Classification and Prediction Model using Hybrid Technique for Medical Datasets

Raghavendra S.
Research Scholar
CSE Research Centre
BMSCE, Bengaluru-19

Indiramma M., PhD
Professor
Department of CSE
BMSCE, Bengaluru-19

ABSTRACT

For processing of large amount of data numerous techniques are used. Data Mining is one of the technique that is used most often. To process these data, Data mining combines traditional data analysis with sophisticated algorithms. Medical data mining is an important area of Data Mining and considered as one of the important research field due to its application in healthcare domain. Classification and prediction of medical datasets poses real challenges in Medical Data Mining. To cope with these challenges Logistic Regression (LR) and Artificial Neural Network (ANN) are commonly used. LR enables us to investigate the relationship between a categorical outcome and a set of explanatory variables. LR explains that there can be one or more independent variables that can determine the problem outcome. ANN resembles the human brain and here the information is processed by simple elements called neurons and signals are transmitted between the neurons. Feature subset selection selects subsets of features that are enough to explain the target concept. In this paper feature selection methods like forward selection and backward elimination using mean evaluation are used on the medical datasets. LR and ANN are applied on feature selection methods using Cross Validation Sample (CVS) and Percentage Split as test options. From the experimental results it is identified that for SPECTF dataset LR using percentage split prediction accuracy of 83.95% is achieved, for Diabetes Dataset LR using percentage split prediction accuracy of 80.46% is achieved, and for Liver Disorder dataset NN using percentage split prediction accuracy of 74.75% is achieved.

Keywords

Cross Validation Sample, Data Mining, Mean Evaluation, Feature Subset Selection, Logistic Regression, Artificial Neural Network, Percentage Split.

1. INTRODUCTION

Medical data mining is an important area of Data Mining and considered as one of the important research field due to its application in healthcare domain. To improve the general quality of healthcare latest achievements in the field of machine learning and data mining are used in biomedical research. In many countries keeping permanent medical records has become a standard practice. In addition to this latest diagnostic techniques generate heterogeneous and huge amount of data. Due to the ill-structured nature of medical data, there is a requirement for intelligent machine learning and data mining algorithms to identify the logical relationship within the stored data which is called Medical Data Mining. Medical Data Mining has a great ability to identify the hidden patterns within the medical datasets [1].

Medical diagnosis is a difficult and completed task and should be carried out efficiently and precisely. Most of the medical decisions are made based on doctor's advice and experience rather than the knowledge hidden in the database. This practice may lead to errors and excess medical cost which can affect quality of medical service provided to the patients. Data Mining methods can improve the quality of medical decisions significantly [1].

Biomedical datasets are usually associated with high dimensional features. The clinical databases that provide datasets may contain systemic and human errors. The classification accuracy of machine learning schemes is affected by noisy nature, sparseness and the missing values in the datasets. So there is a need to improve the accuracy of existing medical diagnosis tools. In addition to this the characteristics of medical data and number of variables must also be considered for developing a new technique. For accurate and efficient implementation of automated system a detailed study of different techniques to be considered. Hence we propose a new hybrid model for medical predictions based on LR and ANN.

ANN is considered as an important field of Artificial Intelligence. The ANN model development was motivated by the neural architecture of human brain. ANN have been used in many fields like biology, psychology, statistics, mathematics, medical science, and computer science, accounting and auditing, finance, management and decision making, marketing and production. Recently diagnosing a disease and predicting survival ratio of patients is done using ANN [2].

Identification and removal of irrelevant and redundant data can be done using Feature Selection methods. The dimensionality of the datasets can be reduced and better analysis can be obtained using Feature Selection methods. Feature Selection in data classification has many advantages: Computational complexity is reduced due to reduction in dimensionality; Classification accuracy is increased due to noise reduction, and makes the algorithms to work faster and effectively [2].

The proposed research work is mainly focused on mean evaluation method and compares results of LR and ANN model with feature selection methods like Forward Selection and Backward Elimination using CVS and percentage split as test options.

2. LITERATURE SURVEY

Cancer Diagnosis and Survival predictions can also be done by Logistic Regression. In statistics and biomedical field Logistic Regression is a powerful and well established method. LR compares categorical outcome and a set of explanatory variables. Neural Network model is also used in the fields like biology, business, auditing etc. Between LR and NN with and without hidden layers, a performance analysis is done on publically available medical datasets. From the analysis it is confirmed that NN without hidden layer performs better [3].

