IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-2, Feb 2019 E-ISSN: 2347-2693

Performance Evaluation of Machine Learning Techniques for the Classification of BUPA Liver Disorder

S. Raghavendra[1*], J. Santosh Kumar [2], B. K. Raghavendra[3], S. K. Shivashankar[4]

[1]Department of Computer Science and Engineering, Christ Deemed To Be University, Karnataka, India
[2,3]Department of Computer Science and Engineering, KSSEM, Karnataka, India
[4]Department of Computer Science and Engineering, PESCE, Mandya, India

[*]*Corresponding Author: raghav.trg@gmail.com, Tel.:+91 9740857501*

DOI: https://doi.org/10.26438/ijcse/v7i2. 864869 | Available online at: www.ijcseonline.org

Accepted: 20/Feb/2019, Published: 28/Feb/2019

***Abstract*—** Liver is an important organ which plays major role in digesting food, removing poisons and stocking energy. One major challenge is to identify the Liver disorder using its ambiguous symptoms due to this many people's are suffering like anything. So to overcome the challenges we have proposed a method to identify the disorder which in turn will help medical field and society. Data mining is nothing but the process of viewing data in different angle and compiling it into appropriate information. Technically the data mining can be considered as the sequence of steps followed for searching patterns or identifying correlations between large numbers of fields within a huge relational database. Recent improvements in the area of data mining and machine learning have empowered the research in biomedical field to improve the condition of general health care. Data mining techniques are applied to different medical domains to improve the medical diagnosis. Improving the accuracy of the classification and improving the prediction rate of medical datasets are the main tasks/challenges of medical data mining. Since the wrong classification may lead to poor prediction, there is a need to perform the better classification which further improves the prediction rate of the medical datasets. When medical data mining is applied on the medical datasets the important and difficult challenges are the classification and prediction. In this proposed work we evaluate the performances of machine learning techniques like Logistic Regression (LR), Artificial Neural Networks (ANN), and ANN with k-fold Cross Validation Sample (CVS) with Feature Selection Methods (FSMs) using Percentage Split (PS) as test option on Liver Disorder Datasets. The performance of the proposed model is measured in the form of classification accuracy. Performance of proposed work is assessed as classification accuracy. The work deliver the better accuracy for reduced set of attributes compared with full set attribute and we state that those are the very important tests compared to all tests to identify the disorder.

Keywords— artificial neural Networks; ANN; classification accuracy; CA; backward elimination; BE; classification accuracy; CA; entropy evaluation (EE): feature subset selection methods; FSM's; forward selection; FS; logistic regression; LR.

I. INTRODUCTION

According to National statistics in UK, Liver disorder has the fifth most common reason of mortality. It is also one of the second reason of death amongst all gastral diseases in the US. International Liver Congress States that the number of sufferers from a chronic Liver condition in the European zone is about 29000 thousand and 30000 thousand have a Liver disorder in America. So the Liver is the biggest strong organ in the human body. Recent improvements in the area of data mining and machine learning have empowered the research in biomedical field to improve the condition of general health care. In many parts of the world the tendency for maintaining long-lasting records consisting of medical data is becoming an accepted practice. In addition to this, the newer medical equipment's and the techniques used in diagnosis, produces composite and huge data. Therefore, to handle these ill-structured biomedical data, intelligent algorithms for data mining and machine learning are required in order to take logical reasoning from the saved raw data, which is considered as medical data mining. Within the medical data, the medical data mining searches for patterns and relationships which can provide useful information for appropriate medical diagnosis [1]. Data mining techniques are applied to different medical domains (health care databases or medical datasets) to improve the medical diagnosis.

To check for any invisible patterns inside the medical datasets, medical data mining is strongly recommended. In

medical data mining, the actual tasks (challenges) are the classification and prediction of medical datasets. To manage these tasks the following methods are used most often.

