International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 1 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT183818

Liver Cancer Detection

L. S. Rohith Anand*, B. Shannmuka, R. Uday Chowdary, K. Satya Sai Krishna
CSE, Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India

ABSTRACT

Machine learning techniques play an important role in building predictive models by learning from Electronic Health Records (EHR). Predictive models building from Electronic Health Records still remains as a challenge as the clinical healthcare data is complex in nature and analysing such data is a difficult task. This paper proposes prediction models built using random forest ensemble by using three different classifiers viz. J48, C4.5 and Naïve Bayes classifiers. The proposed random forest ensemble was used for classifying four stages of liver cancer. Using a feature selection method the reliable features are identified and this subset serves as input for the ensemble of classifiers. Further a majority voting mechanism is used to predict the class labels of the liver cancer data. Experiments were conducted by varying the number of decision trees generated using the J48, C4.5 and Naïve Bayes classifiers and compared with the classification made using decision stump and Adaboost algorithms.

Keywords : Ensemble, Feature Selection, C4.5, J48 and Random Forest

I. INTRODUCTION

In health care industry, patient's medical data size grows day to day. The process of applying computer based information system (CBIS), including new techniques, for discovering knowledge from data is called data mining. The process of machine learning is similar to that of data mining. Machine learning algorithms may be distinguished by either supervised or unsupervised learning methods. Supervised learning methods are widely used for predictive modelling. Predictive modelling is a branch of clinical and business intelligence branch which is used for health risk classification and also to predict the future health status of the individuals. Electronic health records (EHR) are used to store large scale information of patient conditions, treatments etc. The EHR information may be structured or unstructured. Using controlled vocabulary, electronic health records are maintained in structured data format for documenting patient information than narrative text which is unstructured in nature. EHR helps to streamline the clinical workflow information. Ensemble learning is a well-known approach used in machine learning for prediction by combining various ensemble models [1]. Ensemble of classifiers is aggregations of multiple classifiers are J48, C4.5 and Naive Bayes etc. [2]. Ensembles aim for better performance than any of the base classifiers. The proposed work aims to improve the accuracy of healthcare data for prediction and classification, by building a hybrid predictive classifier model using ensemble of classifiers [3][4].

The remaining part of the paper described in the following section.

Section 2 describes related works on classification, feature selection, subset generation, pre- processing and boosting algorithms such as adaptive boosting for electronic health records. Section 3 explains the overall architecture of proposed system. Section 4 reports the experimental results. Section 5 concludes the paper.

II. RELATED WORK

This section discusses the existing methods for pre-processing, feature extraction, boosting methods such as adaptive boosting. Aydin et. al. (2009) investigated the various factors involved on ensemble construction using a wide variety of learning algorithms, data sets and evaluation criteria [5]. They have provided the idea of subset selection to the level of discriminating whether the discrimination is applicable or not at the level of classifier.Ping Li et. al. (2013) surveyed about supervisedmulti-label classification and proposed variable pairwise constraint projection for mutli-label ensemble. They have adopted boosting methods to construct a multi-label ensemble to increase the generalization ability [6].

Jia Zhua et. al. (2015) employed multiple classifier systems (MCS) to improve the accuracy of disease detection for Type-2 Diabetes Mellitus. Multi classifier system performs worse when design is not proper [7]. They have proposed a dynamic weighted voting scheme for multiple classifier decision combination. Yan Li et. al. (2015) stated data mining framework for distributed healthcare information based on privacy preserving constraints [8]. Neesha Jothi et. al. (2015) surveyed the data mining techniques and has classified the articles have suggested data mining plays important role in medical diagnosing for predicting diseases [9].

