A Novel Machine Learning Approach for the Prediction of Subarachnoid Hemorrhage

C. Dheeba*[1], S.Vidhya[2]

[1]Research Scholar, Dept of Computer Science, [2]Head, Dept of Information Technology, K.G.College of Arts and Science, Saravanampatti, Coimbatore-641035, Tamil Nadu, India.
sugow.sd@gmail.com[1]

Abstract

Objectives: To predict outcome of patients with Subarachnoid Hemorrhage effectively by using novel ensemble classification method.
Methods: The different machine learning approaches are used to improve the outcome of patients with SAH prediction. One of such approach utilizes random forest classifier which is used for enhancing the prediction accuracy.
Findings: The outcome of patients with Subarachnoid Hemorrhage (SAH) prediction is helpful for guiding and caring patients. Such type of prediction is the most important in medical research area. Mostly SAH prediction is achieved by classification techniques such as decision rules, naive Bayesian classifiers, support vector machines, nearest neighbor classifiers and etc. However, these classifiers are not efficient for higher number of training cases.
Application/Improvements: In this paper, we propose a novel ensemble classification technique for effective classification. In which, a random forest classifier is introduced for providing efficient classification by integrating various machine learning algorithms. The algorithms used are C4.5, REPTree, and PART. The experimental results show that the best ensemble classifier and effectiveness of the random forest algorithm.
Keywords: Subarachnoid hemorrhage, Decision Tree classifier, Support vector machine, Naive Bayesian classifier, nearest neighbor classifier, Random forest algorithm.

1. Introduction

Subarachnoid hemorrhage (SAH) is a special kind of hemorrhagic stroke occurred by means of bleeding in subarachnoid area occupied by the arteries surrounding the brain. The rupture of a cerebral aneurysm, an irregular and weak dilation of the cerebral artery is the most general reason of SAH. Brain stroke is recognized owing to reduce in cerebral blood perfusion leading to cerebral ischemia. The most of SAH are caused by brain aneurysm bursting.

The diagnosis for SAH is obtained with cranial computerized tomography (CT) scan which is used to provide the level of the bleeding. The presence of an aneurysm is confirmed by cerebral angiography. If the aneurysm is identified then must be diagnosed immediately. The diagnosis can be made by endovascular means or surgically for eliminating the aneurysm preserving regular circulation.

The prediction of disease outcomes is the most imperative in the medical research area. For prediction, data mining methods are used such as prediction based on decision trees, based on decision rules, based on machine learning algorithms, and statistical methods. The ensemble techniques called boosting, bagging and genetic algorithms are used infrequently. However, the issues of different classifiers are like predictive accuracy, ability to perform with higher number of cases, noise, computational cost, effectiveness, interpretability and etc.

The main objective of this paper is to develop an ensemble classification approach for improving the predictive accuracy and effectiveness of the classifiers. This ensemble model utilizes the random forest algorithm for different machine learning algorithms to predict the SAH effectively based on the various attributes.

In [1] investigated about the various grading scales for predicting the outcome of subarachnoid hemorrhage. The modified Massachussetts General Hospital (MGH) scale based on the factors which are applicable for each patient suffering from SAH was introduced and compared to world federation of neurological surgeon scale (WFNS), Glasgow coma scale (GCS) and MGH scale for SAH. However, the predictive accuracy was less and complex to interpret in clinical settings.

In [2] developed neural network model for surgical decisions on traumatic brain injury patients. This model was developed for huge TBI patient database. The model was implemented and compared to the mathematical models for developing traumatic brain injury (TBI), medical decision support system (MDSS). The logistic regression method, multi-layer perceptron neural network model (MLP) and radial basis function (RBF) model were compared. However, this model provides accuracy identical to logistic regression.

