Classification and Feature Selection Approaches by Machine Learning Techniques: Hepatocellular Carcinoma (HCC) Prognosis Prediction

Satish Chandra Reddy Nandipati[1],Haziqah Shamsudin[2], Chew XinYing[3*]
[1,2,3]School of Computer Sciences, Universiti Sains Malaysia, Pulau Pinang, Malaysia
[1]satishnandipati@student.usm.my, [2]haziqahsham@gmail.com, [3]xinying@usm.my

Abstract
The hepatocellular carcinoma being one of the common liver cancer deaths worldwide, which is predominant in Asian and African countries. It is estimated that the survival rate can be increased with the early detection, thus it is important to predict the selected features to avoid this disease. The classification model prediction, and to distinguish the role of features selected using Rapid miner machine learning operators is the aim of this study. The accuracy of SVM algorithm is found to be highest with 81.81%, followed by random forest (79.67%) in classification model, whereas in feature selection both Naïve Bayes and SVM shows similar accuracies (74.90 and 74.10 %). The average accuracies of selected features in comparison to complete dataset, and the number of selected features are useful for the prediction of HCC data and to build a better model performance respectively. Whereby, it is expected that the model performance may drop further if the minimum features are not considered.

Keywords: Hepatocellular carcinoma, Classification, Feature Selection, Prediction, Rapid miner

1. Introduction

Liver cancer is one of the fastest growing cancers, second deadliest cancer and the sixth most diagnosed in the world, leading to the mortality rate of over 8.2 million deaths a year [1] Hepatocellular cells are the source for the start Hepatocellular carcinoma (HCC) which is also called as hepatoma and is the most common type of liver disease and accounts for nearly 75-85% of cases. The chronic liver disease leads to the development of cirrhosis (scarring of the liver) which accounts 80-90% of cases and remains to be the one of the most important risk factors for the development of HCC. The risk factors include infection with hepatitis B or C virus, non-alcoholic fatty liver disease (NAFLD), excess alcohol intake, smoking, type 2 diabetes, obesity and aflatoxin contaminated foodstuffs [2]. The prevalence of HCC in developed countries of the world is lower, whereas it is predominant in Asian and African countries which includes Southeast Asia, China, Mongolia, Western and Eastern Africa and Sub-Saharan respectively[3].The ratio of HCC occurrence is more in males and is said to be 2.4 times in worldwide distribution and the common age range is about 30-50 years [4]. It is estimated that the survival rate can be increased to 35%, if HCC is detected at early stage, thus indicating early detection of HCC is critical for improving disease prognosis. The some of the various epigenetic markers such as GSTP1 genes and genetic markers such as mutations of TP53 249 T etc., have been reported to detect the HCC in the patient's urine [1]. The multiple variables, histopathological images, CT (Computerized Tomography) images are used by researchers and clinicians to predict the HCC disease by using both machine learning algorithms and for selection of features [5], [6], [7], [8]. The main objectives of this study are to build a better classification model, and to distinguish the features selected for further prediction of the HCC disease by using Rapid miner version 9.2.

2. Related Work

The various machine learning classification methods has been used for the detection of various diseases such as heart, breast cancer, ovarian etc. [1]. The prediction of HCC dataset with the classification algorithms are reviewed below:

The five classification algorithms such as SVM, RF, J48, MLP and Bayesian Network and feature selection were used with Weka tool to evaluate the ILPD (Indian Liver Patient Dataset) obtained from UCI Repository. Among five classifications, the SVM showed highest accuracy of 71.35% before feature selection, whereas random forest showed 71.86% after feature selection [9]. Similar dataset was used to evaluate SVM and NB with MATLAB and found to achieve 79.66% and 61.28 % of accuracy respectively [10]. In another study, ILPD has been studied with respect to four classification algorithms which include SVM, Logistic Regression, KNN and Artificial Neural Network (ANN). Among all, highest accuracy of 98% is shown by ANN, followed by SVM (75.04%) [11]. Similar dataset has been considered for another study with particle swarm optimization algorithm (PSO) feature selection method. The included features are albumin, total bilirubin, direct bilirubin, total proteins, A/G ratio, SGOT, SGPT, alkphos. The evaluation was carried out by J48 (95.04%), Bayesnet classification (90.33%), Random Forest (80.22%), MLP (77.54%) and SVM (73.44%). It is noted that J48 and Bayesnet classification are better than other classification algorithms [12]. In another study, the under and over-sampling was performed to bring the balance nature of the ILPD, later both SVM and back propagation multi layered

neural network algorithm (MLP) has been applied. The accuracies of the SVM and back propagation MLP are 71 % and 73.2% respectively [13].

