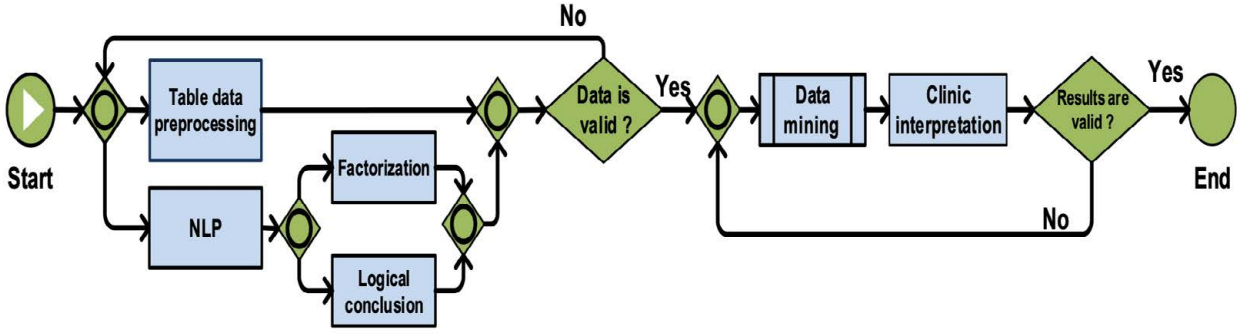

Figure 1: Method of analysis of medical electronic records of patients with echocardiography.

Table 1: Prevalent reasons for echocardiography.

ICD-10	Count of patients
Z03.8 Encounter for observation for other suspected diseases and conditions ruled out	7705
I20.8 Other forms of angina	6971
I11.9 Hypertension	6692
I48 Atrial fibrillation	3935
I10 Primary hypertension	3688
I25.2 Past myocardial infarction	3033
I25.1 Atherosclerosis of heart	2292
I20.0 Unstable angina	2285
R01.0 Cardiac noise	1976
I34.1 Mitral valve prolapse	1388
I50.0 Heart failure	1339
I35.0 Aortic (valve) stenosis	1305
O99.4 Heart disease during pregnancy	1274
Q21.1 Defect of atrial septum	1228
I67.2 Cerebral atherosclerosis	1188

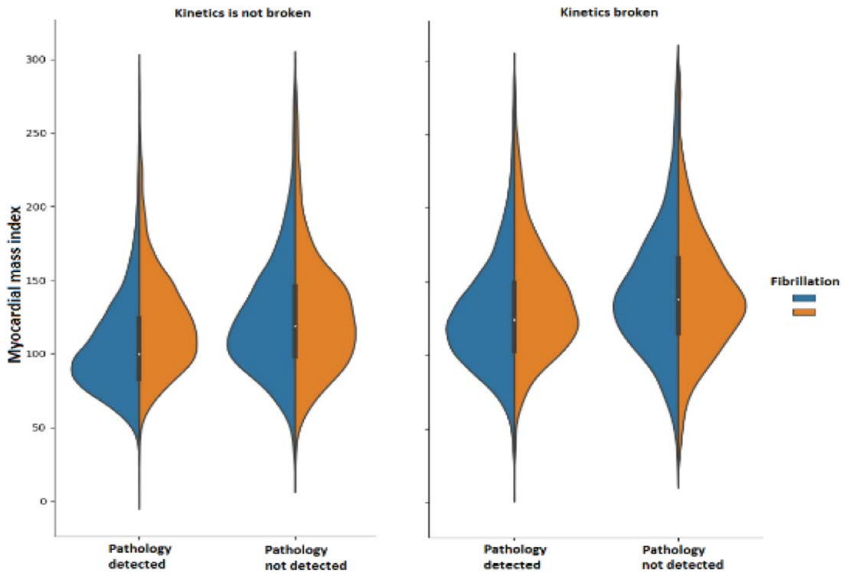

Figure 2: The analysis of echocardiographic data.

ventricle, mitral regurgitation 3-4, 2-3 aortic regurgitation, severe aortic stenosis, severe tricuspid regurgitation, pulmonary hypertension 2-3, aneurysm of aorta) were merged into a single target class „detection of significant pathology". On next step, the first episodes were marked with the target class. Further, repeated episodes were excluded from the data frame to analyze only the first episodes in which pathology was subsequently detected. The feature importance analysis was performed using a Pearson's correlation to reduce the number of parameters for implementing the model. The next step was to train the model using machine learning methods (naive Bayesian classifier [17], k-nearest neighbor's algorithm [18], random forest method [19], and decision trees [20]). The last step was to analyze the features and clinical interpretation of decision trees.