Table 1. Comparative analysis of different ensembles of classifiers

Author	Methods Used	Data Sets Used	Number of Iterations	Performance Metrics
Nikunj C. Oza et.al.method (2008)[10]	K nearest neighbor, Learning vector quantization, Multi-layer perceptron's, Radial basis functions, Support vector machines	Datasets from UCI repository	110	Average, Geometric mean
Yong Seog Kim et.al.method (2009)[11]	Naive Bayes, Support vector machines ,Artificial neural networks, Pruned tree classifier	German credit data and COII 2000 competition data	120	AUC, Accuracy, False positive rate, Hit rate gain
Hesam Sagha et.al.method (2013)[12]	Quadrant discriminant analysis	2 real datasets containing data from body mounted inertial sensors	45	Entropy, Mutual Information
Ping Li et.al.method (2013) [6]	Bagging, Boosting, Random Forest, Random subspace, Rotation forest	12 datasets including test categorization, image	20	Hamming loss, Ranking loss, One error, Coverage, Average

		classification and bioinformatics		Precision, F1-metrics, Recall
Ritaban Dutta et.al.method (2015) [13]	Binary tree, Linear discriminant analysis classifier, Naïve Bayes Classifier, K-nearest neighbor, Adaptive Neuro Fuzzy Inference classifier	24 Holstein-Friesian cows from Tasmanian Institute of Agriculture Dairy Research Facility	20	AUC, Accuracy
Yang Zhang et.al.method (2015)[14]	SVM classifier, BPNN classifier	Benchmark datasets from UCI repository	11	Average regression
Bing Gong et.al.method (2016)[15]	Artificial neural network, Support vector machines, CART	Datasets from UCI repository	30	F measure, G mean
Yan Li et.al.method (2016)[8]	Adaboost algorithm	9948 real world EHRs of diabetes patients	20	F-measure, Sensitivity, Precision
Cátia M. Salgado et.al.method (2016)[16]	Apriori decision, Aposteriori decision	Benchmark datasets from UCI collection, MIMIC II datasets	12	AUC, Accuracy, Sensitivity, Specificity

Based on the literature survey carried out a comparative analysis of the ensemble of classification methods, the data sets used for experiments by different researchers, the number of iterations for which the experiments were conducted and the metrics used for measuring the classification accuracy are tabulated in the table given below.

From the above table the conclusion drawn is an ensemble of C4.5, J48 and Naïve Bayes classifier with majority voting scheme was not studied and hence this work focusses on building a predictive model based on building a random forest using these three classifiers. The proposed system has been compared with the existing decision stump and Adaboost algorithms. The next section discusses about the proposed system and how limitations in existing system is resolved.

III. PROPOSED WORK

The proposed architecture is shown in Figure 1. The Electronic health records contain features like patient id, status, age, sex, hepato, ascites, edema, billi, cholestrol, albumin etc. The data considered have to be clinically transformed i.e. to make it suitable for further processing. The clinical transformation step is also identified as preprocessing step.

The unprocessed has null values, irrelevant values and noisy values. These data errors would lead to misclassification and hence need to be clinically transformed. The missing data in the considered dataset is imputed with values computed using mode function.

After pre-processing of data, for classifying instances under Random forest, three subsets from the datasets are generated. The subset will be generated considering three features like platelet count, alkaline phosphate and cholesterol values.

The random forests are built using three classification algorithms namely C4.5, J48 and Naïve Bayes. There are many voting mechanisms followed for ensemble of classifiers, here we are using majority vote method to perform voting with different classifiers. Here the output will be the final outcome of the majority of classifiers

Figure 1 shows the architecture of proposed system.

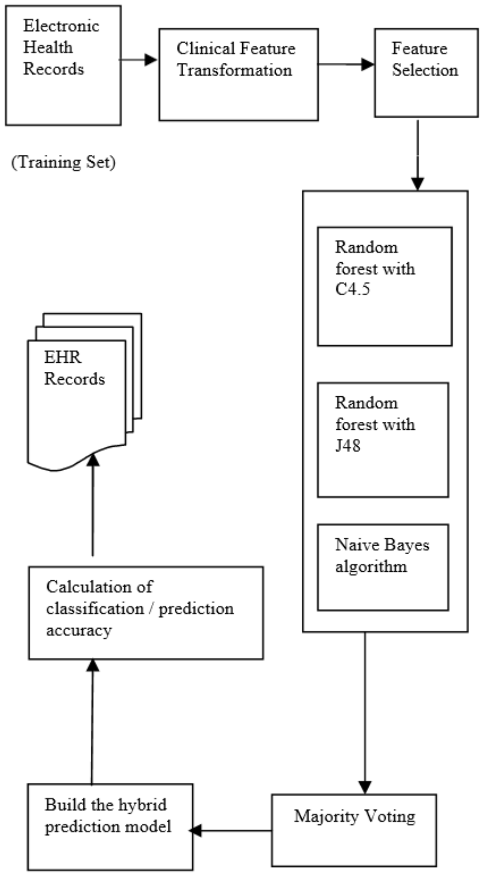

Figure 1. Architecture of Proposed System

The proposed system with its role and advantages is discussed. The experimental result analysis of the proposed work has been discussed in the next section.