In [3] developed neural network predictions of significant coronary artery stenosis in men. The neural network model was introduced for predicting coronary stenosis. The data from male cardiology patients was collected from national cardiac catheterization database. The patient variables were used in neural network as an input. The degree of stenosis was determined and identified coronary stenosis. However, the performance of this model was not better than other techniques such as SVM, decision trees.

In [4] developed an optimized protocol based on Neuro-evolutionary algorithms for classification of dyspeptic patients and predict their treatment. The particular optimized experimental protocol was described for classification and prediction. The applications of Neuro-evolutionary algorithms were developed based on two cases. The prediction of dyspeptic was achieved based on the two different dependent variables. But, it does not have ability to perform better than other techniques.

In [5] investigated about risk classification after aneurysmal subarachnoid hemorrhage. The prognostic rate of two multi-variate methods was evaluated for risk classification. Classification and Regression Trees (CART) and multiple logistic regressions were compared to outcome and level of consciousness from best single predictor. However, the predictive accuracy rate was less.

In [6] investigated about risk stratification for mortality in spontaneous intracerebral haemorrhage. The spontaneous ICH patient's data were collected and the variables from data were abstracted. The prediction method for mortality was developed by means of classification and regression tree technique. The predictive accuracy was evaluated based on ROC curve.

In[7] developed an intelligent brain hemorrhage by developing watershed method on segmented CT scan images. The CT images of brain were transformed into certain format and transmitted to the pre-processing. The objects in the brain images were removed and the features were extracted from every object by utilizing watershed technique. The artificial neural network was constructed by extracted features. However, the computational cost is high and over-fitting problem was occurred.

In [8] investigated about the hemorrhages in brain by means of brain CT images segmentation method automatically. The images were pre-processed and partitioned by using histogram based centroids initialization and k-means clustering algorithm concerning to pixel intensity values. Histogram analysis was performed to identify the center of the clusters and hemorrhage was predicted. However, this technique was sensitive to noise or redundant data.

In [9] developed prediction model rats in hemorrhagic shock by using random forest classifier. The input variables were prioritized by means of Breiman's method. The mean accuracy by backward elimination process was estimated by repeating 5-fold cross validation. The highest variables were sorted and best cross validated accuracy was chosen as optimal variables which are used for creating prediction model. However, the prediction is very slow compared with other techniques.

In [10] investigated about the automatic detection and classification technique to improve the prediction performance of brain hemorrhage. Initially, the skull and brain ventricles were removed and hemorrhages were isolated by using threshold method. The features were selected by genetic algorithm and extracted from every hemorrhage region. Based on multilayer neural network and k-nearest neighbor classification, the hemorrhages were predicted. However, the convergence rate was less for obtaining the better outcome.

2. Materials and Methods

2.1 Data mining methodology

In this paper, the learning process phase is presented and described by Crisp-DMmodel [11] which is defined by Cross-Industry Standard Process for Data Mining Interest Group. The knowledge discovery process involves the series of evolutionary cycles, covering one or more phases such as business and data understanding, data preparation, modelling, evaluation and deployment, repeating tasks like data preparation, feature selection, data mining techniques selection, generation of classifiers and evaluation of the outcomes.

2.2 Data sources

The data are collected from two different data cohorts holding information from all SAH patients admitted in a teaching hospital. The first database keeps information from 431 cases, from 2008 to 2011. The second database keeps information from 190 cases between 2011 and 2015 for which a smaller number of variables are recorded. The strategy followed is to utilize the first database for selecting the attributes and training the classifier and the second database for external validation.

2.3 Data understanding and preparation

The data are collected and considered such as follows:

- Original evaluation variables
- Diagnostic cranial CT scan related variables
- Diagnostic angiography related variables
- The type of diagnosis and level of consciousness before diagnosis related variables
- Outcome variables

2.4 Modeling

For various phases of knowledge discovery process, the open source tool called Weka is used. Weka is the compilation of state-of-the-art data mining algorithms and data pre-processing techniques for wide range of tasks such as data pre-processing, attribute selection, clustering and classification.