LR: LR is one of the data mining methods used for analyzing problems where the outcome is determined based on one or more independent variables. A dichotomous variable is used to measure the outcome. In LR, the non-independent variable is dichotomous or binary i.e., it consists of data represented as 0 (FALSE, failure, etc.) or as 1 (TRUE, success, etc.) [2]. In various biomedical fields such as cancer analysis, survival forecast, kidney transplant etc. [3] [4], LR has been widely used. Even in statistics, it is a well-established and a powerful method. It is suggested that LR has to be compared to data mining techniques while performing medicinal data mining [5]. LR is implemented on the health care databases for detecting the patterns which are useful for either forecasting or determining the diseases along with take the remedial measures for handling such diseases [6].

ANN: ANN is one among the various fields of Artificial Intelligence. The human brain architecture is the main inspiration behind the development of the model. ANNs are successfully used in various disciplines such as environmental science, study of human mind, study of numbers, study of medicine, study of computers etc. ANNs are also being used in many business areas like accounts and audits, funding, managing and decision making, promotion and manufacture etc. ANNs have turned out to be a well-liked model and recently they are used to identify diseases and to forecast the patients' survival proportion [7]. ANN models or "neural nets" are also called by different names. Whatever the name is; each one of these models tries to give good performance through compact interconnection of uncomplicated computational elements. For many years these models have been studied with a hope of achieving the performance like humans in the field of speech and image recognition [8].

SVM: In machine learning SVMs [9] are the models used for supervised learning accompanying with other learning algorithms which can analyze data used for regression and classification. For any set of training examples given, each of them is marked as fitted to one or other group, an SVM training algorithm constructs a model that allocates new examples to a single category or the other, constructing it a non-probabilistic binary linear classifier.

RF: RF are an ensemble learning method for regression, classification and other jobs, that functions by making an assembly of decision trees at training time and generating the class that is the classification or regression of the distinct trees. [10]

II. RELATED WORK

For optimizing the parameter for SVM, an Adjusted Bat algorithm (ABA) is proposed. The experiments are conducted on the liver disorder dataset. The experimental result was compared with the Grid-SVM and other approaches. Based on the result, ABA-SVM is considered as a better classifier than Grid-SVM and compared to other approaches like PSO-SVM and PTVSPSO-SVM, the ABA-SVM achieved better classification accuracy [11].

A method similar to PCA was used to select the important attributes was developed. These attributes are given as an input to the feed forward ANN. The result achieved by the method is measured up with other methods of the feature selection like Tarr's, RUCK's, PCA and t-test. The new model was applied on the liver disorder dataset. Testing is done using 20% of data and remaining 80% is used for training. The proposed method achieved good classification accuracy with less number of attributes [12].

A unique algorithm is presented for the induction of full oblique decision trees (EFTI). The algorithm depends on single and special evolutionary algorithm, which generates a full decision tree by altering the node coefficients and structure of the complete tree at the time of evolution. EFTI algorithm is often used in embedded applications, since it uses small resources for computation when compared with decision tree inference algorithm. The algorithm is implemented on liver disorder dataset and the result was compared with other approaches based on decision tree. The proposed algorithm generates better result the other [13].

A growing-pruning spiking neuron network (GPSNN) consisting of 2 stage learning algorithm is developed for handling the problems of pattern classification. The GPSNN consisted of three layers and two stages of learning algorithm. The GPSNN was experimented on liver disorder dataset. The outcomes are evaluated with batch and online spiking neuron. From the result, it was identified that GPSNN achieved better accuracy that the other [14].

EUCAFES is a robust filter which works on feature weighing approach. Feature weighing approach is used to calculate the weights of the binary feature and gives the detailed information related to feature based on continuous weight. RBF is applied to determine the consequence of feature subsets. The technique is applied on liver disorder dataset. Based on the result we can see that RBF-DDA achieved good accuracy with less number of attributes [15].

In addition to earlier kernel Fisher Discriminent (KFD), by employing heterogeneous kernel model, an iterative

algorithm is proposed for KFD. The new KFD selections of kernels are automatic. The proposed method was implemented on liver disorder dataset. From the experiment it was observed that the new KFD gives better classification than the earlier KFD [16].

