Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 62 (2015) 39 – 46

Procedia Computer Science

The 2015 International Conference on Soft Computing and Software Engineering (SCSE 2015)

Analysis of Demographic Characteristics Creating Coronary Artery Disease Susceptibility using Random Forests Classifier

Kemal AKYOL[a], Elif ÇALIK[b], Şafak BAYIR[a], Baha ŞEN[c]*, Abdullah ÇAVUŞOĞLU[d]

[a] *Karabük University, Engineering Faculty, Baliklarkayasi, Karabük 78050, Turkey*
[b] *Karabük University, School of Health Sciences, Baliklarkayasi, Karabük 78050,Turkey*
[c] *Yıldırım Beyazıt University, Faculty of Engineering and Natural Sciences,Ulus, Ankara 06030,Turkey*
[d] *The Scientific and Technological Research Council of Turkey, Kavaklıdere, Ankara 06100, Turkey*

Abstract

Cardiovascular system diseases are an important health problem. These diseases are very common also responsible for many deaths. With this study, it is aimed to analyze factors that cause Coronary Artery Disease using Random Forests Classifier. According to the analysis, we observed correct classification ratio and performance measure that creates susceptibility to Coronary Artery Disease for each factor. The performance measure results clearly show the impact of demographic characteristics on CAD. Additionally, this study shows that random forests algorithm can be used to the processing and classification of medical data such as CAD.

Peer-review under responsibility of organizing committee of The 2015 International Conference on Soft Computing and Software Engineering (SCSE 2015)

Keywords: Classification, Coronary Artery Disease, Random Forests ;

1. Introduction

Cardiovascular system diseases are an important public health problem in all societies. It is responsible for the death of many people and very common. Coronary Artery Disease is leading causes of death in Turkey as developed

* Corresponding author. Tel.: +90-312-324-1502; fax: +90-312-324-1501.
E-mail address: bsen@ybu.edu.tr

Peer-review under responsibility of organizing committee of The 2015 International Conference on Soft Computing and Software Engineering (SCSE 2015)
doi:10.1016/j.procs.2015.08.407

western countries. It is known to be many risk factors that known and unknown thought to cause this disease and numerous investigations in this area. With this study, it is aimed to analyze the factors that cause CAD using Random Forests Classifier, bordered by correct classification ratios and performance measure results. Used abbreviations are given in Table 1.

Table 1. Abbreviations and explanations.

Abbreviation	Description
CAD	Coronary Artery Disease
RFC	Random Forests Classifier
HDL	HDL cholesterol value
CHOL	Cholesterol value
LDL	LDL cholesterol value
TRIG	Triglyceride value
HT	Diagnosis of hypertension
HL	Diagnosis of hyperlipidemia
DM	Diagnosis of diabetes mellitus
FH	Diagnosis of family history
SM	Diagnosis of smoking
HT_N	Hypertensive patients and normal patients
HL_N	Hyperlipidemic patients and normal patients
DM_N	Diabetes Mellitus patients and normal patients
FH_N	Patients with a family history and normal patients
SM_N	Smoking patients and normal patients

In this study, effect of demographic characteristics such as hypertension and hyperlipidemia is investigated with important test variables for CAD rather than seeking answers to “which test values are important?”. Of course, there are many factors that directly or indirectly affecting the CAD as can be seen in literature. Hence, it is possible to be CAD even if his or her all values are normal. It can be explained as the effect of hidden factors.

The structure of the paper is as follows. Section 2 discusses the literature review related to CAD. Section 3, the information about classification is given. Section 4 describes the material and methods used to develop the model. It is elaborated the application, developed for the analysis of demographic characteristics that cause to CAD. The impact of risk factors that cause to CAD and the obtained results are presented in Section 5. In final section, conclusions and recommendations are presented.