IV. RESULTS AND DISCUSSION

Experiment Results

The proposed system is implemented using Java and Weka tool. The liver cancer dataset having 500 instances and breast cancer dataset are used for the experiments. The ensemble of classifiers is used for classifying these datasets on which voting is performed.

Pre-processing

In this module, pre-processing of data is done. The dataset which we have contains null values, irrelevant values and noisy values. As missing values in the dataset lead to misprediction of the final result, the dataset is pre-processed by filling missed values on basis of mode function. The dataset without pre-processing which contains irrelevant values is shown below in figure 2.

sex	ascites	hepato	spiders	edema	bili	chol	albumin	copper	alk.phos	ast	trig	platelet	protime	stage
f	1	*	*	*	14.5	261	2.6	156	1718	137.95	172	190	12.2	Stage4
f	0	1	1	0	1.1	302	4.14	54	7394.8	113.52	88	221	10.6	Stage3
m	0	0	0	0.5	1.4	176	3.48	210	516	96.1	55	151	12	Stage4
f	0	1	1	0.5	1.8	244	2.54	64	6121.8	60.63	92	183	10.3	Stage4
f	0	1	1	0	3.4	279	3.53	143	671	113.15	72	136	10.9	Stage3
f	0	1	0	0	0.8	248	3.98	50	944	93	63	0	11	Stage3
f	0	1	0	0	1	322	4.09	52	824	60.45	213	204	9.7	Stage3
f	0	0	0	0	0.3	280	4	52	4651.2	28.38	189	373	11	Stage3
f	0	0	1	0	3.2	562	3.08	79	2276	144.15	88	251	11	Stage2
f	1	0	1	1	12.6	200	2.74	140	918	147.25	143	302	11.5	Stage4
f	0	1	1	0	1.4	259	4.16	46	1104	79.05	79	258	12	Stage4
f	0	0	1	0	3.6	236	3.52	94	591	82.15	95	71	13.6	Stage4
f	0	0	0	0	0.7	281	3.85	40	1181	88.35	130	244	10.6	Stage3
m	1	1	0	1	0.8	0	2.27	43	728	71	0	156	11	Stage4
f	0	0	0	0	0.8	231	3.87	173	9009.8	127.71	96	295	11	Stage3
f	0	0	0	0	0.7	204	3.66	28	685	72.85	58	198	10.8	Stage3
f	0	1	0	0	2.7	274	3.15	159	1533	117.8	128	224	10.5	Stage4
f	0	1	1	1	11.4	178	2.8	588	961	280.55	200	283	12.4	Stage4
f	0	1	0	0.5	0.7	235	3.56	39	1881	93	123	209	11	Stage3
f	0	1	0	0	5.1	374	3.51	140	1919	122.45	135	322	13	Stage4
m	0	1	1	0	0.6	252	3.83	41	843	65.1	83	336	11.4	Stage4
f	0	0	1	0	3.4	271	3.63	464	1376	120.9	55	173	11.6	Stage4
f	1	1	1	1	17.4	395	2.94	558	6064.8	227.04	191	214	11.7	Stage4
m	0	1	0	0	2.1	456	4	124	5719	221.88	230	70	9.9	Stage2

Figure 2. Dataset before preprocessing

The dataset is preprocessed to fill missing and irrelevant values as shown in figure 3

Figure 3. Dataset after preprocessing

Feature selection

Feature selection method used for model construction by choosing a subset of relevant predictors. It also called as variable selection or attribute selection [17].