2.4.1 Attribute selection

The most important part for achievement in generation of model is attribute selection. Different subset evaluators such as classifier subset evaluator, Wrapper and search methods are combined together. The different search methods are used such as greedy stepwise, genetic search, exhaustive search and race search.

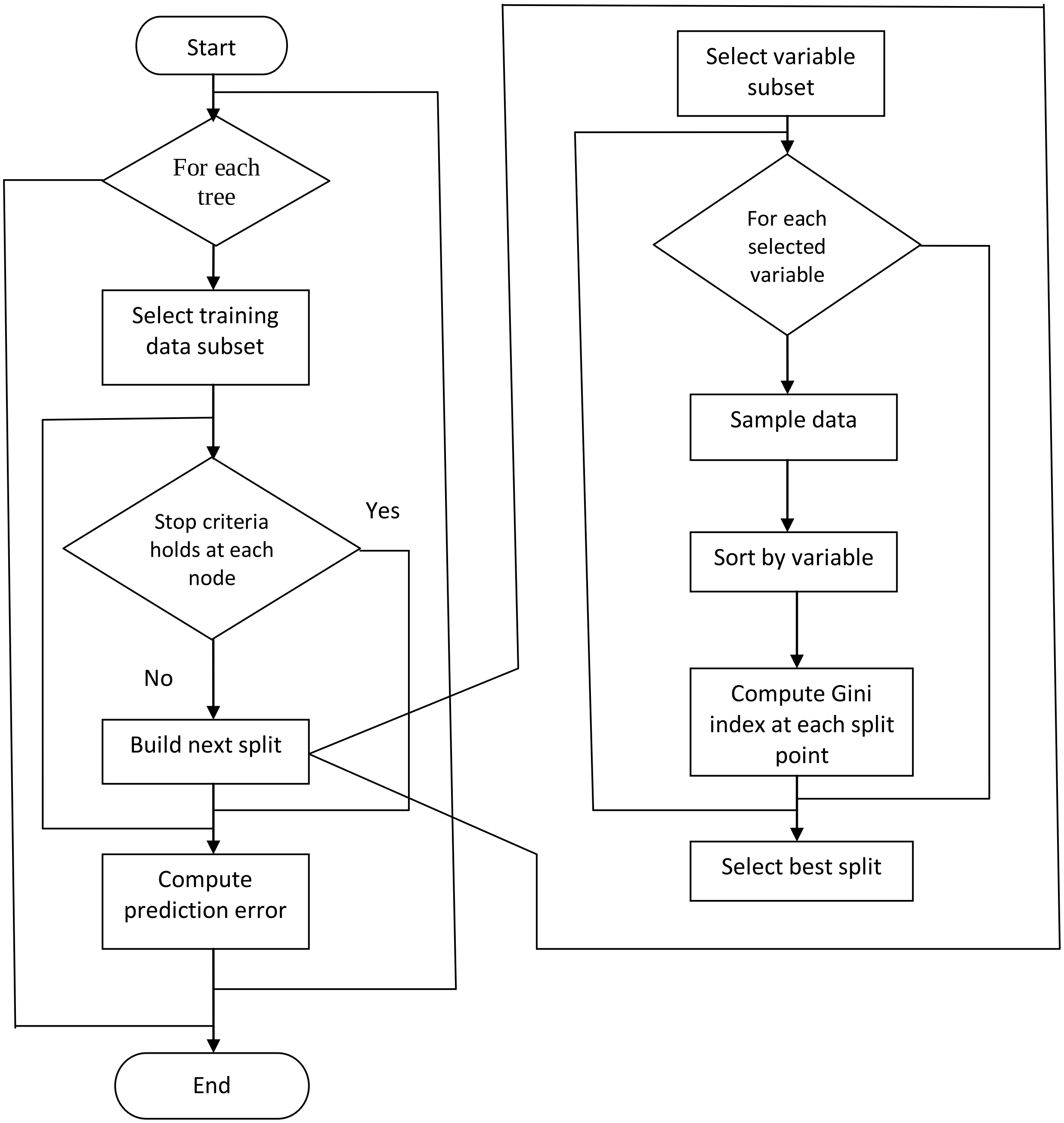
2.4.2 Classification Algorithms

The different machine learning algorithms such as decision trees and decision rules are ensemble based on random forest algorithm. The classification algorithms used are C4.5, fast decision tree learner (REPTree), partial decision trees (PART). Initially, C4.5, REPTree and PART classifiers are generating the classification rules. Then these different classification rules are optimized by random forest algorithm. Each tree is constructed based on the following steps:

1. Consider N is the number of training cases and M is the number of variables in classifiers.
2. The number of input attributes m used for determining decision at node of tree and$m < M$.
3. Select training set for this tree by choosing N times with replacement (Bootstrap samples) from all N available training cases.
4. Apply the remaining training cases to compute the error of the tree by means of predicting their classes.
5. For each node in tree, select m variables randomly.