In remote sensing data processing Feature selection is a key task. LR model is used with both feature selection and classification of remotely sensed images. LR model with fewer restrictive assumptions can reduce feature substantially without any significant decrease in classification accuracy [4].

Feature Selection methods like forward Selection and Backward Elimination is evaluated on publicly available medical datasets. Predictive model for classification using LR is developed using selected features. Classification accuracy, root mean square error and mean absolute error are used to measure the performance of the model. LR model with Forward Selection and Backward Elimination is more reliable than LR model [5].

Feature Selection is the process of selecting most useful features that can produce results as the original set. Feature selection algorithm based on fast clustering (FAST) is proposed and experimentally evaluated. FAST is compared with other Feature Selection algorithms like FCBF, ReliefF etc, with respect to classifiers, namely the probability based Naïve-Bayes, rule based RIPPER, instance based IB1 and tree based C4.5. FAST relatively produces smaller subsets and also improves the performance of above four types of classifiers [6].

For prediction and diagnosis of various diseases with good accuracy Data Mining techniques are widely used. The two most successful data mining tools, Neural Networks and Genetic Algorithms are used for prediction of heart disease. To initialize the Neural Network weights global optimization advantage of Genetic Algorithms is used. Using this technique the learning is faster, more stable and accurate [7].

While diagnosing a disease the patient has to undergo various tests which are costly and sometimes all the tests are not required. For automated detection of diabetes mellitus an intelligent and effective methodology is designed based on Neural Network. There exists many methods to diagnose diabetes mellitus but the main drawback is that the patient has to undergo various tests. Using this method user can check whether he/she is suffering from diabetes mellitus or not [8].

By using entropy technique and ANN automatic prediction system for Dengue Haemorrhagic-Fever outbreak risk is developed. To reduce the data redundancy and to retain only relevant data the information is preprocessed. The external factors such as temperature, relative humidity, and rainfall are considered during information extraction. To predict the risk of Dengue Hemorrhagic- Fever outbreak a supervised Neural Network is used and good accuracy was achieved [9].

To provide prognosis and detailed understanding of the classification of neurodegenerative diseases Data Mining technique is used. Major risk factors responsible for Alzheimer’s disease and Parkinson’s disease are considered and a new model for classification is developed. Neural Network and machine learning methods are also developed. For Alzheimer’s disease genetic factors, diabetes, age and smoking were the strongest risk factors and for Parkinson’s disease stroke, diabetes, genes and age were considered as risk factors [10].

Pattern Recognition and Data Mining techniques are used in risk prediction of cardiovascular medicine. The data to be modeled is classified using classification Data Mining technique. The problem with conventional medical scoring system is that there exists intrinsic linear combination and they can’t model the nonlinear complex interactions [11].

A large amount of heterogeneous data is usually generated from the modern medicine. The transformation of this huge quantity of data to useful information and knowledge is the biggest challenge. The Data Mining techniques permit the discovery of medicine and support the predictions on the individual [12].

The prediction of heart disease is treated as most difficult task in the field of medical sciences. A hybrid model consisting of Genetic Algorithm with back propagation technique is developed for prediction of heart disease which can’t be observed by the naked eye is the most threatening one. The hospitals stores information about patients in the form of images, text, charts and numbers. A prototype is developed to determine and extract patterns and relations in heart disease record database [13].

In recent years the heart attacks are increasing in India. About 60% of daily deaths are due to this. Ten risk factors for Coronary Heart Disease have been identified and these factors are clinically validated by cardiologists based on their medical experience [14].

Using Data Mining Technique an intelligent system is developed which can retrieve hidden data from stored database. The system can answer complex queries for diagnosing heart disease and can help doctors to make intelligent clinical diagnosis [15].

In our previous work we evaluated the performance of LR and ANN with Feature Selection methods using CVS and Percentage Split as test options on publicly available medical datasets based on entropy evaluation. From the experimental result it was identified that ANN with Backward Elimination feature selection method using percentage split gave better result [16].