2. Literature Review

There have been many studies in this field. Akyol et al [2] have tested that significance of biochemistry and laboratory data for CAD with logistic regression method. Turkish Society of Cardiology data is referenced for this test. CAD analysis is realized with inclusion of important variables in to the model. In Framingham study, a person's risk of CAD in ten years was analyzed in framework of criteria obtained with knowledge discovery [6]. In PROCAM study, a system has been developed for the prevention of CAD. The studies were conducted to be able to prevent disease and get early diagnosis with this system [3]. In TEKHARF study, the search of scoring CAD risk has been done especially on adults of Turkish individuals. This study was conducted with 10 years of knowledge of individuals [10]. Wang et al [14] have had a significant result that, traditional risk factors have different size of effect on CAD with using Framingham function and also they have found another result that other risk factors can also cause this disease. Karaolis et al [6] have developed a data mining system based on Apriori algorithm. This system is used to identify the risk factors associated with CAD. According to this study, smoking is an important risk factor which that directly affecting the CAD for all cases. Kunc et al [8] developed simulation which can be used for evaluation of patients such as coronary heart disease, congestive heart failure, end-stage renal disease. Additionally, treatment costs were calculated for each of disease. The results of this study showed that it is possible to estimate the potential savings for the treatment of chronic diseases. Srinivas et al [12] have tried to effectively predict heart attack using data mining techniques such as decision tree, artificial neural networks and Bayesian classifiers. Abdullah and Rajalaxmi [1] have developed data mining techniques to help doctors in prediction of CAD using random forests algorithm. Another study called logistics risk score, the risk of CAD is predicted. This study

has an important place in the treatment of patients with acute coronary syndrome and CAD within framework of identification of the risk factors for this disease [5].

3. Classification

Classification is the process of finding which data belongs to the variety of classes through training and observation. The created model is trained with the training data and is asked to decide which test data are belong to which class. The best sampling of data space is created the most performance is achieved by training data.

4. Materials and Methods

4.1. Data Description

Actualized as a retrospective and case-control, this study's data was obtained from Ankara Ataturk Training and Research Hospital Cardiology and other services of Yildirim Beyazit University on the dates between 01.01.2011 and 11.10.2011. It includes information of laboratory and demographic characteristics of patients.

ICD-10(International Statistical Classification of Diseases and Related Health Problems) booklet is a reference for the patient and control groups. Feature definition is an important step of the classification process and in fact it constitutes the core of problem solution. In this context, in order to analyze laboratory tests variables, it is based on the manual of Turkish Society of Cardiology published in 2002; "Coronary Heart Disease Prevention and Treatment Manual". The risk value is determined in this manual and used in our study are as follows:

- Age limit should be 45 for men and 55 for women and above,
- LDL serum cholesterol levels should be equal 130 mg/dl and above and/or Cholesterol levels should be equal 200 mg/dl and above,
- Serum triglyceride level should be equal 150 mg/dl and above,
- Serum HDL cholesterol level should be less than 40 mg/dl.

4.2. Data Preprocessing

Data preparation is essential for successful classification. Typically, poor quality data are incorrect and unreliable data mining results. Data preparation process consists of the following actions: determination of variables in accurately, cleaning their respective values, built into the form which is desired and formatted. After the data cleaning is completed, the categorical variables are transformed into proper representational form for classification. For instance, for the gender type variable which depicts male or female, "0" is used for male and "1" used for female. In addition, the reference values of test variables are introduced to the system. According to Turkish Society of Cardiology data, reference value can be below or above specific threshold. The list of variables in dataset is shown in Table 2.

Table 2. The list of variables in dataset.

Variable name	Data type	Description
GENDER	Categorical	Patient gender
AGE	Categorical	Patient age
DIAGNOSIS_OF_DISEASE	Categorical	The situation of CAD
LDL	Continuous	LDL cholesterol value
HDL	Continuous	HDL cholesterol value
TRIGLYCERIDE	Continuous	Triglyceride value
CHOLESTEROL	Continuous	Cholesterol value
HYPERTENSION	Categorical	The situation of hypertension
HYPERLIPIDEMIA	Categorical	The situation of hyperlipidemia
DIABETES_MELLITUS	Categorical	The situation of diabetes mellitus
SMOKING	Categorical	The situation of smoking
FAMILYHISTORY	Categorical	The situation of family history

