Boosted Decision Trees for Lithiasis Type Identification

Boutalbi Rafika
Computer Science Departement, Badji Mokhtar University
LABGED Laboratory
Annab, Algeria

Chitibi Kheir Eddine, Boutefnouchet
Urology department, CHU Ibn Rochd
Badji Mokhtar University Hospital
Annaba, Algeria

Farah Nadir
Computer Ssience Departement, Badji Mokhtar University
LABGED Laboratory
Annab, Algeria

Tanougast Camel
University of lorraine
LCOMS-ASEC
Metz, France

***Abstract*—Several urologic studies showed that it was important to determine the lithiasis types, in order to limit the recurrence residive risk and the renal function deterioration. The difficult problem posed by urologists for classifying urolithiasis is due to the large number of parameters (components, age, gender, background ...) taking part in the classification, and hence the probable etiology determination. There exist 6 types of urinary lithiasis which are distinguished according to their compositions (chemical components with given proportions), their etiologies and patient profile. This work presents models based on Boosted decision trees results, and which were compared according to their error rates and the runtime. The principal objectives of this work are intended to facilitate the urinary lithiasis classification, to reduce the classification runtime and an epidemiologic interest. The experimental results showed that the method is effective and encouraging for the lithiasis type identification.**

Keywords—urinary lithiasis; classification; Boosting; Decision Trees

I. INTRODUCTION

Urinary lithiasis are hard crystals that form in the urinary tract, mostly in the upper urinary tract. From a cooperation with the hospital university center of Annaba (CHU), we obtained a significant related data set. However the major problems of these data resided in their analysis and their interpretations to well define the problem. Physics laboratory of CHU has provided us with data concerning the patients (age, sex, …) and urolithiasis composition. Most collected data are important in determining urinary lithiasis type.

The urolithiasis composition plays a significant role [1] [2] in determining the lithiasis types and their etiologies, which will allow to know the reasons of their occurrence, and help to prescribe a diet or appropriate treatment.

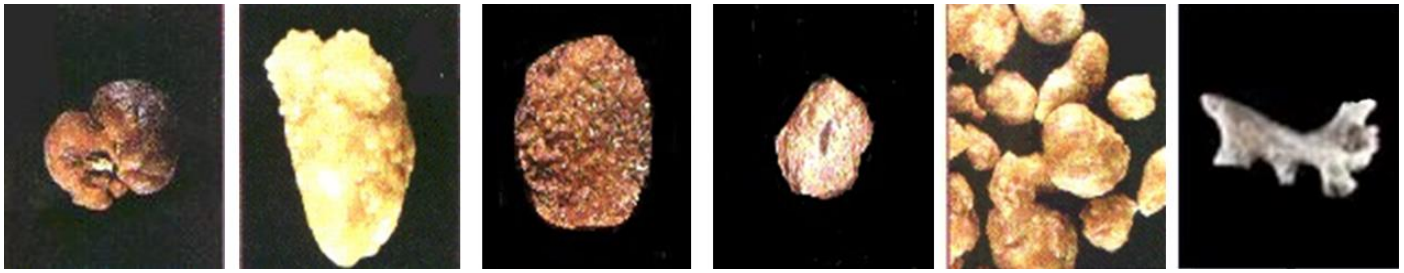
The problem posed in this study was to identify urinary lithiasis type according to their compositions and the patient's profile.

There exist 6 types of urolithiasis which differ according to their morphological and chemical compositions, the six types of urolithiasis are presented in the following figure (Fig. 1). Most studies [3] is based on the four most dominant urinary lithiasis types, namely types 1,2,3 and type 4, because 80% of the urinary lithiasis are part of these four types. In this work the six types of existing urinary lithiases , have been included in the classification, and which correspond respectively to the following etiologies: hyperoxaluria, hypercalciuria, Hyperuricosuria, urinary infections, Cystinuria, Proteinuria.

Each of the six types is composed of the following substances: C1 for type 1. C2 and C1 for the type2. C1, C2 and AU for type 3. C1, C2 and CA for type 4. Cystine and CA for type 5. C1 and Protein for type 6 [3, 4]. However, these six types are not only composed of the quoted components but contain tens other components, with relatively low amounts, which make it possible to effectively distinguish the six types , which is not the case for the components present with large amounts (appendix 1).

In this article, a boosted decision tree system was used to determine the urinary lithiasis types.