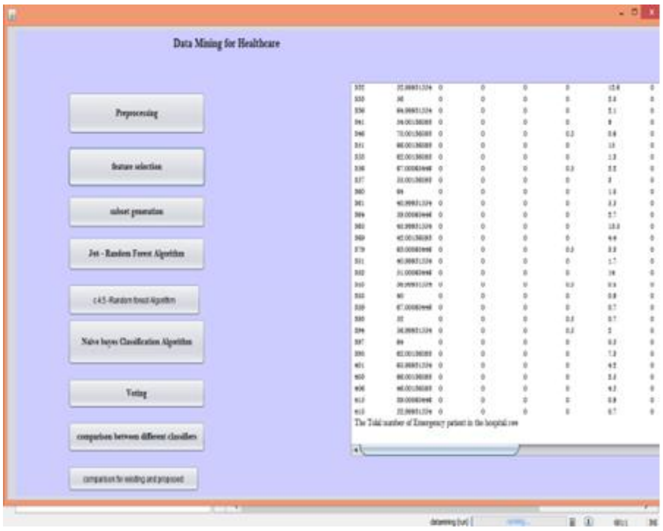

Figure 4. Dataset after feature selection

Subset Generation

For classifying instances under Random forest, generate three subsets from the datasets. Subset will be generated considering some features as first, middle and last stages as shown in figure 5.

Figure 5. Results achieved after implementing subset generation

Performance Evaluation

This section evaluates the performance of J48 Random forest classifier, C4.5 Random Forest classifier and Naïve Bayes classifier using Accuracy, True positive, False positive, Precision, Recall and F-measure. Figure 6 shows the performance of J48 Random Forest classifier for prediction different stages of liver cancer. Figure 7 show the performance of C4.5 Random forest classifier for prediction different stages of liver cancer and Figure 8 shows the performance of Naïve Bayes classifier for prediction different stages of liver cancer.

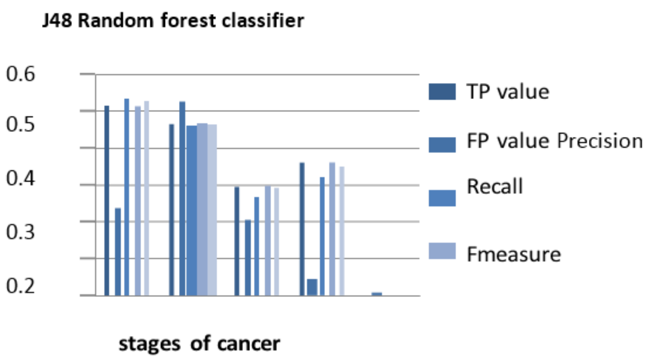

Figure 6. Comparison of J48 Random forest classifier at different stages

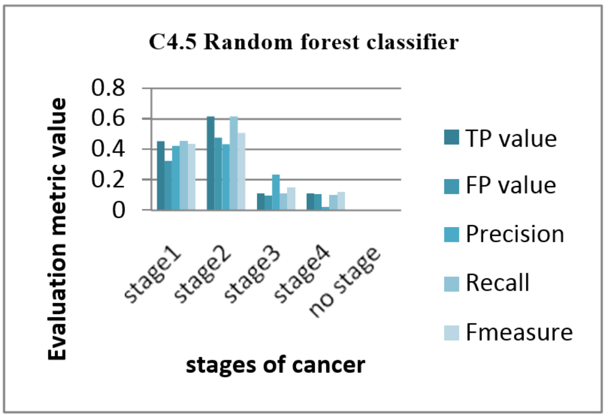

Figure 7. Comparison of C4.5 Random forest classifier at different stages

Table 2. Accuracy of Existing and Proposed system based on different threshold values

System	Weighted Threshold value 100 in %	Weighted Threshold value 200 in %	Weighted Threshold value 500 in %	Weighted Threshold value 1000 in %
Existing	46	48	46	44
Proposed	51	53	50	48