5. Results

The EMR of 145,966 echocardiographic studies of 82,421 patients with different diagnoses (Table 1) were analyzed.

The most frequent reasons for echocardiography were: arterial hypertension, observation of patients with abnormal heart sound, atherosclerosis and stenosis, heart attacks, congenital and acquired defects of membranes and valves (Figure 2).

Figure 2 demonstrates that there are no patients with a significant disease in younger age. It is suggested that younger patients with fibrillation will necessarily be diagnosed with any structural pathology, or there are just very few of them.

Pearson's correlation analysis revealed the most significant predictors for the target class (Figure 3). Top 20 predictors were used for the implementation of the model.

The model based on the decision trees provided 85% AUC of ROC on the test dataset. Dissection of the aneurysm of the ascending aorta was most dangerous condition because it was covered with pericardium, the rupture causes pericardial tamponade with no bleeding.

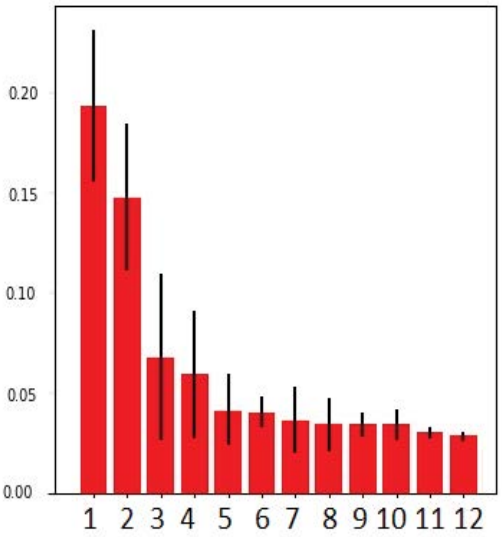

Figure 3: Features importance of aortic aneurysm model, where 1: congenital heart defect, 2: age, 3: male gender, 4: height, 5: Body-mass index (BMI), 6: body surface area, 7: weight, 8: Simpson ejection fraction, 9: cardiosclerosis , 10: end-diastolic volume, 11: end-systolic volume, 12: the mass of the right ventricle