The proposed system evaluates the performance of LR and ANN with Feature Selection Methods using CVS and Percentage Split as test options for Diabetes Dataset, SPECTF Dataset and Liver Disorder Datasets using Mean Evaluation Method. Further the result can be improved for better prediction.

3. PROPOSED FRAMEWORK

The proposed hybrid framework is shown in Figure 1. This involves the following steps:

1. The medical datasets used in the proposed framework is of ARFF format.
2. The preprocessing is done for missing values in the attributes, if found.
3. Calculate the mean value for the entire attributes.
4. Compare the mean value of an attribute with each and every value present in that attribute. If the value is greater than or equal to mean value then the column value is considered as 1, otherwise 0.
5. Count the numbers of 1’s in each attribute and apply feature selection methods based on maximum count of 1’s in the attributes.
6. Compare performance of ANN and LR model using cross validation sample and percentage split for the subset of attributes obtained.

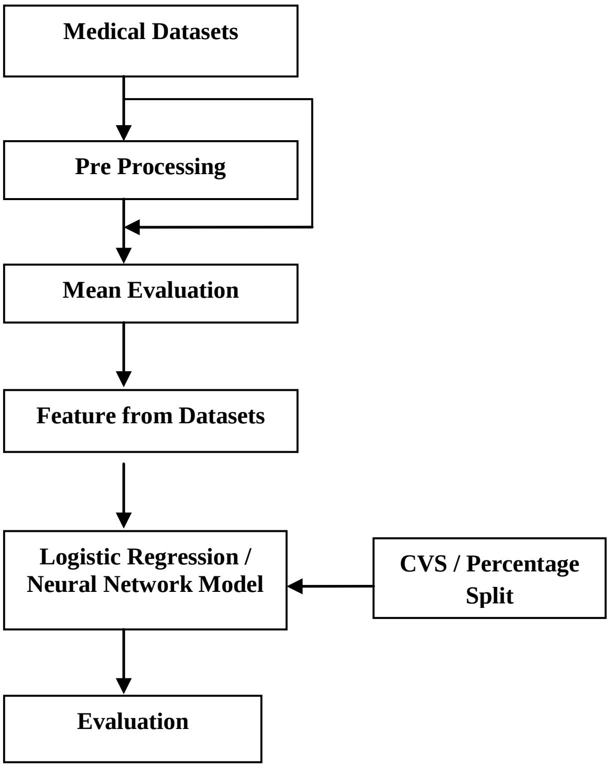

Figure 1: Hybrid Framework for ANN and LR using CVS and percentage split

4. RESULTS AND DISCUSSION

The proposed work uses mean evaluation method to compare the performance of LR and ANN model with feature selection methods using cross validation sample and percentage as test options. The specifications of the datasets are as shown in table 1.

Table 1. Specification of medical datasets

Sl No.	Medical Datasets	No. of Instances	Total Number of Attributes	No. of Classes
1	Diabetes	768	9	2
2	Liver Disorders	345	7	2
3	SPECTF	269	45	2

The performance of LR and ANN using CVS and percentage split is calculated for the entire dataset and subset of attributes obtained from feature selection methods.

Table 2 gives the result of evaluation for LR and ANN model with feature selection methods using CVS and percentage split based on mean evaluation for SPECTF Dataset for all attributes.

.Table 2. Evaluation for SPECTF datasets for full attributes

LR (Full Datasets)				NN (Full Datasets)			
CVS	% Split			CVS	% Split		
	66%	70%	75%		66%	70%	75%
80.29	80.21	82.71	80.59	78.43	73.62	79.01	80.59

Table 3 gives the result of evaluation for SPECTF Dataset for 20 attributes.

Table 3. Evaluation for SPECTF dataset for 20 attributes

LR (20 Attributes)				NN (20 Attributes)			
CVS	% Split			CVS	% Split		
	66%	70%	75%		66%	70%	75%
82.52	75.82	80.24	79.1	79.55	78.02	75.3	77.61

Table 4 gives the result of evaluation for SPECTF Dataset for 25 attributes.

Table 4. Evaluation for SPECTF dataset for 25 attributes

LR (25 Attributes)				NN (25 Attributes)			
CVS	% Split			CVS	% Split		
	66%	70%	75%		66%	70%	75%
80.66	81.31	80.24	80.59	81.78	80.21	77.77	82.08

Table 5 gives the result of evaluation for SPECTF Dataset for 30 attributes.