This paper is organized as follows. Section II presents the related works. Section III discusses data analysis and data reduction. In Section IV the different methods and tools used were explained. In Section V the results of these methods are presented and compared to other models of learning, according to their classification accuracy, thus etiologies determination. Finally, Section VI concludes the paper.



(a) (b) (c) (d) (e) (f)

Fig. 1. The six urolithiasis types, (a) Type I, (b) Type II, (c) Type III, (d) Type IV, (e) Type V, (f) Type VI

II. Related Works

Work on various urolithiasis types recognition, have been the subject of some studies, in particular the work performed by Igor Kuzmanovski et al., in their article "Classification of Urinary Calculi using Feed-Forward Neural Network", they carried out, using a neural network, the urolithiasis classification based on lithiasis spectrophotometric analysis. Genetic algorithms have been used to optimize the selection of the most suitable spectral areas in order to improve the classification.

We realized at the beginning of this project, a first work on the classification of urolithiasis, which gave promising results. We presented the results of three classifiers system used: neural network, SVM and a neuro-fuzzy system, and were compared according to their effectiveness.

III. Data Analysis

In this work the data of 528 patients were used, each sample (for each patient) has 23 features which are: age, sex, and twenty-one components (C1, C2, CA, AU0, AU2, WFP, Br, Cystine, URAM, Calcite, Protein, Trg, Mps, Wk, pacc, ocp, urna, Inc, oxypurinol, nexbrit, polysa) (TABLE I).

After having standardized (TABLE II) the 378 patients data, they were divided into two subsets: 378 samples for the training stage and 150 for the validation stage.

TABLE II. Data Codification

Data	Codification
Age and Quantity of components	Integer
Sex	0 for woman 1 for man

Data normalization is an important step especially for classifiers based on distance calculation between two samples like KNN. The normalization ensures that no variable takes too much importance simply because of its measurement unit and it also allows to give equal weight to all the variables. The Normalization of our features was realized using the following formula:

$$fn = \frac{f - f_{min}}{f_{max} - f_{min}}$$

Where fn is a normalized feature value, f is the original feature value, f_{min} is a minimum of feature values and f_{max} is a maximum of feature values.

Several data analysis in particular statistical analysis were performed to better evaluate and interpret data.

Of the 378 cases recorded, there is a ratio man/woman equal to 1.6 i.e. 3 men for 2 women suffering from renal lithiasis (Fig. 2).

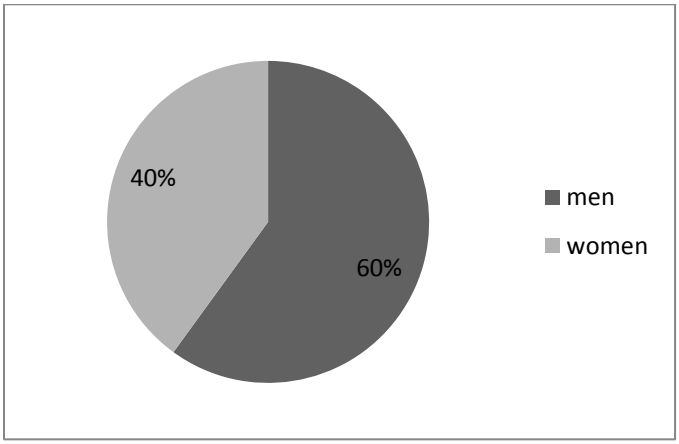


Fig. 2. Proportion of the patients with urolithiasis according to the sex

We found that the average age of calculi appearance in the men population is 47 years and 45ans for the women population(Fig. 3).

Statistical analysis based on urolithiasis type distribution according to the sex, showed that Type 2 mostly dominates in men population while type 4 dominates in women population (Fig. 4).

TABLE I. Data Table

C1	C2	CA	AU0	AU2	PAM	BR	Cys	Prot	Uram	Cal	Trg	Mps	Wk	Pacc	ocp	urna	inc	oxyp	nex	poly
97	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
85	0	5	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0
72	20	5	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
7	0	80	0	0	10	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 3. Number of Cases listed by age, group and sex