The kNN classifiers are delicate to noise and the outliers present inside the training dataset. Two approaches of depuration algorithm are employed to edit training data. For kNN classifier, to edit the data neural network ensemble is made use of. The method was implemented on liver disorder dataset. From the result, we can see that kNN is much better than the two methods of depuration algorithm [17]. For data reduction or compression a method based on multidimensional scaling is proposed. This method can produce shorter vectors from data vectors of high dimension, but with some loss of information. The formal model for data reduction in Bayesian framework is the Bayesian networks. The method was applied on liver disorder datasets. The result of kNN is compared with Naïve Bayes. Naïve Bayes performs better than kNN [18].

Data mining algorithms is also used during generating the insights of the agnostic analytics data. From the result it was identified that Naïve Bayes' Classifier is identified as the better algorithm [19].

III. PROPOSED FRAMEWORK

The proposed model is shown in the following Figure 1.
Proposed model consists of the following steps:

i) First step is the collection of BUPA liver dataset.
ii) Preprocessing is done for any missing values.
iii) For preprocessed data we apply entropy evaluation method.
iv) Based on the entropy value we apply the FSMs like FS and BE. This results in generating different subsets of attributes.
v) For each attribute we evaluate the performance of LR, ANN, and ANN with 10-fold CVS with percentage split as test option.

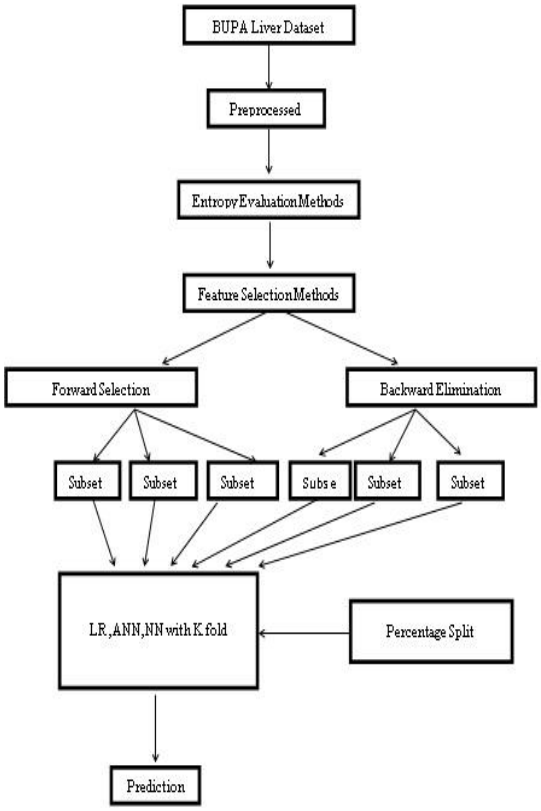

Figure 1: Proposed model for Prediction

Finally we identify the subset that achieves the best CA as the best attributes for the prediction of Liver disorder.

IV. RESULTS AND DISCUSSION

For Liver Disorder Dataset, we apply FSMs like FS and BE after applying finding the entropy values of each attribute using the Entropy Evaluation Method and we get different subsets of attributes.

Table 1 shows Different subsets of attributes obtained after applying FS method based on entropy value each attribute of Liver disorder dataset.

Table 1: After FS different subset organization

Subset No.	Subset of Attributes	No. of Attributes
1	drinks, selector	2
2	sgot, drinks, selector	3
3	sgpt, sgot, drinks, selector	4
4	alkphos, sgpt, sgot, drinks, selector	5
5	mcv, alkphos, sgpt, sgot, drinks, selector	6

For full attribute set of liver disorder dataset the classification accuracy attained is LR 60.70% with 66% split ratio, NN without CVS is 78.77% with 66% split ratio, **NN with CVS is 78.77%** with 80% split ratio as shown in Table 2.