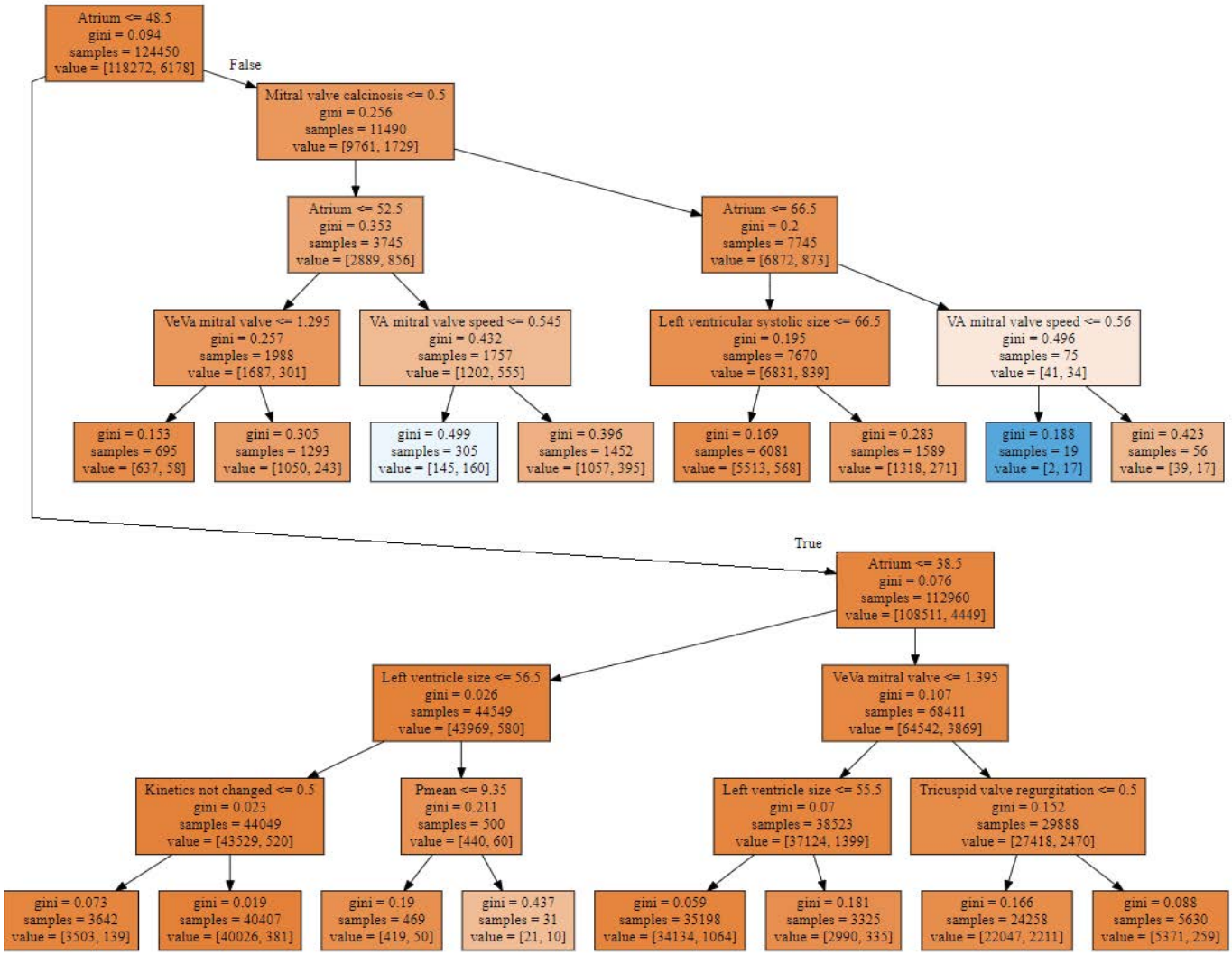

Figure 4: Decision tree classification to finding of significant pathology.

Early detection of risks and „accurate" appointment of diagnostic procedures can make a significant contribution to the value-oriented medicine. Moreover, it was also possible to identify the class and characteristics of patients for whom repeated diagnostic procedures were justified. Calculation of personal risks by empirical retrospective data contributes to the detection of the disease in the early stages, the decision on timely treatment, which in turn can significantly improve the quality of treatment, optimize costs and improve the quality of life of the patient. The decision trees provide 85% ROC cover, but in addition provides a clear clinical interpretation (Figure 4).

One of the critical parameters of the classification is the characteristics of the left ventricle, which reflects the severity of changes in the cardiovascular system associated with hypertension. Further, anthropometric indicators (height, weight), sex and age allow distinguishing a group of patients with a high probability of finding a significant pathology. In clinical practice, the decision tree can be used to determine the boundaries of reference indicators for risk calculation. Clinicians consider these models as the most interpreted to identify patterns and components of pathophysiological processes. Further detailed study may be useful in the clinical workflow to prove clinical decision-making in different situations as well as for the objective appointment of repeated echocardiography procedures.

6. Discussion

Echocardiography is a subjective method of diagnostic and is highly dependent on the skills of an operator. So, a human factor has a high influence on the results of a test. Our study provides real world evidence based on a large dataset that the machine learning methods can become an efficient tool to support doctors in the interpretation of echocardiography. At the same time this method of analysis has significant limitations, given the retrospective analysis and the lack of direct contact with the patient. It focuses only on the data that are included in the electronic health record. However, the results of the study are of a high value primarily because of the large samples of patients, and the possibility of multi-factor analysis. In can lead to the detection of risks and the factors of disease progression. The results of the study allow to develop a personalized approach to patient treatment based on the patient similarity approach.

7. Conclusions

Our study of echocardiography data allowed answering two essential questions: who is exposed to echocardiographic studies and what the probabilities of detecting pathologies are? It was possible to identify personal characteristics of an average patient in the population. By deviations from the average patient, it is possible to understand the causes and calculate the individual risk for each individual patient. Moreover, the model of appointment of the procedure considering the probability of finding a significant pathology makes a substantial contribution to the modern healthcare. Further research will be focused on the identification of the risks of mortality using echocardiography data.