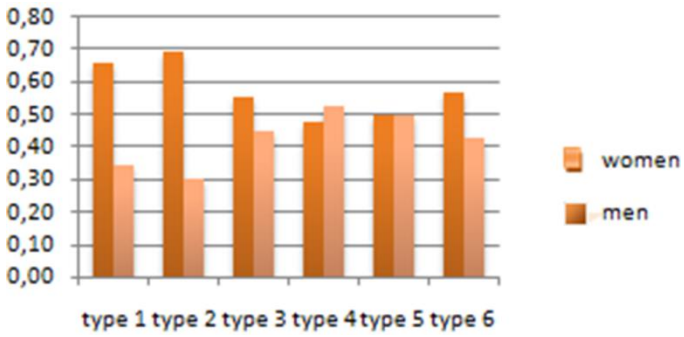


Fig. 4. The urolithiasis type distribution according to the sex

The principal component analysis (PCA) is a method of data analysis family and more generally of multivariate statistics, which involves transforming variables linked together (called "correlated") in new variables decorrelated from each other. These new variables are named "principal components" or "principal axis" . It allows the practitioner to reduce the number of variables and produce the least redundant information[5].

In order to extract as much information as possible and to have a global view on the data, a principal components analysis (PCA) was applied to urolithiasis features to reduce the components number (Fig. 5). The PCA showed that there are correlations between the various components. It is therefore a classification problem with primary variables reduction.

By applying a PCA, we have been able to reduce the number of components from 21 to 11, added to age and sex information, we obtained 13 features for the model (TABLE II).

IV. Methods and Tools

A. Decision trees

Decision trees are a type of structures that may deduce a final result from successive decisions. To span a decision tree searching for a solution it is necessary to start from the root. Each node is an atomic decision. Each sub-tree answer allows to move in the one of the child node direction. Gradually, we go down in the tree up to finding a leaf. The leaf represents the answer which the tree gives to the tested sample [6].

The algorithm used to generate decision trees is the C4.5 algorithm, it completely depends on the ID3 algorithm, but has been proposed to overcome the ID3 algorithm limitations.

B. Boosting

Boosting algorithm [7,8] is a machine learning method, precisely it belongs to meta algorithm family. There are several variations of boosting algorithms, some of them are applied to multiclass problems like AdaBoost.MH [9].

One of the main ideas of AdaBoost, is to set at each steps $1 \leqslant t \leqslant T$, a new prior probability distribution Dt for learning samples based on the algorithm results in the previous step. The weight to "t" step for example (xi, ui) of index i, where xi is sample and ui is a class, is denoted pt (i). Initially, all examples have the same weight, then at each step the weights of misclassified examples are increased, forcing the learner to focus on the difficult examples of the training sample.

Many classification studies [8, 9] showed only the Boosting algorithm effectiveness on simple decision rules.

In order to experimentally select a best decision trees for the boosting algorithm, many decision trees have been generated separately. The boosting algorithm was performed on these decision trees.