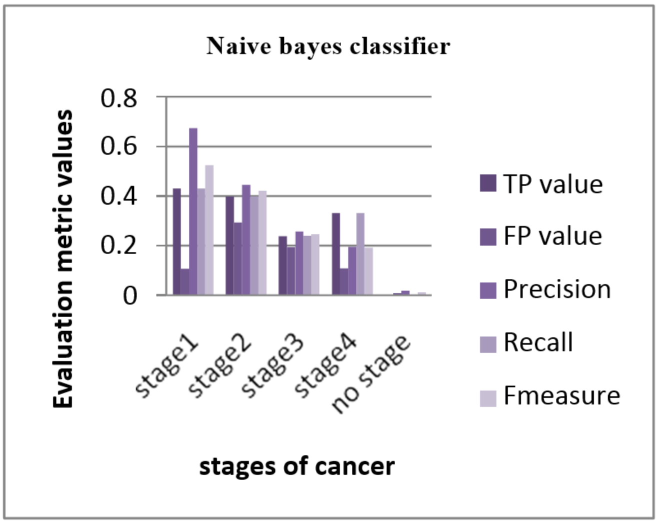

Figure 8. Comparison of Naïve Bayes classifier at different stages

Comparison of Existing and Proposed System

The comparative analysis of classification accuracy for Liver and Breast cancer dataset shown in figure 9 and also existing and proposed system based on different threshold values shown in Table 2 and figure 10.

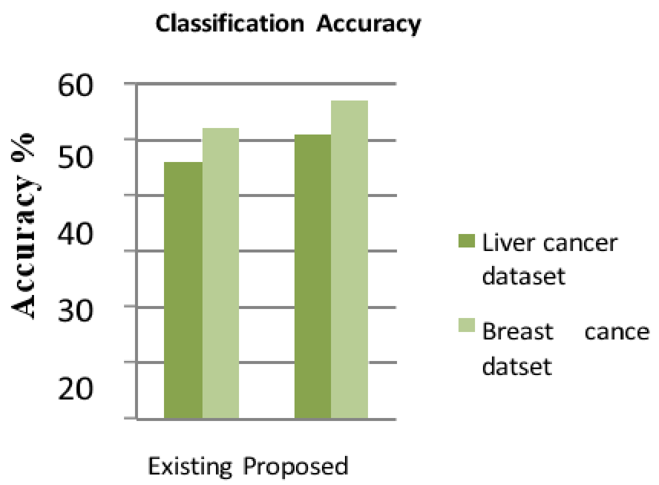

Figure 9. Comparison of existing and proposed system for two different datasets

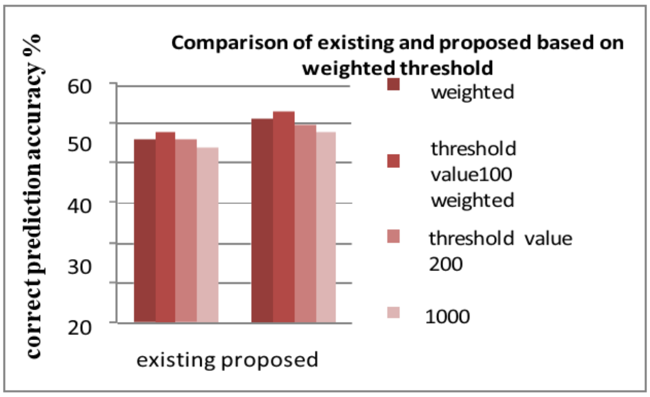

Figure 10. Comparison of existing and proposed system based on weighted threshold

In this section, we had a brief discussion about implementation and experimental results of Clinical feature transformation, Feature selection, Subset generation, J48 classifier, C4.5 classifier, Naive bayes classifier and Majority voting.

V. CONCLUSION

The prediction model is built by series of steps such as clinical feature transformation, feature selection, ensemble of classifiers and Majority voting which aimed to improve rate of correct predictions. The prediction accuracy is improved by an ensemble of classifiers and when majority voting mechanism was applied on them. The proposed system here achieve an accuracy of 40% for C4.5 Random forest classifier, 43% for J48 Random forest classifier, 38% for Naïve bays classifier when tested for Liver cancer dataset which is more than existing system. The accuracy of 45% for C4.5 Random forest classifier, 52% for J48 Random forest classifier and 47% for Naïve Bayes classifier was achieved when tested for Breast cancer dataset which is more than existing system which has an accuracy of 46%. The number of trees generated was varied and the prediction accuracy of the proposed work was studied.