Table 5. Evaluation for SPECTF dataset for 30 attributes

LR (30 Attributes)				NN (30 Attributes)			
CVS	% Split			CVS	% Split		
	66%	70%	75%		66%	70%	75%
81.04	80.21	83.95	79.1	78.81	74.72	79.01	77.61

Table 6 gives the result of evaluation for Diabetes Dataset for full set of attributes.

Table 6. Evaluation for Diabetes dataset for full attributes

LR (Full Dataset)				NN (Full Dataset)			
CVS	% Split			CVS	% Split		
	66%	70%	75%		66%	70%	75%
77.21	80.08	80.43	81.77	75.39	74.32	75.21	77.08

Table 7 gives the result of evaluation for Diabetes Dataset for 4 attributes.

Table 7. Evaluation for Diabetes dataset for 4 attributes

LR (4 Attributes)				NN (4 Attributes)			
CVS	% Split			CVS	% Split		
	66%	70%	75%		66%	70%	75%
73.96	77.78	77.82	77.60	75.13	75.1	77.39	79.16

Table 8 gives the result of evaluation for Diabetes Dataset for 5 attributes.

Table 8. Evaluation for Diabetes dataset for 5 attributes

LR (5 Attributes)				NN (5 Attributes)			
CVS	% Split			CVS	% Split		
	66%	70%	75%		66%	70%	75%
76.95	80.08	80	80.20	76.56	74.71	73.48	79.16

Table 9 gives the result of evaluation for Diabetes Dataset for 6 attributes.

Table 9. Evaluation for Diabetes dataset for 6 attributes

LR (6 Attributes)				NN (6 Attributes)			
CVS	% Split			CVS	% Split		
	66%	70%	75%		66%	70%	75%
76.69	80.46	80.43	79.69	75.65	71.64	74.78	77.60

Table 10 gives the result of evaluation for Liver Disorder Dataset for full set of attributes.

Table 10. Evaluation for Liver Disorder dataset for full attributes

LR (Full Dataset)				NN (Full Dataset)			
CVS	% Split			CVS	% Split		
	66%	70%	75%		66%	70%	75%
68.11	69.23	69.9	68.6	71.59	65.81	64.07	67.44

Table 11 gives the result of evaluation for Liver Disorder Dataset 4 attributes.

Table 11. Evaluation for Liver Disorder dataset for 4 attributes

LR (4 Attributes)				NN (4 Attributes)			
CVS	% Split			CVS	% Split		
	66%	70%	75%		66%	70%	75%
67.82	66.66	68.93	70.93	69.85	73.5	74.75	72.09

Figure 2 through to Figure 11 shows the classification accuracy details after evaluation process.

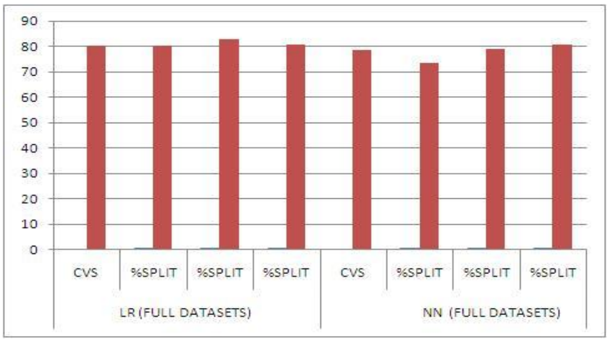

Figure 2: Classification accuracy for SPECTF dataset for full set of attributes

Figure 2 gives the graphical representation of result of evaluation for Table 2.

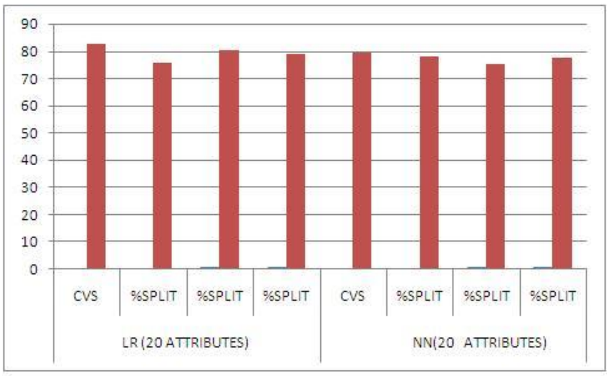

Figure 3: Classification accuracy for SPECTF dataset for 20 attributes

Figure 3 gives the graphical representation of result of evaluation for Table 3.