C. Proposed method

	C1	C2	CA	AU0	AU2	PAM	BR	Cystine	Prot	UrAm	Calcite	Trg	Mps	WK	pacc	ocp	urna	inc	oxypurinol	nexbrit	polysa
C1	1																				
C2	0,0757	1																			
CA	-0,5565	-0,0896	1																		
AU0	-0,101	-0,1798	-0,16609	1																	
AU2	-0,2679	-0,1912	-0,1612	0,7983	1																
PAM	-0,2869	-0,2936	-0,0964	-0,0961	-0,0656	1															
BR	-0,0694	0,1008	0,0198	-0,0054	0,01683	-0,0184	1														
Cystine	-0,2215	-0,1411	-0,1335	-0,0751	-0,0468	-0,0455	-0,0131	1													
Prot	0,1401	-0,0601	0,3889	-0,3186	-0,2523	-0,0145	-0,0685	-0,1957	1												
UrAm	-0,1412	0,0018	-0,0208	0,1307	0,0608	-0,0034	-0,0143	-0,0426	-0,0078	1											
Calcite	-0,0565	0,5261	-0,0454	-0,0248	-0,0608	-0,0161	-0,0046	-0,0115	0,0253	-0,0151	1										
Trg	-0,0465	-0,0951	-0,1492	-0,0557	-0,0371	0,8156	-0,0104	-0,0258	0,1068	-0,0328	-0,0091	1									
Mps	0,0684	-0,0448	-0,0247	-0,0326	-0,0218	-0,0212	-0,0061	-0,0151	0,0391	0,8912	-0,0054	-0,0121	1								
WK	-0,0065	0,0178	0,2659	-0,0474	-0,0354	-0,0161	-0,0099	-0,0246	0,0707	-0,0297	-0,0087	0,9432	-0,0114	1							
pacc	-0,1064	-0,0951	0,2421	-0,1369	-0,0932	0,1061	-0,0211	-0,0646	0,9167	-0,0416	-0,0229	-0,0423	-0,0301	0,2309	1						
ocp	-0,0274	-0,1779	0,0906	-0,0281	-0,0188	-0,0183	-0,0053	-0,0134	0,0181	-0,0171	-0,0046	0,0842	-0,0061	0,981	0,0217	1					
urna	-0,0641	-0,0561	0,0689	-0,0123	0,0074	-0,0181	-0,0052	-0,0129	0,1031	0,7941	-0,0046	-0,0103	-0,006	-0,0098	-0,0257	-0,0052	1				
inc	-0,0759	0,0003	0,1602	-0,0458	-0,0306	-0,0298	-0,0086	-0,0212	-0,0423	-0,0279	-0,0075	0,8341	-0,0099	-0,0161	0,0313	-0,0854	-0,0084	1			
oxypurinol	-0,0632	-0,0402	0,8135	-0,0199	-0,0133	0,0129	-0,0037	-0,0092	-0,0041	-0,0121	-0,0033	-0,0073	-0,0043	-0,0073	-0,0184	-0,0037	-0,0037	-0,0062	1		
nexbrit	0,0663	-0,0188	-0,0378	-0,0199	-0,0133	0,7829	-0,0037	-0,0092	-0,0602	-0,0121	-0,0033	-0,0073	-0,0043	-0,0073	-0,0184	-0,0037	-0,0037	-0,0062	-0,0026	1	
polysa	0,0434	-0,0082	-0,0186	-0,0199	-0,0133	0,0129	-0,0037	-0,0092	0,8251	-0,0121	-0,0033	-0,0073	-0,0043	-0,0073	-0,0184	-0,0037	-0,0037	-0,0062	-0,0026	-0,0026	1

Fig. 5. correlation matrix

TABLE III. FINAL DATA

Age	Sex	C1	C2	CA	AU0	AU2	PAM	Cystine	Prot	Uram	pacc	Other
41	1	0	85	15	0	0	3	0	0	0	0	0
79	1	5	0	0	70	25	0	0	0	0	0	0
50	0	0	40	55	0	0	0	0	5	0	0	0
22	0	78	15	3	0	0	0	0	2	0	2	0
34	1	7	3	40	0	0	15	0	12	0	20	3

The idea of this work consists in implementing the algorithm of boosting on decision trees. Two stages are required, first decision trees creation and then boosting algorithm application. Decision trees were directly generated from data files randomly created. However, two important factors must be taken into consideration:

- Tree depth

- Number of trees

We decided to experiment the boosting on small decision trees, constrained by their depth. Deep enough to separate data and make a decision, but not too deep to maintain general rules and avoid over-learning.

The number of trees used in boosting must be fixed, not too large for not to slow down training step, and not too small for boosting algorithm powerfulness.

V. RESULTS AND DISCUSSIONS

The used model, boosting of decision tree, generates different results according to the selected parameters. For evaluated system, two parameters must be fixed: the tree depth and the number of trees. The results are presented in terms of training rate error and runtime.

A. Evaluation according to trees depth variation

The decision tree depth used in boosting algorithm varies from 3 to 5.

TABLE IV shows the obtained results of our model while varying the depth for a boosting with 15 decision trees.

The results in TABLE IV shows that when using low trees depth (depth 3) i.e. with the simpler rules, we obtain better performances than trees with large depth (depth 5), however the runtime for small tree depth is almost doubled; 500ms for the tree with depth 3 and 249 ms for the tree with depth 5.

TABLE IV. RESULTS ACCORDING TO DECISION TREE DEPTHS

Depth	Error rate	Time(ms)
Depth 5	9%	249
Depth 4	4,5%	430
Depth 3	1,59%	500

B. Evaluation according to the number of trees variation

The number of decision trees used in the Boosting algorithm takes on the three following values: 10, 15 and 20 trees.

TABLE V illustrates the results obtained by our model while varying the number of trees for Boosting under a fixed depth equal to 4. It is shown that with a greater number, learning gives better results and therefore a lower error rate.

TABLE V. RESULTS ACCORDING TO NUMBER OF DECISION TREES

Number of decision trees	Error rate	Time(ms)
10 trees	8%	249
15 trees	4,5%	374
20 trees	2%	455

Compared to our parameters, trees depth (three possible values) and the number of trees (three possible values), you can have 9 different combinations and therefore 9 different systems according to their error rate and their runtime (Fig. 5).