The analysis of classification algorithms for liver disease diagnosis from two datasets (UCLA and AP) is carried out by Naive Bayes, KStar, Logistic and REP tree, bagging using Weka 3.6.10. The results showed the highest accuracy is achieved by KStar (100%), followed by bagging (88%). [14].

The below three studies were conducted on HCC dataset obtained from UCI repository. During the preprocessing stage the missing values were imputed by using HEOM distance and K-means clustering was applied, later synthetic minority over-sampling technique (SMOTE) method was applied to obtain the balanced dataset. Both neural networks and logistic regression algorithms were applied on the balanced dataset. The accuracy of the neural networks with respect to without cluster and with cluster is in the range of 68.7%–75.2, similar for logistic regression it is found to be in the range of 70%–73% respectively [6]. In another study, k-NN imputation method is applied to deal with missing data, five classifications such as Decision tree, random forest, logistic regression, bagging and boosting were studied on HCC dataset with Python with scikit-learn library, among these random forest showed the highest accuracy (74%), and followed by 72% of boosting [15]. In another study, during preprocessing a Markov Blanket-based clustering method was applied, where the redundancy among the features is computed based on the ranking. A total of six different classifiers were used to study the evaluation of the proposed method, and SVM showed the accuracy of 76.25%, followed by Naïve Bayes (73.95%) and KNN of 72.10% [16]. Similar HCC dataset with 165 patients has been considered in another study, where ten machine learning algorithms are used with Python language along with Pandas, Deep, and Sklearn libraries. Normalization approach is used in the preprocessing step. Initially, for parameter optimization the genetic algorithm coupled with stratified 5-fold cross-validation (twice), and then feature selection was applied. The results showed that 2 level genetic optimizer and feature selection with SVM (type C-SVC) achieved highest accuracy (88.49%) and F1-Score (87.62%) respectively [8].

3. Data Sources

The actual dataset is from UCI machine learning repository webpage [17]. The dataset consists of 165 patients diagnosed with Hepatocellular Carcinoma (HCC), and includes both missing values and imbalance nature of class label. In this study, the Hepatocellular Carcinoma complete balanced dataset is used, in this dataset the missing values were imputed by KNN (K=1) using HEOM distance, and the balance nature of the dataset (205 rows from 167) is through SMOTE (k = 3) with oversampling method. This dataset is collected from available source [18].The balanced dataset consists of 50 attributes with 204 instances. Out of 204 cases, 120 cases labeled as “lives (No), 63 as “dies (Yes). Table 1 shows the dataset characteristics.

Table 1: Characteristic of Hepatocellular carcinoma (HCC) Datasets

Datasets	**Attributes**	**Instances**	**Missing Values**
UCI- HCC Survival Data Set	50	165	Yes
HCC balanced dataset	50	204	No

4. Attributes Description

The dataset consists a total of 50 attributes which includes 26 qualitative variables + 23 quantitative variables (referred as predictable attribute or input attributes), one as a class label [“lives (No), “dies (Yes)]. They are categorized in Nominal, Continuous, Ordinal and Integer. The attribute descriptions are shown in Table 2.

Table 2: Description of attributes and their codes

Description: Nominal	**Code**	**Description: Continuous**	**Code**
Gender	Gender	Grams of Alcohol per day	Grams_day
Symptoms	Symptoms	Packs of cigarets per year	Packs_year
Alcohol	Alcohol	International Normalized Ratio	INR
Hepatitis B Surface Antigen	HBsAg	Alpha-Fetoprotein (ng/mL)	AFP
Hepatitis B e Antigen	HBeAg	Haemoglobin (g/dL)	
Hepatitis B Core Antibody	HBcAb	Mean Corpuscular Volume (fl)	MCV
Hepatitis C Virus Antibody	HCVAb	Leukocytes(G/L)	
Cirrhosis	Cirrhosis	Platelets (G/L)	
Endemic Countries	Endemic	Albumin (mg/dL)	
Smoking	Smoking	Total Bilirubin(mg/Dl)	Total Bil
Diabetes	Diabetes	Alanine transaminase (U/L)	ALT
Obesity	Obesity	Aspartate transaminase (U/L),	AST
Hemochromatosis	Hemochro	Gamma glutamyl transferase (U/L)	GGT