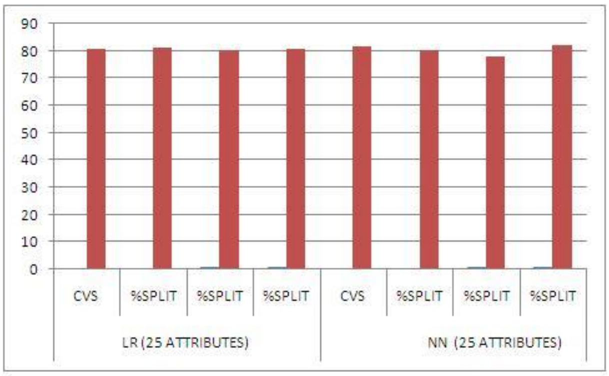

Figure 4: Classification accuracy for SPECTF dataset for 25 attributes

Figure 4 gives the graphical representation of result of evaluation for Table 4.

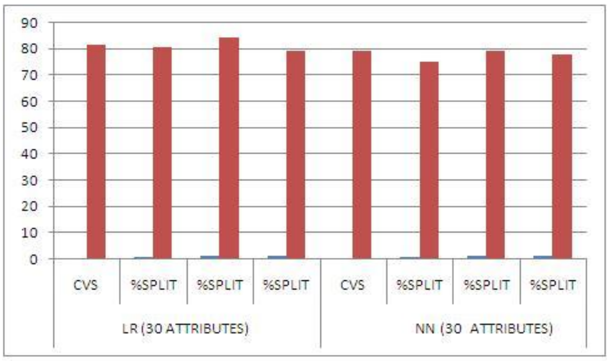

Figure 5: Classification accuracy for SPECTF dataset for 30 attributes

Figure 5 gives the graphical representation of result of evaluation for Table 5.

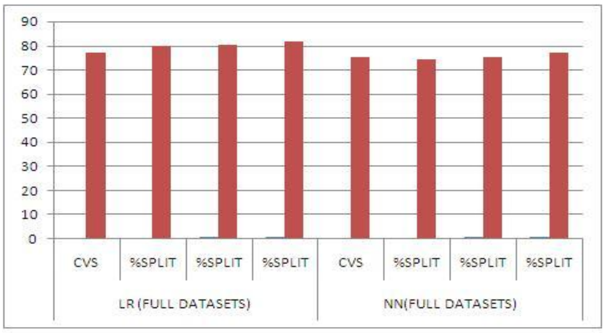

Figure 6: Classification accuracy for Diabetes Dataset for full attributes

Figure 6 gives the graphical representation of result of evaluation for Table 6.

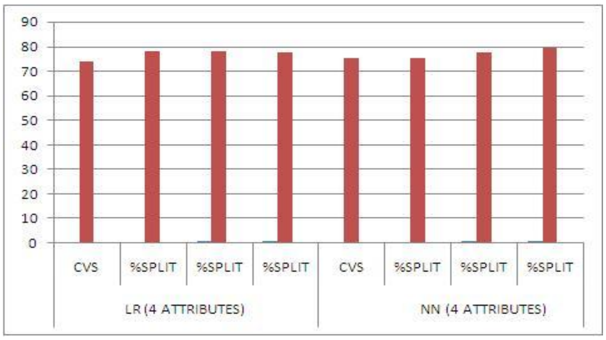

Figure 7: Classification accuracy for Diabetes Dataset for 4 attributes

Figure 7 gives the graphical representation of result of evaluation for Table 7.

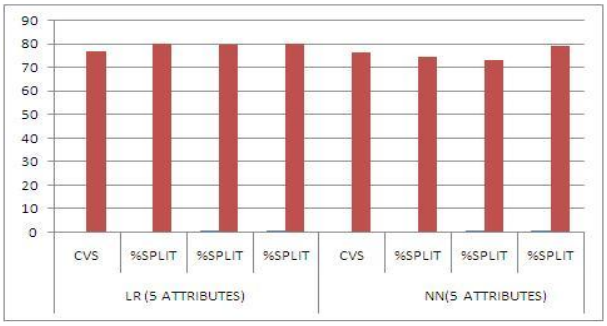

Figure 8: Classification accuracy for Diabetes Dataset for 5 attributes

Figure 8 gives the graphical representation of result of evaluation for Table 8.