VI. REFERENCES

[1] Dietterich TG.(2000), Ensemble methods in machine learning. In: Proceedings of Multiple Classifier System, vol. 1857, Springer (2000), pp. 1–15.

[2] Zhi-Hua Zhou, Ensemble Methods: Foundations and Algorithms, Machine Learning & Pattern Recognition Series, 2012.

[3] Yongjun Piao, Minghao Piao, Keun Ho Ryu, Multiclass cancer classification using a feature subset- based ensemble from microRNA expression profiles, Computers in Biology and Medicine 80 (2017) 39–44.

[4] ManjeevanSeera, Chee Peng Lim, A hybrid intelligent system for medical data classification, Expert Systems with Applications 41 (2014) 2239– 2249.

[5] Aydın Ulas, Murat Semerci, Olcay Taner Yıldız, Ethem Alpaydın, Incremental construction of classifier and discriminant ensembles, Information Sciences 179 (2009) 1298–1318.

[6] Ping Li , Hong Li , Min Wu , Multi-label ensemble based on variable pairwise constraint projection, Information Sciences 222 (2013) 269–281.

[7] Jia Zhua,, Qing Xie, Kai Zheng,An improved early detection method of type-2diabetes mellitus using multiple classifier system, Information Sciences 292 (2015) 1–14.

[8] Yan Li, Changxin Bai, Chandan K. Reddy, A distributed ensemble approach for mining healthcare data under privacy constraints, Information Sciences (2015).

[9] Neesha Jothi, Nur'Aini Abdul Rashid, Wahidah Husain, Data Mining in Healthcare – A Review, Procedia Computer Science 72 (2015) 306 – 313.

[10] Nikunj C. Oza , Kagan Tumer, Classifier ensembles: Select real-world applications, Information Fusion 9 (2008) 4–20.

[11] YongSeog Kim, Boosting and measuring the performance of ensembles for a successful database marketing, Expert Systems with Applications 36 (2009) 2161–2176.

[12] Hesam Sagha, Hamidreza Bayati, José del R. Millán, Ricardo Chavarriaga, On-line anomaly detection and resilience in classifier ensembles, Pattern Recognition Letters (2013).

[13] Ritaban Duttaa, Daniel Smitha, Richard Rawnsley, Greg Bishop-Hurley, James Hills, Greg Timms, Dave Henry, Dynamic cattle behavioural classification using supervised ensemble classifiers, Computers and Electronics in Agriculture 111 (2015) 18–28.

[14] Yang Zhang, Li Zhang, M.A. Hossain, Adaptive 3D facial action intensity estimation.

[15] Bing Gong, Joaquín Ordieres-Mere, Prediction of daily maximum ozone

threshold exceedances by preprocessing and ensemble artificial intelligence techniques: Case study of Hong Kong, Environmental Modelling & Software 84 (2016) 290-303.

[16] Cátia M. Salgado, Susana M. Vieira, Luís F. Mendonça, Stan Finkelstein, João M.C. Sousa, Ensemble fuzzy models in personalized medicine: Application to vasopressors administration, Engineering Applications of Artificial Intelligence (2015).

[17] B. Seijo-Pardo, I. Porto-D'iaz, V. Bol'on-Canedo, A. Alonso-Betanzos, Ensemble Feature Selection: Homogeneous and Heterogeneous Approaches, Knowledge-Based Systems (2016).

[18] J.Prathyusha, G.Sandhya, V.Krishna Reddy," An Improvised Partition-Based Workflow Scheduling Algorithm", International Innovative Research Journal of Engineering and Technology, vol 02, no 04,pp.120-123,2017.

Cite this article as :

L. S. Rohith Anand, B. Shannmuka, R. Uday Chowdary, K. Satya Sai Krishna, "Liver Cancer Detection", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 1, pp. , January-February 2019.

Available at doi :

https://doi.org/10.32628/CSEIT183818

Journal URL : http://ijsrcseit.com/CSEIT183818