Arterial Hypertension	AHT	Alkaline phosphatase (U/L)	ALP
Chronic Renal Insufficiency	CRI	Total Proteins (g/Dl)	TP
Human Immunodeficiency Virus	HIV	Number of Nodules	Nodule
Nonalcoholic Steatohepatitis	NASH	Creatinine (mg/dL)	
Esophageal Varices	Varices	Major dimension of nodule (cm)	Major_Dim
Splenomegaly	Spleno	Direct Bilirubin (mg/dL)	Dir_Bil
Portal Hypertension	PHT	Iron (mcg/dL)	
Portal Vein Thrombosis	PVT	Oxygen Saturation (%)	Sat
Liver Metastasis	Metastasis	Ferritin (ng/mL)	
Radiological Hallmark	Hallmark	**Description: Ordinal**	
1= Survives, 0 = Died	Class	Performance Status	PS
Description: Integer		Encefalopathy degree	Encefalopathy
Age at diagnosis	Age	Ascites degree	Ascites

5. Methodology

5.1 Data Preprocessing

The missing values are not present in the HCC balanced dataset (Table 1). However, the presence of different measuring units in the dataset has been noticed and need to rescale (i.e., the variable values between 0 and 1) using normalization. Thus, the normalization method which is included in Rapid miner machine learning techniques is used during the model building.

5.2 Data Analysis

The pre-processing steps (normalization, percentage split of 70–30% as training-testing) and classification model building is carried out in Rapid miner studio version 9.2. The Rapid miner with a total of seven machine learning operators such as K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB) and Auto Multilayer Perceptron (AutoMLP, for Neural network), and two ensemble classifiers such as Bagging (method = decision tree) and Adaboost (method = decision tree) with 10-fold cross validation and with default values parameter settings available in Rapid miner were evaluated on the training and testing data. The two feature selection methods included in this study are forward election (method = Naïve Bayes) and backward elimination (method = Naïve Bayes and decision tree). The accuracy, precision and recall are used to check the performance of a model on the test data.

6. Results

6.1 Performance on 49 features (attributes)

The balanced dataset used in this study, has been studied with respect to logistic regression and neural networks algorithms, this approach shows better detection of HCC (Santos et al. 2015). However, we have not come across the evaluation of algorithms such as bagging along with 6 mentioned ML algorithms using Rapid miner. In this study we applied Rapid miner ML operators to understand the better performance from the selected classification models with complete and feature selected datasets. The accuracy has been taken into consideration for the performance measures of each classification operator. Initially, the 49 attributes were used to find the performance of seven ML operators. The highest accuracy has been observed in SVM (81.81%), followed by random forest (79.67%), and average accuracies (75.19%) respectively (Table 3).

Table 3: The performance comparisons of different classification operators with 49 features

Algorithms	Accuracy	Precision	Recall
KNN	74.24	79.37	67.57
SVM	81.81	79.27	87.84
RF	79.67	78.48	83.78
NB	72.41	75	68.92
Auto MLP	76.81	77.33	78.38
Bagging	63.86	63.41	70.27
AdaBoost	77.57	77.63	79.73
Average	75.19	75.78	76.64

6.2 Performance on 7 selected features

To imporve the model accrucay the feature selections approach was performed. In the view if this a subset of relevant features has been selected based on the two feature selection approaches such as forward (method = naïve bayes) and backward elimination (method = naïve bayes and decision tree) to bulid a better model. A total of 7 features has been selected by forward selection, whereas in backward elimination with methods naïve bayes and decision tree is found to be 47 and 48 respectively (Table 4).

Table 4: Different types of Feature seelctions and selected features

Feature Selection Method	Selected Features

Forward Selection	Symptoms, Smoking, CRI, Grams_day, Hemoglobin, Albumin, GGT
Backward Elimination (method = NB)	47, except creatin and HbeAg
Backward Elimination (method = DT)	48, except AST

Since the feature selection approach by backward elimination did not show much variation in selection of features, thus we did not considered for further analysis. However the 7 features selected by forward selection has taken into consideration to evaluate the performance of seven ML operators. The highest accuracy is shown in Naïve Bayes (74.90%) followed by SVM (74.10%), and with average accuracies (71.93%) respectively (Table 5). The average accuracies comparison between 49 and 7 features show a significant decrease of 3.26%, this wide decrease could be due less performance in the 5 ML operators such as KNN, SVM, RF, Auto MLP and AdaBoost performances (Figure 1). However, both Naïve Byes and Bagging showed better performance in 7 selected features in comparison with 49 features (Figure 1), even though bagging do not show highest or followed performance in 7 selected features (Table 5).