Table 2: Classification accuracy achieved for full set of attributes

Technique Used for Finding Classification Accuracy	Percentage Split			
	66%	**70%**	**75%**	**80%**
LR	60.70	58.6	53.90	54.60
NN	78.77	78.76	76.25	73.63
NN with fold CVS	78.29	77.76	76.54	**78.77**

The classification accuracy achieved by different subsets of attributes of Table 1 based on forward selection is shown from Table 3 through Table 9.

Table 3: Classification accuracy achieved for drinks and selector combination of attributes

Technique Used for Finding Classification Accuracy	Percentage Split			
	66%	70%	75%	80%
LR	52	45.60	47.80	49.50
NN	76.13	76.22	**76.70**	75.99
NN with 10 Fold	75.56	75.50	74.54	75.53

Table 4: Classification accuracy achieved for sgot, drinks and selector combination of attributes

Technique Used for Finding Classification Accuracy	Percentage Split			
	66%	**70%**	**75%**	**80%**
LR	59.60	56.80	62.80	62.20
NN	77.29	77.41	**77.58**	75.30
NN with 10 Fold	75.81	75.10	75.30	76.57

Table 5: Classification accuracy achieved for sgpt, sgot, drinks and selector combination of attributes

Technique Used for Finding Classification Accuracy	Percentage Split			
	66%	**70%**	**75%**	**80%**
LR	62.70	61.40	59.70	61.60
NN	77.20	76.69	74.54	72.69
NN with 10 Fold	77.04	76.32	76.81	**77.38**

Table 6: Classification accuracy achieved for alkphos, sgpt, sgot, drinks and selector combination of attributes

Technique Used for Finding Classification Accuracy	Percentage Split			
	66%	**70%**	**75%**	**80%**
LR	58.10	56.20	54.00	55.60
NN	77.37	76.35	73.94	72.20
NN with 10 Fold	77.24	77.12	76.66	**77.71**

Table 7: Classification accuracy achieved for mcv, alkphos, sgpt, sgot, drinks and selector combination of attributes

Technique Used for Finding Classification Accuracy	Percentage Split			
	66%	**70%**	**75%**	**80%**
LR	58.30	55.50	53.10	56.60
NN	76.64	75.66	73.23	70.50
NN with 10 Fold	77.28	76.84	77.07	**78.36**

Table 8 shows Different subsets of attributes obtained after applying BE method based on Entropy value each attribute of Liver disorder dataset.

Table 8: After BE different subset organization

Subset No.	Subset of Attributes	No. of Attributes
1	alkphos, sgpt, sgot, gammagt, drinks, selector	6
2	sgpt, sgot, gammagt, drinks, selector	5
3	sgot, gammagt, drinks, selector	4
4	gammagt, drinks, selector	3
5	gammagt, selector	2

The classification accuracy achieved by different subsets of attributes of Table 8 based on backward elimination is shown from Table 9 through Table 13.

Table 9: Classification accuracy achieved for alkphos, sgpt, sgot, gammagt, drinks and selector combination of attributes

Technique Used for Finding Classification Accuracy	Percentage Split			
	66%	**70%**	**75%**	**80%**
LR	60.50	58.70	54.90	54.10
NN	**80.59**	80.04	77.98	76.69
NN with 10 Fold	78.53	78.32	78.63	80.10

Table 10: Classification accuracy achieved for sgpt, sgot, gammagt, drinks and selector combination of attributes

Technique Used for Finding Classification Accuracy	Percentage Split			
	66%	**70%**	**75%**	**80%**
LR	65.10	63.90	59.30	59.70
NN	**80.14**	79.80	78.53	76.81
NN with 10 Fold	78.90	77.37	78.48	79.09

Table 11: Classification accuracy achieved for sgot, gammagt, drinks and selector combination of attributes

Technique Used for Finding Classification Accuracy	Percentage Split			
	66%	**70%**	**75%**	**80%**
LR	62.80	60.00	63.40	62.70
NN	77.53	77.42	**77.60**	77.18
NN with 10 Fold	75.70	75.01	75.07	75.68

Table 12: Classification accuracy achieved for gammagt, drinks and selector combination of attributes