Each dataset includes two categorical and four continuous variables. The resume information of patients, which causes to increase the susceptibility of CAD are as follows: diabetes mellitus, hypertension, hyperlipidemia, family history and smoking. The models are generated with datasets which are created with each risk factor that creating CAD susceptibility. 40% of each data set was reserved for training and the rest was reserved for the model testing. One half of training and test data is composed of demographic characteristics and other half is composed of normal patients' data. However, the number of patients with a diagnosis of CAD can vary in each dataset. According to the Table 3, 30 patients are diagnosed as HT and 30 of them are diagnosed non-HT. On the other hand according to Table 4, it is observed that 41 of the 60 patients have CAD and 19 have non-CAD.

Table 3. Created datasets.

HT (Dataset 1)		HL (Dataset 2)		DM (Dataset 3)		FH (Dataset 4)		SMO (Dataset 5)		All Factors (Dataset 6)	
0	1	0	1	0	1	0	1	0	1	0	1
30	30	10	10	16	16	7	7	10	10	73	73

Table 4. Training and test data.

		Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Dataset 6
CAD	0	19	8	10	6	8	51
	1	41	12	22	8	12	95
Train		24	8	12	5	8	58
Test		36	12	20	9	12	88
Sum		60	20	32	14	20	146

4.3. Random Forest Classifier

Random forest algorithm is a very efficient and popular algorithm for classification and regression problems introduced by Breiman [4]. RFC is a machine learning technique that constructs a forest of classification trees wherein each tree is expand on a bootstrap sample of the data and the attribute at each tree node is selected from a random subset of all attributes. The final classification of an individual is determined by voting over all trees in the forest [11]. The algorithm uses a haphazard subset of forecaster variables to split an observation data into homogenous subsets [9]. Increase the value of classification is aimed using multiple decision trees during the classification process with this algorithm.

4.4. The Developed Application

In this study, the analysis of demographic characteristics which cause the CAD is handled with the help of RFC. As shown in Figure 1, for the analysis belong to each model which is edited, the operations performed on the data set as follows:

- Laboratory and diagnostic information is loaded and data is prepared,
- Meaningful information is obtained and dataset is prepared,
- The success of the model is evaluated by testing the constructed model with training and test datasets.
- Performance measure analysis is made.

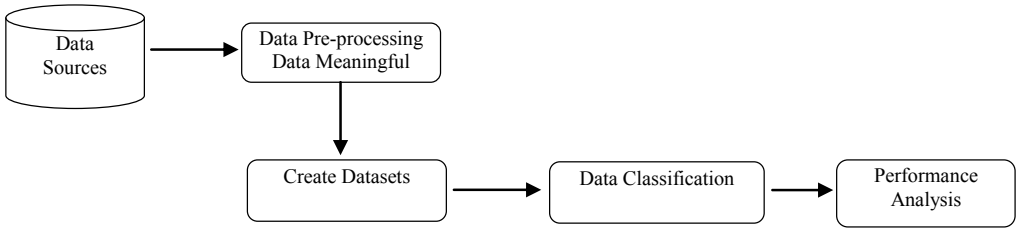

Fig. 1. Schematic representation of data analysis.

5. Results and Discussion

Obtained data and performance analysis in this study are presented in table format. Since the output variable had two categorical values, the confusion matrixes show 2x2 square matrix. In the confusion matrixes the rows represent the expected and the columns represent the observed. The last column shows the prediction accuracies for each of the two output variable values. The overall accuracy of each model is presented at the bottom of the last columns.

Before examining the impact on the CAD of the demographic characteristics, a classification was made about the laboratory test variables, referred to Turkish Society of Cardiology. Being the total number of data is 306, which was in classification, the number of patients with positive CAD diagnose is 153. Likewise, the number of patients with negative CAD diagnose is 153 as well. The number of data to be used for training was 122; the number of data to be used for testing was 184.

Table 5. Prediction results that effect of laboratory test values for CAD.