Fig. 6 presents the error rate of each of the nine system combination. The system that gives the best result is the one with 15 decision trees and depth equal to 3.

The most powerful model, a compromise between execution time minimization and error rate, is the one with 15 decision trees and depth equal to 3. The details of this model and its confusion matrix are presented in Fig. 7. It happens to reach a classification accuracy equal to 98.41% , with a correct classification rate of 100% for types 3 and 5 and 99% for types 1 and 2. The execution time is 500ms.

In Fig. 8, the blue curve represents the error rate of the learning stage and the red curve represents the error rate of the validation step. The validation error is almost equivalent to the learning error, our system is efficient, with an error rate equal to 1.59% for the training step and a rate error equal to 1.35% for the validation step. The iteration number is approximately equal to 400 iterations for the two steps.

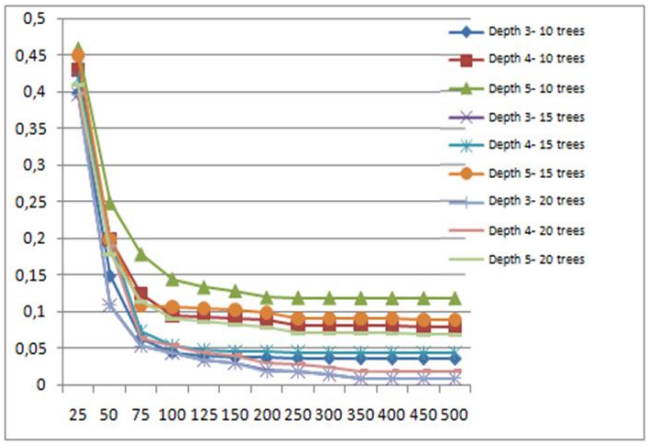


Fig. 6. Curves of results according to depth and number of trees used for Boosting based on error rate and iterations number

VI. CONCLUSION

In conclusion, this work has allowed us to achieve our objectives, namely the effective classification of urolithiasis. The boosting model proposed using 15 decision trees with a depth equal to 3 is the best one for this classification problem. Its accuracy is 98.41% for the urolithiasis classification. He correctly classified 372 cases of 378 cases.

This model with a validation error equal to 1.35%, can be considered as a promising model for the identification of urinary tract stones and determination of étiologies.

REFERENCE

[1] A. Hesse, M. Gergeleit, P. Schüller and K. Möller, 'Analysis of Urinary Stones by Computerized Infrared Spectroscopy', *Clinical Chemistry and Laboratory Medicine*, vol. 27, no. 9, 1989.

[2] V. M, d. JC and G. HM, 'Infrared analysis of urinary calculi by a single reflection accessory and a neural network interpretation algorithm.', *Clinical chemistry*, vol. 47, no. 7, pp. 1287-1296, 2000.

[3] J. Guerra-López, J. Güida and C. Della Védova, 'Infrared and Raman studies on renal stones: the use of second derivative infrared spectra', *Urological Research*, vol. 38, no. 5, pp. 383-390, 2010.

[4] I. Kuzmanovski, M. Trpkovska and B. optrajanov, Maked. Med. Pregled, 1999, 53, 251–255.

[5] H. Abdi and L. Williams, 'Principal component analysis', *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433-459, 2010.

[6] J. Quinlan, *C4.5*. San Mateo, Calif.: Morgan Kaufmann Publishers, 1993.

[7] Y.FreundandR.Schapire, "Experimentswithanewboostingalgorithm," Proc. 13th Int. Conf. Mach. Learn.,San Mateo, CA: Morgan Kaufmann, 1996,pp.148–156.

[8] J. Quinlan, 'Bagging, Boosting, and C4.5', *In Proceedings of the Thirteenth National Conference on Artificial Intelligence*, 2015.

[9] Y. Freund and R. E. Schapire, "A short introduction to boosting", *J. Jpn. Soc. Artif. Intell.*, vol. 14, no. 5, pp.771 -780 1999.