Technique Used for Finding Classification Accuracy	Percentage Split			
	66%	70%	75%	80%
LR	63.20	60.60	62.50	61.60
NN	77.59	77.60	**78.12**	76.00
NN with 10 Fold	76.45	76.10	76.36	77.54

Table 13: Classification accuracy achieved for drinks and selector combination of attributes

Technique Used for Finding Classification Accuracy	Percentage Split			
	66%	70%	75%	80%
LR	64.80	63.70	64.50	61.50
NN	77.51	77.56	**77.89**	77.39
NN with 10 Fold	76.83	76.36	75.76	76.37

V. CONCLUSION AND FUTURE SCOPE

In the proposed research work Feature Selection Methods like forward selection and backward elimination is applied on the Liver Disorder dataset. This results in different subsets of attributes. For each subset, we evaluate the performances of the machine learning techniques like LR, ANN, and ANN with 10-fold CVS. From the experimental result it was found that the classification accuracy achieved is 80.59% by reduced set of attributes of only 5 (alkphos, sgpt, sgot, gammagt, drinks) by reduced set of attributes by ANN using percentage split and is better compared to the classification accuracy achieved by full set of attributes is 78.77%. By this we can reduce the number of tests that is required to predict the Liver disorder and avoid the unnecessary tests which reduces the cost and the time.

ACKNOWLEDGMENT

I Acknowledge every one for supporting me for doing research.

REFERENCES

[1] S. K. Wasan, V. Bhatnagar,H. Kaur, *"The Impact of Data Mining Techniques on Medical Diagnostics"*, Data Science Journal, Vol. **5**, pp.119-126, **2006**.

[2] B. K. Raghavendra, J. B. Simha, *"Evaluation of Logistic Regression Model with Feature Selection Methods on Medical Datasets"*, ACS-International Journal on Computational Intelligence, Vol.**1**, Issue.**2**, pp.35-42, **2010**.

[3] J. Chhatwal, O. Alagoz, M. J. Lindstorm, C. E. Kahn, K.A. Shaffer, E.S. Burnside, *"A Logistic Regression Model Based on the National Mammography Database Format to Aid Breast Cancer Diagnosis"*, American Journal of Roentgenology, Vol.**192**, Issue.4, pp.1117-1127, **2009**.

[4] H. Khedmat, G. R. Karami, V. Pourfarziani, S. Assari, M. Rezailashkajani, M. M. Naghizadeh. *"A Logistic Regression Model for Predicting Health-Related Quality of Life in Kidney Transplant Recipients"*, Transplantation Proceedings, Elsevier, Vol.**39**, pp.917-922, **2007**.

[5] R. Bellazzi, B. Zupan, *"Predictive Data Mining in Clinical Medicine: Current Issues and Guidelines"*, International Journal of Medical Informatics, Elsevier, Vol.**77**, Issue.**2**, pp.81-97, **2008**.

[6] S. Raghavendra, M. Indiramma M, *"Performance Evaluation of Logistic Regression and Artificial Neural Network Model with Feature Selection Methods using Cross Validation Sample and Percentage Split on Medical Datasets"*, Proceedings of the 2nd International Conference on Emerging Research in Computing, Information Communication and Applications, Vol.**2**, Elsevier Publication, pp.750-755, **2014**.

[7] B. K. Raghavendra, J. B. Simha, *"Performance Evaluation of Logistic Regression and Neural Network Model with Feature Selection Methods and Sensitivity Analysis on Medical Data Mining"*, International Journal of Advanced Engineering Technology, Vol.**2**, Issue.**1**, pp.289-298, **2011**.

[8] R.P. Lippmann, *"An Introduction to Computing with Neural Nets"*, IEEE ASSP Magazine, Vol.**4**, Issue.**2**, pp.4-22, **1987**.

[9] C. Cortes, V. Vapnik, *"Support-vector networks"*, Machine Learning, Vol.**20**, Issue.**3**, pp.273–297, **1995**.