	No	Yes	Accuracy (%)
No	47	45	51.08
Yes	19	73	79.34
		Overall	65.22

As seen in Table 5, classification performance is about 65.22%. As the reason for this level of the successful classification rate, it can be interpreted that data space could not represent CAD in a good manner. On the other hand, it should be remembered that there are a lot of reasons of CAD directly or indirectly. Demographic characteristics of the patients can be one of these reasons. In this study, analysis was performed to examine the effects of these demographic characteristics on CAD. For example, when hypertension data were examined in Tables 3 and 4, we can see the total the number of patients is 60. Within these patients, 30 is diagnosed as HT, 41 is diagnosed as CAD. Classification analysis results are given in Table 6. The results show that diagnostic information which is cause to CAD. Among the six analyses, hypertension data were examined, the number of patients with a diagnosis of CAD is 22 and the correct classification ratio is 80.55%.

Table 6. Prediction results for all factors that creating CAD susceptibility.

Datasets		No	Yes	Accuracy (%)
CAD Analysis with HT_N				
	No	7	3	70.0
	Yes	4	22	84.61
			Overall	80.55
CAD Analysis with HL_N				
	No	4	1	80.0
	Yes	1	6	85.71
			Overall	83.33
CAD Analysis with DM_N				
	No	5	0	100.0
	Yes	5	10	66.66
			Overall	75.0
CAD Analysis with SM_N				
	No	4	1	80.0
	Yes	1	6	85.71
			Overall	83.33
CAD Analysis with FH_N				
	No	3	0	100.0
	Yes	2	4	66.66
			Overall	77.77
CAD analysis with all of diagnoses and normal patients.				
	No	32	0	100.0
	Yes	2	54	96.42
			Overall	97.72

In Figure 2, it's seen the importance degrees of each test variable in created models. At the top right of each graph, we can see the significance of the test variables.