Results

Classifier performances

Error rate	0,0159
Values prediction	Confusion matrix

Value	Recall	1-Precision
C1	0,9899	0,0101
C4	0,9663	0,0000
C2	0,9918	0,0320
C3	1,0000	0,0185
C6	0,6667	0,0000
C5	1,0000	0,0000

	C1	C4	C2	C3	C6	C5	Sum
C1	98	0	1	0	0	0	99
C4	0	86	2	1	0	0	89
C2	1	0	121	0	0	0	122
C3	0	0	0	53	0	0	53
C6	0	0	1	0	2	0	3
C5	0	0	0	0	0	12	12
Sum	99	86	125	54	2	12	378

Fig. 7. Boosting model results with 15 decision trees and depth equal to 3

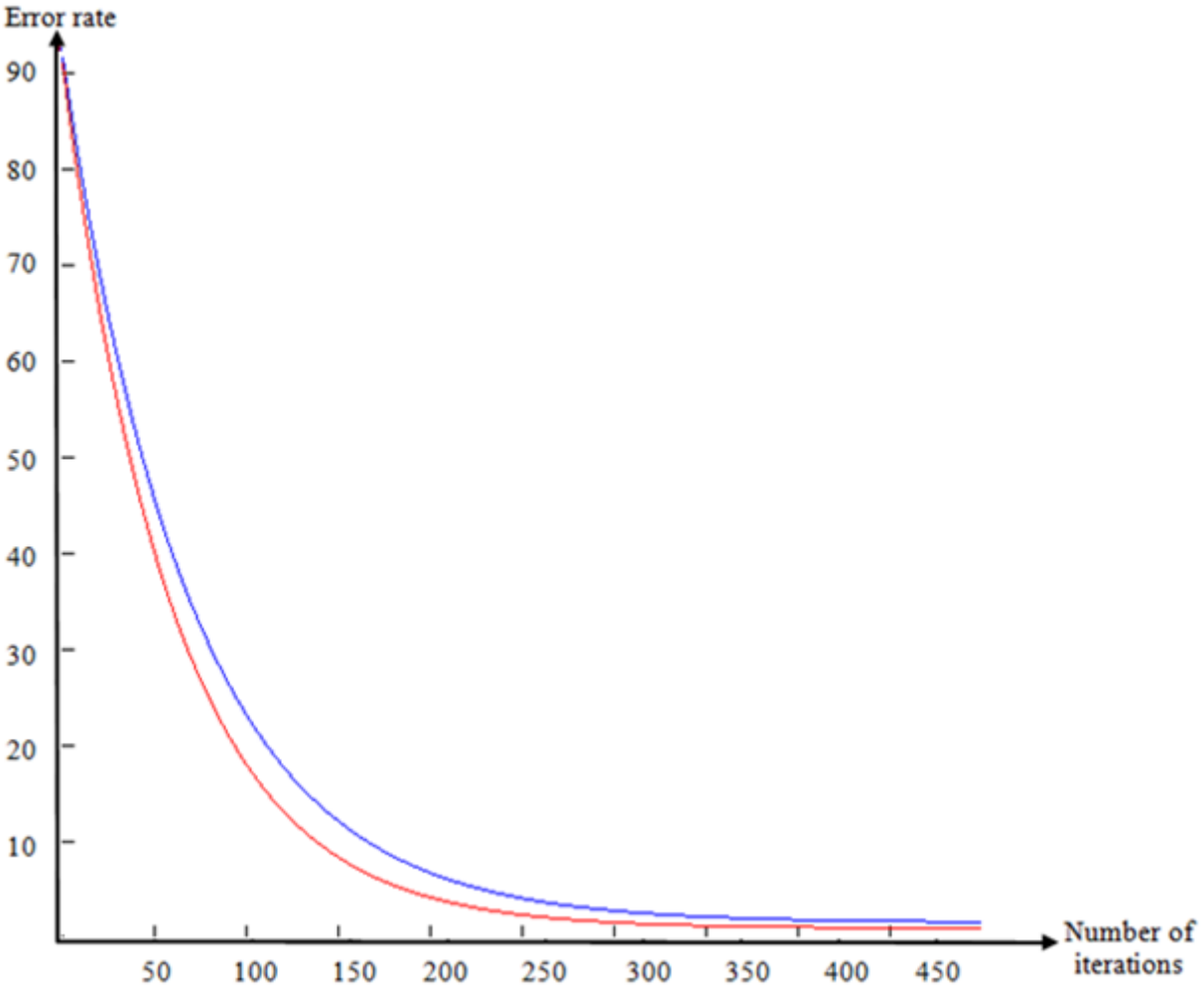


Fig. 8. Error rate Curves of training and validation steps