A comparative study of machine learning algorithms for diagnosis of hepatocellular carcinoma

By

Mohamed Masoud[1], Hoda Anwar Ibrahim[1] and Kamel Hussein Rahouma[2]

[1]Department of Biomedical Engineering, Faculty of Engineering, Minia University, Egypt

[2]Department of Electrical Engineering, Faculty of Engineering, Minia University; Currently: Vice Dean of the Faculty of Computer Science, Nahda University in Beni Suef, Egypt.

Corresponding email: h.anwar@mu.edu.eg

Abstract

In adults, hepatocellular carcinoma (HCC) is the most frequent type of liver cancer. Due to the disease's modest symptoms, it is a very difficult challenge for medical experts to identify the disease in its early stages. Frequently, the symptoms show up only when it's too late. Using machine learning approaches such as logistic regression, Gaussian naïve Bayes, random forest and artificial neural network, a comparative diagnostics strategy is proposed in this study.

Artificial Neural Network (ANN) is shown to have the highest accuracy 95% followed by Random Forest (RF) 79% in the classification model. Also, in the feature selection model, ANN has the highest performance followed by RF giving accuracy of 92.5% and 77.4% respectively. Our algorithm demonstrates how the proposed model might present a distinct and accurate method for diagnosing HCC using machine learning techniques.

1. Introduction

Hepatocellular carcinoma (HCC) is one of the most leading causes of cancer-related mortality globally[1]. Early detection is one of the most effective prevention strategies , but it can be challenging to distinguish cancer from other diseases, especially at an early stage, due to the multitude of diverse causes that can cause cancer[2]. Hepatitis B or C infection, non-alcoholic fatty liver disease (NAFLD), excessive alcohol consumption, smoking, type 2 diabetes, obesity, and foods contaminated with aflatoxin are among the risk factors[3]. Alpha-fetoprotein (AFP) blood tumor marker, radiographic imaging, liver biopsy, and biomarkers are common diagnostic methods to improve patient survival[4]. Egypt was rated the country with the highest rate of adult viral hepatitis C (HCV) infection in 2015, at 7%. Because all infected persons could not receive treatment, Egypt's government started a countrywide campaign called "100 Million Seha" (seha is an Arabic term that means "health") from October 2018 to April 2019. Around 35 million persons had been tested for HCV by the end of March 2019. We think that a variety of issues can be solved using machine learning techniques, which would speed up analysis and reduce mistake risks[5]. Early identification of HCC is crucial for improving the prognosis of the disease because it is predicted that the survival rate can rise to 35% if it is achieved correctly[6].

The principal objective of this research is to distinguish between liver patients and healthy people using classification algorithms. In this investigation, the performance of four classification algorithms—Logistic Regression (LR) , Gaussian Naïve Bayes (NB), Random Forest (RF), and artificial neural networks (ANN)—was compared using data from liver patient.

The paper is organized in five sections. Section (1) is an introduction and section (2) give a literature review of the related work. Section (3) explains the methods applied in this paper and section (4) illustrates the experiment and results of the paper. It also discusses these results and compares them with the previous research results. Section (5) highlights some conclusions and suggests some points for future work. A list of the sources used is provided at the end of the paper.

2. Related Work

In recent years, multiple machine learning models have been developed to early liver cancer detection and diagnosis, particularly hepatocellular carcinoma[7]. There are various different types of classification algorithms used for cancer diagnosis, or, often known as classifiers. Some of these are the Genetic Algorithm (GA), Fuzzy Set (FS), Rough Set (RS), Artificial Neural Network (ANN), Support Vector Machine (SVM), and Fuzzy Set[8]. In order to provide a representative data set, [7]developed a new cluster-based oversampling strategy for HCC detection based on the K-means clustering and SMOTE algorithm. Additionally employed in this investigation were logistic regression and neural networks. The most effective model has a classification accuracy of 75.19%.The LDA-GA-SVM approach, which combines the dimensional reduction LDA method with genetic algorithms and a support vector machine, was proposed by [9]. The proposed classifier had a 90.30% accuracy rate. [10] provided a model with the same accuracy. [11]suggested a combination of the support vector machine and the Lasso approach. A dataset of 331 patients from Sir Run Shaw Hospital, Zhejiang University, China, served as the basis for the study. The classification accuracy of the suggested model was 89.18%. The missing values were imputed using HEOM distance during the preprocessing stage, and K-means clustering was used. Later, synthetic minority over-sampling technique (SMOTE) was used to obtain the balanced dataset. On the balanced dataset, techniques for logistic regression and neural networks were both used. The accuracy of neural networks is determined to be between 68.7% and 75.2% for both without and with clusters, while for logistic regression it is found to be between 70% and 73% [7]. In a separate article, a Markov Blanket-based clustering technique was used during preprocessing, where the redundancy among the features is calculated based on ranking. The suggested technique was evaluated using a total of six different classifiers, and SVM demonstrated the highest accuracy at 76.25%, followed by Naive Bayes (73.95%) and KNN at 72.10% [12]. The following deep learning techniques were employed to detect liver cancer[13].