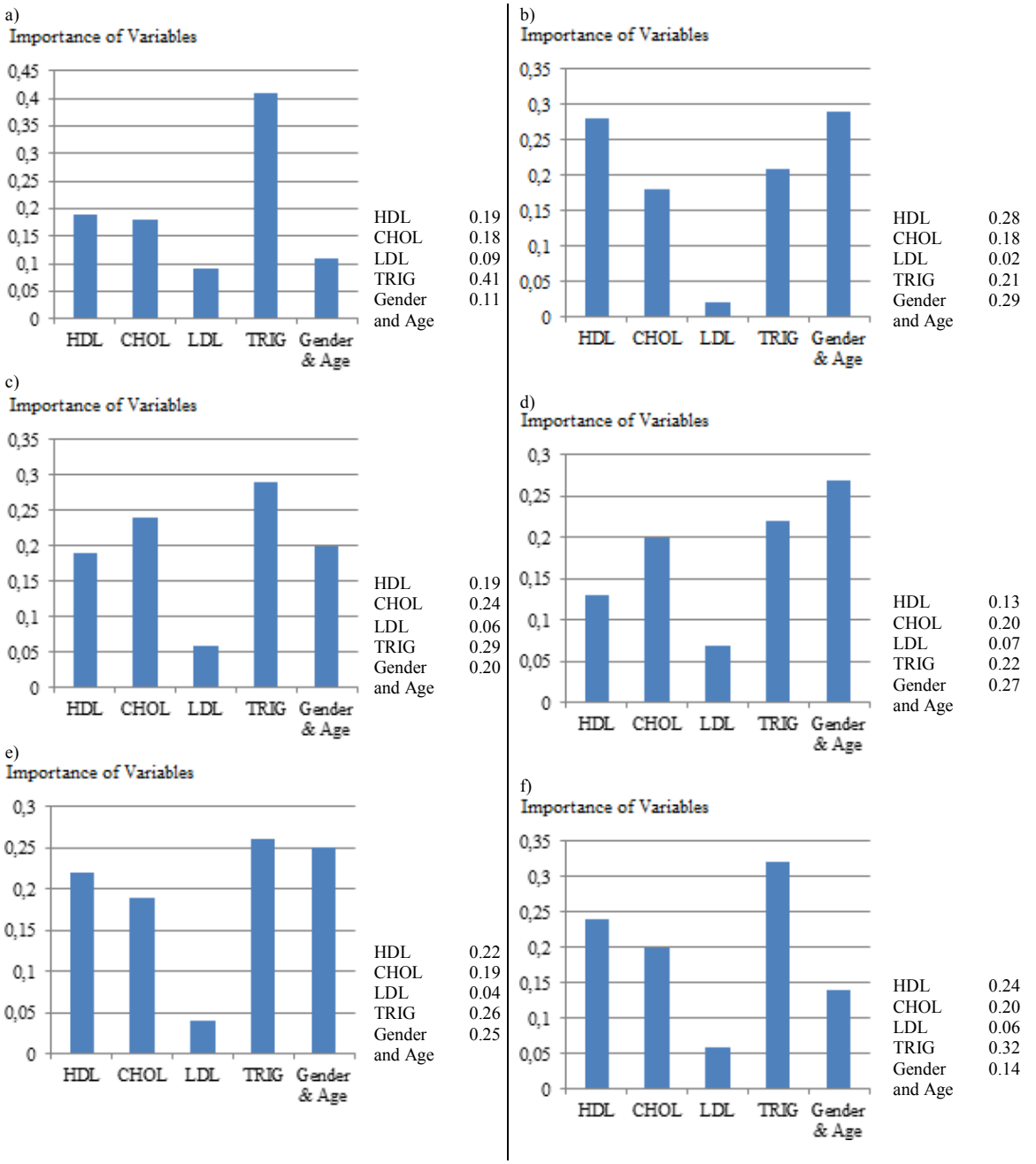

Fig. 2. A graphical representation of the importance of the variables in all analysis: (a) CAD analysis with hypertensive patients and normal patients; (b) CAD analysis with hyperlipidemia patients and normal patients; (c) CAD analysis with diabetes mellitus patients and normal patients; (d) CAD analysis with patients with a family history and normal patients; (e) CAD analysis with smoking patients and normal patients; (f) CAD analysis with all of diagnoses and normal patients.

In order to evaluate algorithm performance, we used sensitivity (SN), specificity (SP), true positive rate (TPR) and accuracy (AC). These metrics are computed using Equation 1 respectively [13].

$$SN = TP/TP+FN, \qquad SP = TN/TN+FP, \qquad TPR = TP/TP+FP \tag{1}$$

All these measures can be calculated based on four values: true positive (TP), the number of patients with CAD correctly classified as patients with CAD; false positive (FP), the number of non-CAD patients wrongly classified as diagnosis of CAD; false negative (FN), the number of patients with CAD wrongly classified as non-CAD patients and true negative (TN), the number of non-CAD patients correctly classified as non-CAD patients.

The successful results are obtained from analysis with the diagnosis of hypertension, hyperlipidemia, diabetes mellitus, smoking and family history which thought to be caused to CAD. These data are given in Table 7. It is considered these diagnoses to be important for CAD. Thereupon, it is analyzed with a dataset which is established that contains all of the diagnoses which is analyzed before. Correct classification ratio of generated model is 97.72%.

Table 7. Classification evaluation results.

Datasets	SN	SP	TPR	AC
HT_N	63.63	88.0	70.0	80.55
HL_N	80.0	85.71	80.0	83.33
DM_N	50.0	100.0	100.0	75.0
FH_N	60.0	100.0	100.0	77.77
Smoking_N	80.0	85.71	80.0	83.33
All Factors	94.11	100.0	100.0	97.72

6. Conclusions and Recommendations

A way to obtain meaningful information from data is classification. Classification is the process of finding where the data belongs to the several of classes by training and observation. RFC, which helps to make classification and assignment process, is a very popular machine learning technique.

While CAD occurs in 5-10% of healthy individuals it can reach 80% in patients whose laboratory test values exceed the reference value. Moreover, the risk of CAD increases regarding the age.