6. Compute the best split according to the m variables in training set.
7. Each tree is completely grown and not pruned.

These steps are iterated entire trees in the ensemble and average vote of all trees are reported as random forest prediction.

Flow Diagram



3. Results and Discussion

This section presents the experimental results that are performed to prove the proposed random forest algorithm based classifiers achieves higher prediction accuracy. The performance of the proposed Random forest classifier based feature classification is evaluated in terms of precision, recall, accuracy and Kappa coefficient with existing classifiers C4.5, PART, REPTree.

3.1 Precision

Precision is defined as the value which is evaluated for feature classification at true positive prediction and false positive prediction. It is measured as follows:

$$Precision = \frac{Truepositive(TP)}{Truepositive(TP) + Falsepositive(FP)}$$

ISSN (Online) : 2277-5374
ISSN (Print) : 2277-5366

Figure 1. Comparison of precision

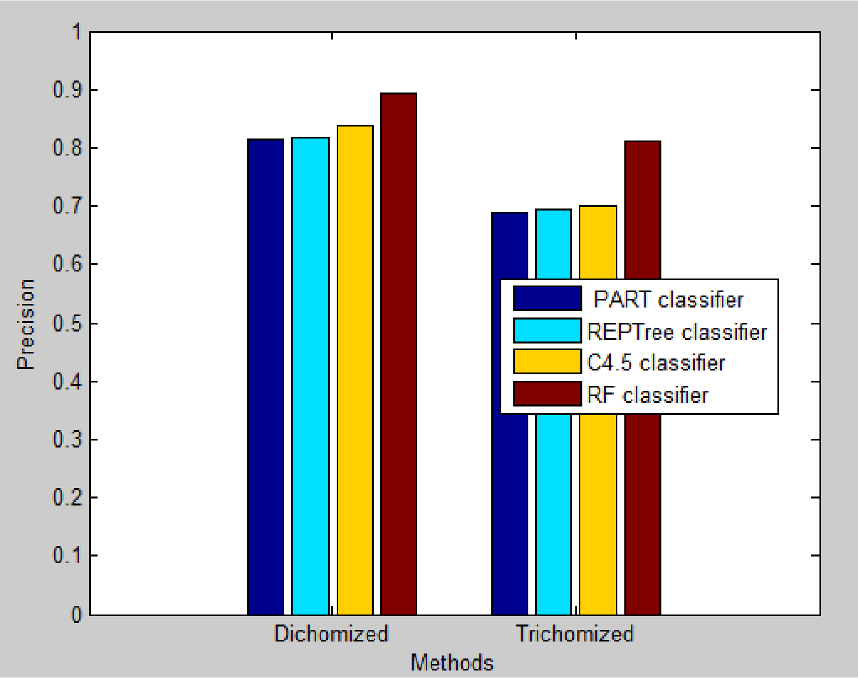


Figure 1 shows that the comparison of precision values for C4.5, PART and REPTree classifiers with Random forest classifier. It shows that the random forest classifier has high precision values for both favourable outcome and poor outcome prediction than other classifiers. The result proves that the performance of random forest classifier is better than the other classifiers.