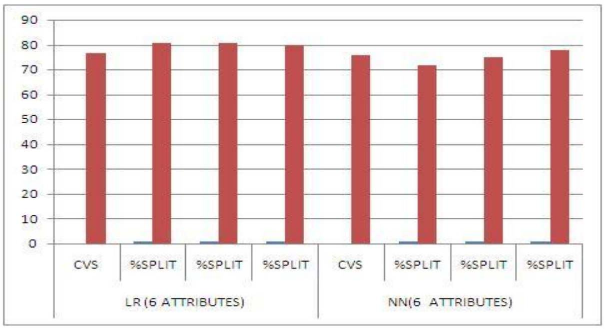

Figure 9: Classification accuracy for Diabetes Dataset for 6 attributes

Figure 9 gives the graphical representation of result of evaluation for Table 9.

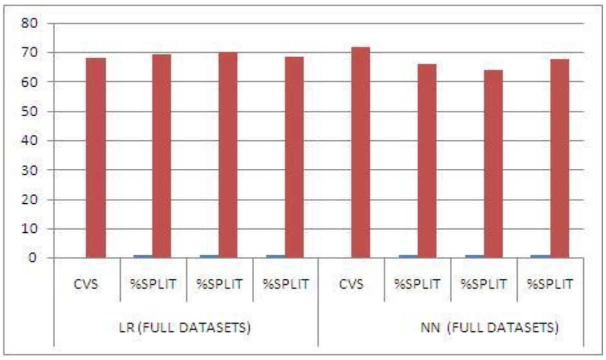

Figure 10: Classification accuracy for Liver Disorders Dataset for full attributes

Figure 10 gives the graphical representation of result of evaluation for Table 10.

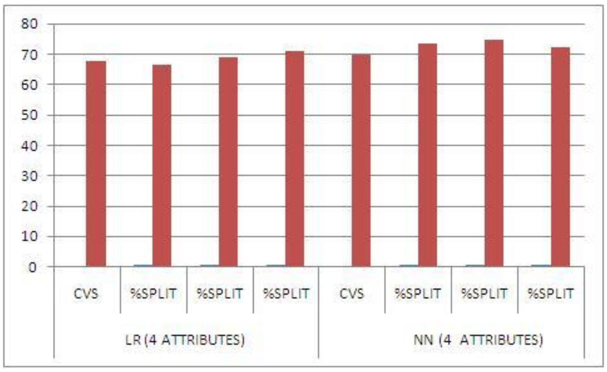

Figure 11: Classification accuracy for Liver Disorders Dataset for 4 attributes

Figure 11 gives the graphical representation of result of evaluation for Table 11.

From the Table 2 through Table 5 we note that for SPECTF Dataset:

i. For full set of attributes we get a maximum prediction accuracy of 82.71% for LR using Percentage Split.
ii. For 20 attributes we get a maximum prediction accuracy of 82.52% for LR using CVS.
iii. For 25 attributes we get a maximum prediction accuracy of 82.08% for ANN using Percentage Split.
iv. For 30 attributes we get a maximum prediction accuracy of 83.95% for LR using Percentage Split.

From Table 6 through Table 9 we note that for Diabetes Dataset:

i. For full set of attributes we get a maximum prediction accuracy of 81.77% for LR using Percentage Split.
ii. For 4 attributes we get a maximum prediction accuracy of 79.16% for ANN using Percentage Split.
iii. For 5 attributes we get a maximum prediction accuracy of 80.20% for LR using Percentage Split.
iv. For 6 attributes we get a maximum prediction accuracy of 80.46% for LR using Percentage Split.

From Table 10 and Table 11 we note that for Liver Disorder dataset:

i. For full set of attributes we get a maximum prediction accuracy of 71.59% for ANN using CVS.

ii. For 4 attributes we get a maximum prediction accuracy of 74.75% for ANN using Percentage Split.