In this study, the factors which are creating CAD susceptibility are analyzed with RFC. CAD data are composed of processing received data, which was diagnosed by specialists in cardiology department. The data of the control group, on the other hand is consisted of patients in the other departments except the cardiology department. Although the control group data is not obtained from the department of cardiology, it should be considered that patients may even have heart disease in other sections. For this reason, in the medical research, it would be more successful that the data should be processed with a cardiologist. It is understood from obtained results by analysis that diabetes mellitus, hypertension, hyperlipidemia, smoking and family history are risk factors for CAD. In contrast to cannot be changed risk factors such as age, gender and family history, the modifiable risk factors diabetes mellitus, hypertension, hyperlipidemia and smoking are taken under the control, it is inevitable that avoided CAD and it's complications. The obtained results are found to have a successful data classification with random forest and can help the medical practitioners for predicting CAD.

In addition to this, another important detail must be taken into consideration that the number of cases is quite few in the obtained data set. So, the correct classification percentage of the created model depends on enough number of cases in the data set.

We also note that is worth noting: The data in this study are actual data obtained from hospitals but that would be wrong to say that the results obtained from analysis are valid in all circumstances. Because, it is considered that many known and unknown factors are caused to this disease.

References

1. Abdullah AS, Rajalaxmi RR. A Data mining Model for predicting the Coronary Heart Disease using Random Forest Classifier. *International Conference on Recent Trends in Computational Methods, Communication and Controls (ICON3C 2012)*; 2012. p. 23.
2. Akyol K, Şen B, Çalık, E, Analysis of Biochemistry and Hemogram Laboratory Test Results with Logistic Regression Method, *Akademik Bilişim;* 2012. p. 350-354.
3. Assmann G, Cullen P, Schulte H. Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the Prospective Cardiovascular Münster(PROCAM) study. *Circulation* 2002;**105**:310-5.
4. Breiman L. Random forests. *Machine Learning* 2001; **45**. p.18.
5. Chambless LE, Dobson AJ, Patterson CC, Raines B. On the use of a logistic risk score in predicting risk of coronary heart disease. *Stat Med 9* 1990;**385**-96.
6. Kannel WB, McGee D, Gordon T. A general cardiovascular risk profile: the Framingham Study. *Am J Cardiol* 1976; **38**:1-1.
7. Karaolis M, Moutiris JA, Pattichs L. Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees. *IEEE Transactions on IT in Biomedicine* 2010;**14**:3.
8. Kunc MA, Drinovec J, Rucigaj S, Mrhar A. Simulation analysis of coronary heart disease, congestive heart failure and end-stage renal disease economic burden. Mathematics and computers in simulation, *6th Vienna International Conference on Mathematical Modelling* 2011;**82**:3 p. 494–507.
9. Mellor A, Haywood A, Stone C, Jones S. The performance of random forests in an operational setting for large area sclerophyll forest classification. *Remote Sens* 2013; **5**:6 p. 2838–2856.
10. Onat A, Keleş İ, Çetinkaya A, Başar Ö, Yıldırım B, Erer B, Ceyhan K, Eryonucu B, Sansoy V. Prevalence of Coronary Mortality and Morbidity in the Turkish Adult Risk Factor Study: 10-year Follow-up Suggests Coronary "Epidemic". *Türk Kardiyol Arş* 2001; ***29****:1-1*.
11. Reif DM, Motsinger AA, McKinney BA, James E, Crowe JR, Moore JH. Feature Selection using a Random Forests Classifier for the Integrated Analysis of Multiple Data Types. *Computational Intelligence and Bioinformatics and Computational Biology*; 2006. p.2.11
12. Srinivas K, Rao GR, Govardhan A. Analysis of Coronary Heart Disease and Prediction of Heart Attack in Coal Mining Regions Using Data Mining Techniques. *5th IntConf on Computer Science and Education*; 2010. p. 1344-1349.
13. Tripathi S, Singh KK, Singh BK, Mehrotra A. Automatic Detection of Exudates in Retinal Fundus Images using Differential Morphological Profile, *International Journal of Engineering and Technology (IJET)* 2013; **5**:3, p. 2028-2029.
14. Wang Z, Hoy WE. Is the Framingham coronary heart disease absolute risk function applicable to Aboriginal people? *Med.J.Australia* 2005;**182**:2-66.