3.2 Recall

Recall is defined as the value which is evaluated in terms of true positive rate (TP) also equivalent to hit rate. It is measured as follows:

$$Recall = \frac{Truepositive(TP)}{Truepositive(TP) + Falsenegative(FN)}$$

Figure 2. Comparison of hit rate

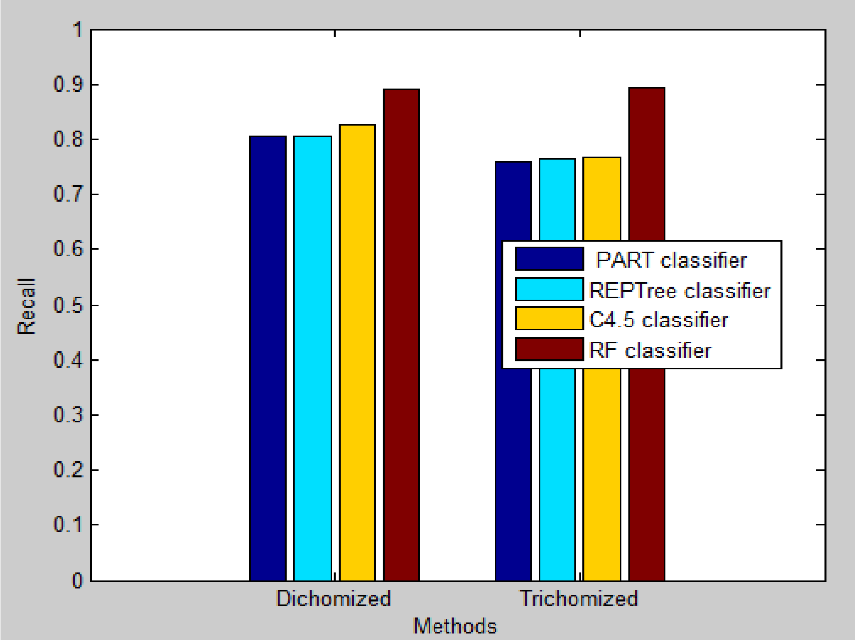


Figure 2 shows that the comparison of hit rates for C4.5, PART and REPTree classifiers with Random forest classifier in both DICHOT problem and TRICHOT problem. It shows that the random forest classifier has high hit rate for both problems than other classifiers. The result proves that the performance of random forest classifier is better than the other classifiers.

3.3 Accuracy

The accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined. Accuracy can be calculated from formula given as follows

$$\text{Accuracy} = \frac{\text{True positive + True negative}}{\text{True positive + True negative+False positive+ False negative}}$$

Figure 3. Comparison of Accuracy (%)