5. CONCLUSION

The proposed research work uses mean evaluation method on feature selection methods like forward selection and backward elimination on publicly available medical datasets. LR and ANN are applied on feature selection methods using Cross Validation Sample and Percentage Split as test options. From the experimental results it is identified that for SPECTF dataset LR using percentage split a prediction accuracy of 83.95% is achieved, for Diabetes dataset LR using percentage split a prediction accuracy of 80.46% is achieved, and for Liver Disorder dataset NN using percentage split a prediction accuracy of 74.75% is achieved. For all datasets used in the research work gives better classification accuracy with reduced subset of features. From the experimental results it is observed that the reduced subsets of attributes gives more efficient results than that obtained by using full set of attributes.

Further, the results of the hybrid method can be improved by using threshold method with feature selection methods like forward selection and backward elimination.

6. REFERENCES

[1] Sunita Soni, Ujma Ansari, Dipesh Sharma and Jyoti Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer application (0975-8887), vol. 17, no.8, March (2011).

[2] Raghavendra B.K., Jay 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. II, Issue: I, January-March 2011), pp. 288-298.

[3] Raghavendra B.K., S.K. Srivatsa, Raghavendra S, Shivashankar S.K., "Comparison of Logistic Regression and Neural Network Model with and without hidden Layers", Universal Journal of Applied Computer Science and Technology, Vol.1, 2011, pp. 49-53.

[4] Qi Cheng, Pramod K. Varshney, and Manoj K. Arora, Logistic Regression for Feature Selection and Soft Classification of Remote Sensing Data", Geoscience and Remote Sensing Letters, IEEE, Vol. 3, No. 4, pp. 491-494.

[5] Raghavendra B.K., Jay B. Simha, "Evaluation of Logistic Regression Model with Feature Selection on Medical Dataset", International Journal of Computational Intelligence (Vol.1, Issue 2, Dec 2010), pp. 35-42.

[6] Qinbao Song, Jingjje Ni and Guangtao Wang, "A Fast Clustering Based Feature Subset Selection Algorithm for High Dimensional Data", IEEE Transactions on Knowledge and data engineering 2013, Vol 25, Issue 1, pp 1-14.

[7] Amin S.U., Agarwal, K. and Beg, R.," Genetic neural network based data mining in prediction of heart disease using risk factors", IEEE Conference on Information & Communication Technologies (ICT), Page(s):1227–1231, 2013.

[8] Kumari, Sonu Singh and Archana, "A data mining approach for the diagnosis of diabetes mellitus", 7th International Conference on Intelligent Systems and Control (ISCO), Page(s): 373 – 375, January 2013.

[9] Rachata N., Charoenkwan P., Yooyativong T. Chamnongthal K., Lursinsap C. and Higuchi, K." Automatic Prediction System of Dengue Haemorrhagic Fever Outbreak Risk by Using Entropy and Artificial Neural Network", International Symposium on Communications and Information Technologies (ISCIT), Page(s): 210 – 214, October 2008.

[10] Sandya Joshi, Deepa Shenoy, Vibhudendra Simha G.G., P. L. Rashmi, and K. R. Venugopal, "Classification of Alzheimer's Disease and Parkinson's Disease by Using Machine Learning and Neural Network Method", Second International Conference on Machine Learning and Computing, page(s): 218- 222, 2010.

[11] T. John Peter, and K. Somasundaram, "An Empirical Study on Prediction of Heart Disease Using Classification Data Mining Techniques", IEEE International Conference On Advances In Engineering.

[12] R.Robu and C. Hora", "Medical Data Mining with Extended WEKA", IEEE International Conference on Intelligent Engineering System (INES 2012), June 2012, page(s): 347-350.

[13] Ankita Dewan and Meghna Sharma, "Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification", 2015, page(s): 704-706.

[14] Sikander Singh Khuri, and Gurpreet Singh, "Ranking Early Signs of Coronary Heart Disease Among Indian Patients", 2015, page(s): 840-844.

[15] Sana Shaikh, Amit sawant, Shreerang Paradkar, Kedar Patil, "Electronic Recording System-Heart Disease Prediction System", International Conference on Technologies for Sustainable Development (ICTSD 2015), February 2015.

[16] Raghavendra S, and 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", Second International Conference on Emerging Research in Computing, Information, Communication and Applications (ERCICA- 2014), August-2014.