[10] T. K. Ho, "*Random decision forests"*, Proceedings of the Third International Conference on Document Analysis and Recognition Vol.**1**, pp.278, **1995**.

[11] E. Tuba, M. Tuba, D. Simian, *"Adjusted Bat Algorithm for Tuning of Support Vector Machine Parameters"*, IEEE Congress on Evolutionary Computation, pp.2225-2232, **2016**.

[12] S. J. Perantonis, V. Virvilis, *"Input Feature Extraction for Multilayered Perceptrons Using Supervised Principal Component Analysis"*, Neural Processing Letters, Vol.**10**, Issue.**3**, pp.243-252, **1999**.

[13] B. Vukobratovic, R. Struharik, *"Evolving Full Oblique Decision Trees"*, IEEE 16th International symposium on Computational Intelligence and Informatics (CINTI), pp.95-100, **2015**.

[14] S. Dora, S. Sundaram, N. Sundararajan, *"A Two Stage Learning Algorithm for a Growing-Pruning Spiking Neural Network for Pattern Classification Problems"*, International Joint Conference on Neural Networks (IJCNN), pp.1-7, **2015**.

[15] M. Scherf, W. Brauer. *"Feature Selection by Means of a Feature Weighting Approach"*, Technical Report FKI-221-97, Institut fur Informatik, Technische Universitat Munchen, **1997**.

[16] G. Fung, M. Dundar, J. Bi, B. Rao, *"A Fast Iterative Algorithm for Fisher Discriminant Using Heterogeneous Kernels"*, Proceedings of the 21st International Conference on Machine Learning, ACM, **2004**.

[17] Y. Jiang, Z. Zhou, *"Editing Training Data for kNN Classifiers with Neural Network Ensemble"*, International Symposium on Neural Networks, Springer, pp.356-361, **2004**.

[18] P. Kontkanen, J. Lahtinen, P. Myllymaki, H. Terri, *"Unsupervised Bayesian Visualization of High-Dimensional Data"*, Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.325-329, ACM, **2000**.

[19] A. J. Jebamalar, *"Efficiency of Data Mining Algorithms Usesd in Agnostic Data Analytics Insight Tools"*, International Journal of Scientific Research in Network Secutity and Communication, Vol.**6**, Issue.**6**, pp.14-18, **2018**.

Authors Profile

Mr.Raghavendra S. Pursed Bachelor of Engineeriing and Master of Technology and Ph.D from VTU, Belagavi, Karnataka and currently working as Associate Professor in the Department of Computer Science and Engineering, Christ Deemed To Be University, Bengaluru, Karnataka. He has published more than 10 research papers in reputed international journals. His main research work focuses on Data mining and Big Data Analytics. He has 14 years of teaching experience and 5 years of Research Experience.

Mr. Santosh Kumar Jankatti Pursed Bachelor of Engineering and Master of Technology from VTU, Belagavi, Karnataka. He is pursuing Ph.D from VTU Belagavi, Karnataka and currently working as Associate Professor in Department of Computer Science and Engineering, KSSEM, Bengaluru, Karnataka. He has published more than 5 research papers in reputed international journals. His main research work focuses on Data mining and Big Data Analytics, IoT. He has 10 years of teaching experience and 3 years of Research Experience.

Mr. Raghavendra B. K. Pursed Bachelor of Engineeriing from Bangalore University, Bengaluru, and Master of Technology from VTU, Belagavi, Karnataka. He pursued Ph.D from VTU, Belagavi, Karnataka and currently working as Professor in the Department of Computer Sciences and Engineering, KSSEM, Bengaluru. He has published more than 10 research papers in reputed international journals. His main research work focuses on Data mining and Big Data Analytics. He has 15 years of teaching experience and 10 years of Research Experience.

Mr. Shivashankar S. K. Pursed Master of Technology from VTU, Belagavi, Karnataka. He is currently working as Assistant Professor in the Department of Computer Sciences and Engineering, PESCE, Mandya. His main research work focuses on Data mining and Big Data Analytics. He has 15 years of teaching experience.