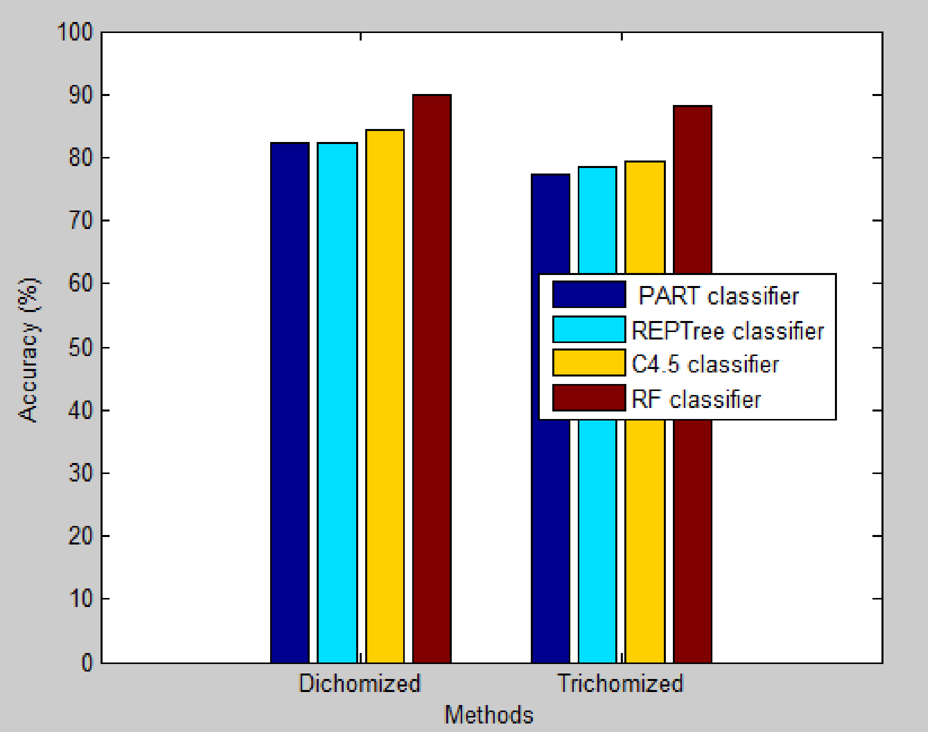


Figure 3 show that the comparison of accuracy for C4.5, PART and REPTree classifiers with Random forest classifier in both DICHOT problem and TRICHOT problem. It shows that the random forest classifier has high accuracy rate for both problems than other classifiers. The result proves that the performance of random forest classifier is better than the other classifiers.

3.4 Kappa coefficient

Kappa coefficient is a statistic which is used for correcting the degree of agreement between the classifier's predictions and reality by considering the fraction of predictions that occurred by the chance. It is computed as follows:

$$K = \frac{P_{observed} - P_{chance}}{1 - P_{chance}}$$

Figure 4. Comparison of kappa coefficient

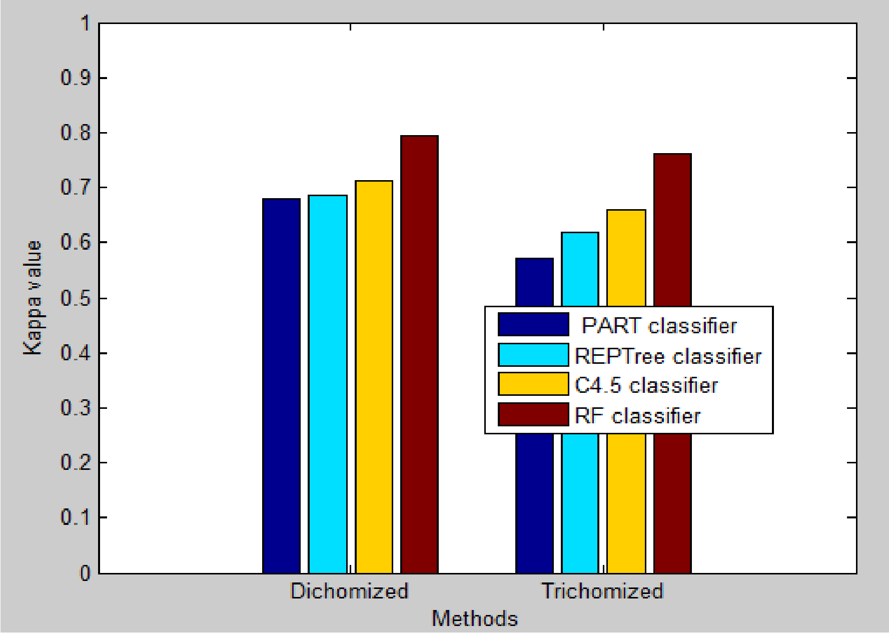


Figure 4 shows that the comparison of kappa coefficient for C4.5, PART and REPTree classifiers with Random forest classifier in both DICHOT problem and TRICHOT problem. It shows that the random forest classifier has high kappa coefficient for both problems than other classifiers. The result proves that the performance of random forest classifier is better than the other classifiers.