8. Competing interests

The authors declare no competing interests

9. Authors' contribution

Sergey Kovalchuk was responsible for the conceptualization of the study, Oleg Metsker and Aleksandr Ilin were responsible for running the experiments and data analysis, Alexey Yakovlev was responsible for the clinical interpretation, Georgy Kopanitsa was responsible for writing and editing the manuscript.

10. Acknowledgments

This work financially supported by Ministry of Education and Science of the Russian Federation, Agreement #14.575.21.0161 (26/09/2017). Unique Identification RFMEFI57517X0161.

References

1. Pellikka PA, Nagueh SF, Olson AA, Chaudhry FA, Chen MH, Marshall JE, et al. American Society of Echocardiography Recommendations for Performance, Interpretation and Application of Stress Echocardiography. J Am Soc Echocardiogr. 2017; 1021-1040.
2. Gulati V, Katz W, Follansbee WP, Gorcsan J. Mitral annular descent velocity by tissue Doppler echocardiography as an index of global left ventricular function. 2008; 77(11): 979-984.
3. Sahn DJ, DeMaria A, Kisslo J, Weyman A. Recommendations regarding quantitation in M-mode echocardiography: results of a survey of echocardiographic measurements. Circulation. 1978; 58(6): 1072-1083.
4. Plummer D, Brunette D, Asinger R, Ruiz E. Emergency department echocardiography improves outcome in penetrating cardiac injury. Annals Emergency Med. 1992; 21(6): 709-712.
5. Via G, Hussain A, Wells M, Reardon R, ElBarbary M, Noble VE, et al. International evidence-based recommendations for focused cardiac ultrasound. J Am Soc Echocardiogr. 2014; 27(7):683.e1-683.e33.
6. Evangelista A, Flachskampf FA, Erbel R, Antonini-Canterin F, Vlachopoulos C, Rocchi G, et al. Echocardiography in aortic diseases: EAE recommendations for clinical practice. Eur J Echocardiogr. 2010; 11(8): 645-658.
7. Knottnerus JA, Tugwell P. Real world research. J Clin Epidemiol. 2010; 63(10):1051-1052.
8. Bossone E, D'Andrea A, D'Alto M, Citro R, Argiento P, Ferrara F, et al. Echocardiography in pulmonary arterial hypertension: From diagnosis to prognosis. J Am Soc Echocardiogr. 2013; 26(1):1-14.
9. Metsker O, Yakovlev A, Bolgova E, Vasin A, Koval-chuk S. Identification of Pathophysiological Subclinical Variances During Complex Treatment Process of Cardiovascular

Patients. Procedia Computer Science. 2018; 138: 161-168.

10. Krikunov AV, Bolgova EV, Krotov EM, Abuhay T, Yakovlev AN, Kovalchuk SV. Complex data-driven predictive modeling in personalized clinical decision support for Acute Coronary Syndrome episodes, Procedia Computer Science. 2016; 80: 518-529.
11. Weston AD, Hood L. Systems Biology, Proteomics, and the Future of Health Care: Toward Predictive, Preventative, and Personalized Medicine. J Proteome Res. 2004; 3(2): 179-196.
12. Narula S, Shameer K, Omar A, Dudley JT, Sengupta PP. Machine-learning algorithms to automate morphological and functional assessments in 2D echocardiography. J American Col Cardiol. 2016; 68(21): 2287-2295.
13. Lempitsky V, Verhoek M, Noble JA, Blake A. Random Forest Classification for Automatic Delineation of Myocardium in Real-Time 3D Echocardiography. 2009; 5528: 447-456.
14. Perera P, Lobo V, Williams SR, Gharahbaghian L. Cardiac echocardiography. Crit Care Clin. 2014; 30(1): 47-92.
15. Metsker O, Bolgova E, Yakovlev A, Funkner A, Kovalchuk S. Pattern-based Mining in Electronic Health Records for Complex Clinical Process Analysis. Procedia Computer Science. 2017, 119: 197-206.
16. Han J, Kamber M, Pei J. Data mining: concepts and techniques. 2012; 3: 1-740.
17. Sahami M. Learning Limited Dependence Bayesian Classifiers. KDD-96 Proceedings. 1996; 335-338.
18. Seidl T, Kriegel HP. Optimal multi-step k-nearest neighbor search. ACM. 1998; 1-12.
19. Pal M. Random forest classifier for remote sensing classification. Int J Remote Sens. 2005; 26(1): 217-222.
20. Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. IEEE. 1991; 21(3): 660-674.