Table 5: The performance comparison of classification operators with 7 selected Features

Algorithms	Accuracy	Precision	Recall
KNN	72.76	71.60	78.38
SVM	74.10	71.76	82.43
RF	72.71	72.15	77.03
NB	74.90	73.75	79.73
Auto MLP	72.10	70.73	78.38
Bagging	69.86	67.82	79.73
AdaBoost	67.09	66.67	75.68
Average	71.93	70.64	75.765

7. Discussion

In study, a total of seven classification and 2 feature selection operators has been studied with respect to the HCC dataset using Rapid miner. The 49 features highest accuracy is shown by SVM (81.81%), followed by random forest (79.67%) indicating the two best performance classifiers, this results shows better improvement with the previous studies performed on HCC Survival Data Set with 78.71% for SVM [8], 74% for random forest [15] which used Python tool. On the other hand, SVM also showed the highest accuracy (74.10%) in 7 selected features this resultare in agreement with previous studies [16]. However, the analysis showed the less performance of the 7 selected features classification models (except Naïve Bayes and Bagging, Table 5) in comparison to 49 features, and also not in agreement with the previous studies [6], [8], apart from these we are unable to know which selected features are considered for model building [8]. The classifier performance under default conditions and under fitting of the model could be the possible reasons for the less performance of the 7 selected features which has been noticed in previous studies conducted on heart disease prediction [19], [20]. Thus, this study indicates further drop in the number of features, and as a least number of features for better model performance. However, we cannot rule out since other features do play a role, it should be noted that the ML techniques, tools, parameters along with nature of the data plays a role.

8. Conclusion

The present study showed the performance of the SVM with an accuracy of 74.10%-81.81% and can be used as a good classifier in both 49 and selected features (i.e., 7) for the prediction of HCC disease. The average accuracies differences between 49 and 7 selected shows a small variation and also as a least number of selected features can be useful for the prediction and to build a consistent model. Apart from this, under fitting effect can be noticed with a further drop of the 7 selected features. However, it is dependent on the dataset, techniques, parameters and tools used.

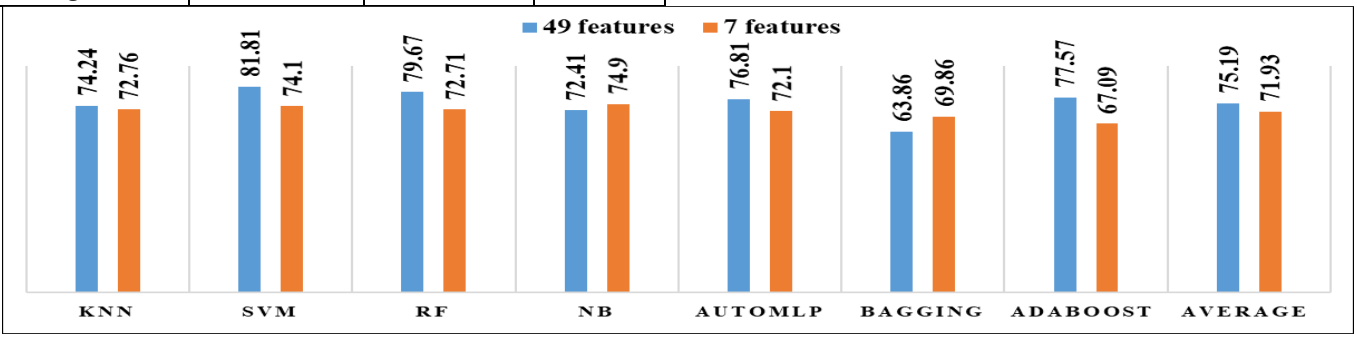

Fig. 1. The accuracies of 7 classification operators and average accuracy

REFERENCES

[1] Wang J, Jain S, Chen D, Song W, Hu C.T, Su Y.H, "Development and Evaluation of Novel Statistical

Methods in Urine Biomarker-Based Hepatocellular Carcinoma Screening", Sci Rep, 8(1):3799, 2018.