4. Conclusion

This paper presented an effective predicting the outcome of patients with subarachnoid hemorrhage. The prediction is achieved based on the classification algorithms. The different machine learning algorithms are used for classification and optimized by random forest algorithm which improves the detection accuracy and also performance of the classifiers. The experimental results are showed that the random forest algorithm based ensemble method has better prediction accuracy than other classifiers. Comparison results demonstrate that our proposed method outperforms the existing techniques.

5. References

1. A. Lagares, P. A. Gomez, J. F. Alen, R. D. Lobato, J. J. Rivas, R. Alday, A. G De La Camara. A comparison of different grading scales for predicting outcome after subarachnoid haemorrhage. Acta neurochirurgica.2005; 147(1), 5-16.
2. Y. C. Li, L. Liu, W. T. Chiu, W. S. Jian. Neural network modeling for surgical decisions on traumatic brain injury patients. International journal of medical informatics. 2000; 57(1), 1-9.
3. B. A. Mobley, E. Schechter, W. E. Moore, P. A. McKee, J. E. Eichner. Neural network predictions of significant coronary artery stenosis in men. Artificial intelligence in medicine. 2005; 34(2), 151-161.
4. M. Buscema, E. Grossi, M. Intraligi, N. Garbagna, A. Andriulli.M. Breda. An optimized experimental protocol based on neuro-evolutionary algorithms: application to the classification of dyspeptic patients and to the prediction of the effectiveness of their treatment. Artificial intelligence in medicine. 2005;34(3), 279-305.
5. T. P. Germanson, G. Lanzino, G. L. Kongable, J. C. Torner, N. F. Kassell. Risk classification after aneurysmal subarachnoid hemorrhage. Surgical neurology, 1998;49(2), 155-161.
6. O. Takahashi, E. F. Cook, T. Nakamura, J. Saito, F. Ikawa, T. Fukui. Risk stratification for in-hospital mortality in spontaneous intracerebral haemorrhage: a Classification and Regression Tree analysis. Qjm. 2006;99(11), 743-750.
7. U. Balasooriya, M. S. Perera. Intelligent brain hemorrhage diagnosis system. In IT in Medicine and Education (ITME), 2011 International Symposium onIEEE. 2011, December; 2, 366-370.

8. B. Sharma, K. Venugopalan. Automatic segmentation of brain CT scan image to identify hemorrhages. International Journal of Computer Applications.2012; 40(10), 1-4.
9. J. Y. Choi, S. K. Kim, W. H. Lee, T. K. Yoo, D. W. Kim. A survival prediction model of rats in hemorrhagic shock using the random forest classifier. In 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, 2012, August, 5570-5573.
10. B.Shahangian, H. Pourghassem. Automatic brain hemorrhage segmentation and classification in CT scan images. In Machine Vision and Image Processing (MVIP), 2013 8th Iranian Conference on IEEE, 2013, September, 467-471.
11. P. de Toledo, P. M. Rios, A. Ledezma, A. Sanchis, J. F. Alen, A. Lagares. Predicting the outcome of patients with subarachnoid hemorrhage using machine learning techniques. IEEE Transactions on Information Technology in Biomedicine. 2009; 13(5), 794-801.

The Publication fee is defrayed by Indian Society for Education and Environment (iSee). www.iseeadyar.org

Cite this article as:

C. Dheeba, S.Vidhya. A Novel Machine Learning Approach for the Prediction of Subarachnoid Hemorrhage. Indian Journal of Education and Information Management. Vol 5 (3), March 2016.