[2] Bray F, Ferlay J, SoerjomataramI, Siegel R.L, Torre L.A, Jemal A, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries", CA Cancer J Clin , 68(6), 394-424, 2018.

[3] McGlynn KA, Petrick JL, London WT, "Global epidemiology of hepatocellular carcinoma: an emphasis on demographic and regional variability", Clin Liver Dis, 19(2):223-238, 2015

[4] Balogh J, Victor D 3rd, Asham EH, Burroughs SG, Boktour M, Saharia A, Li X, Ghobrial RM, Monsour HP Jr, "Hepatocellular carcinoma: a review. J Hepatocell Carcinoma. 3, 41–53. 2016

[5] Aziz MA, Kanazawa H, Murakami Y, Kimura F, Yamaguchi M, Kiyuna T, Yamashita Y, Saito A, Ishikawa M, Kobayashi N, Abe T, Hashiguchi A, Sakamoto M, "Enhancing automatic classification of hepatocellular carcinoma images through image masking, tissue changes and trabecular features", J Pathol Inform, 6:26, 2015

[6] Santos MS, Abreu PH, García-Laencina PJ, Simão A, and Carvalho A, "A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients", Journal of biomedical informatics, 58, 49-59, 2015.

[7] Okamoto Shingo, Yokota Takehiro, Lee Jae, Takai Akihiro, Kido Teruhito, Matsuda Megumi, "Detection of Hepatocellular Carcinoma in CT Images Using Deep Learning", Proceedings of the 4th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'18) Madrid, Spain, August 21 – 23, ICBES 133-1 to 133-7 , 2018.

[8] Książek Wojciech, AbdarMoloud, Acharya U Rajendra, Pławiak, Paweł, "A Novel Machine Learning Approach for Early Detection of Hepatocellular Carcinoma Patients", Cognitive Systems Research, 54, 116-127, 2019.

[9] Gulia A, Vohra R, Rani P, "Liver Patient Classification Using Intelligent Techniques", International Journal of Computer Science and Information Technologies, 5 (4), 5110-5115, 2014.

[10] Vijayarani S, Dhayanand S, "Liver Disease Prediction using SVM and Naïve Bayes Algorithms". International Journal of Science, Engineering and Technology Research, 4(4), 816-820, 2015.

[11] Joel Jacob, Joseph Chakkalakal Mathew, Johns Mathew, Elizabeth Issac, "Diagnosis of Liver Disease Using Machine Learning Techniques", International Research Journal of Engineering and Technology (IRJET), 05, 04, 4011-4014, 2018.

[12] Banu Priya M, Laura Juliet P, P.R. Tamilselvi PR, "Performance Analysis of Liver Disease Prediction Using Machine Learning Algorithms" International Research Journal of Engineering and Technology (IRJET), 05, 01, 2018.

[13] Sontakke S, Lohokare J and Dani R, "Diagnosis of liver diseases using machine learning," 2017 International Conference on Emerging Trends & Innovation in ICT (ICEI), 129-133, 2017.

[14] Ghosh SR, Waheed S, "Analysis of classification algorithms for liver disease diagnosis", Journal of Science, Technology and Environment Informatics, 05(01), 361-370, 2017.

[15] Karolina Alicja Sala "Comparison of Machine Learning Algorithms for prediction mortality in patients with Hepatocellular Carcinoma" International Journal of Scientific Engineering and Technology, 7, 9, 85-89, 2018.

[16] Preetam Pal, Birmohan Singh and Manpreet Kaur, "Prediction of Accuracy for Hepatocellular Carcinoma Patients using Cluster based Feature Ranking", International Journal of Medical Research & Health Sciences, 7(8): 130-140, 2018.

[17] http://archive.ics.uci.edu/ml/datasets/HCC+Survival

[18] https://www.kaggle.com/mrsantos/hcc-dataset

[19] Dominic, Vinitha, Deepa Gupta, Sangita Khare, "An Effective Performance Analysis of Machine Learning Techniques for Cardiovascular Disease", Applied Medical Informatics, 36(1), 23-32, 2015.

[20] Satish Chandra Reddy N, Song Shue Nee, Lim Zhi Min & Chew XinYing " Classification and Feature Selection Approaches by Machine Learning Techniques: Heart Disease Prediction", International Journal of Innovative Computing, 9(1), 39-47, 2019.