Due to previously explained various machine learning techniques, the main objective of this work is to present a comparison between machine learning algorithms that have proven success for hepatocellular carcinoma classification and find the best model which is rapidly and accurately classify liver cancer.

3. Materials and Methods

This section discusses the material and techniques used in our proposed work. We also discussed the HCC database that we utilized to implement our strategies into practice.

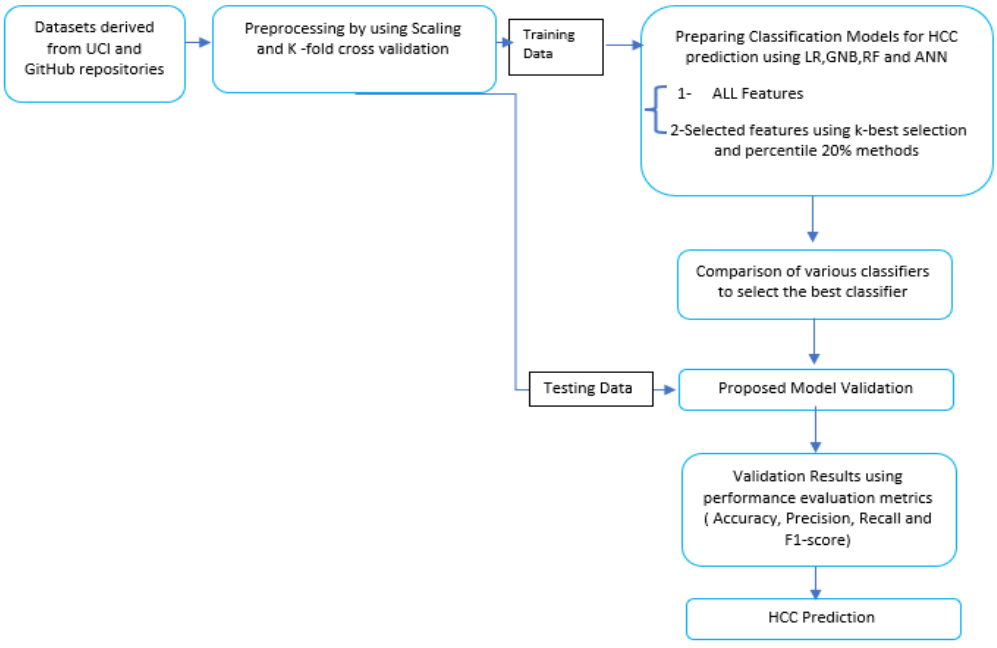


Figure 1: ML classification algorithms for the prediction of liver illness

This study aims to provide an analysis of the performance of ML classification algorithms for the prediction of liver illness using dataset collected from Coimbra Hospital (at the university center) in Portugal taken from the UCI ML repositories[14] and GitHub[15]. The process depicted in Figure 1 is used to prepare the entire research.

Proposed Methodology

3.1 Dataset:

Hepatocellular Carcinoma Complete Balanced Dataset is used in this study. The dataset's missing values were filled in using KNN (K=1) and HEOM distance, and the dataset's balance (205 rows from 167) was achieved using SMOTE (k = 3) and the oversampling approach. This data set was assembled using available resources [18].

The HCC balanced dataset does not contain any missing values (Table 1). The dataset has 49 features/attributes in total, of which 26 are qualitative variables and 23 are quantitative variables. According to the assessment of the 1-year result, the target class is a binary variable with the values 0 (Dies) and 1 (Lives).

Table 1. HCC dataset identification

#	Features / Attributes	Value/Range
1	Gender	0,1
2	Symptoms	0,1
3	Alcohol	0,1
4	Hepatitis B Surface Antigen	0,1
5	Hepatitis B e Antigen	0,1
6	Hepatitis B Core Antibody	0,1
7	Hepatitis C Virus Antibody	0,1
8	Cirrhosis	0,1
9	Endemic Countries	0,1
10	Smoking	0,1
11	Diabetes	0,1
12	Obesity	0,1
13	Hemochromatosis	0,1
14	Arterial Hypertension	0,1
15	Chronic Renal Insufficiency	0,1
16	Human Immunodeficiency Virus	0,1
17	Nano alcoholic Steatohepatitis	0,1
18	Esophageal Varices	0,1
19	Splenomegaly	0,1
20	Portal Hypertension	0,1
21	Portal Vein Thrombosis	0,1
22	Liver Metastasis	0,1
23	Radiological Hallmark	0,1
24	Age at diagnosis	20-93
25	Grams of alcohol per day	0-500
26	Packs of cigarettes per year	0-510
27	Performance status	0,1,2,3,4
28	Encephalopathy degree	0,1,2,3
29	Ascites degree	0,1,2,3
30	International Normalized ratio	0.84-4.82
31	Alpha-Fetoprotein	1.2-1810348
32	Hemoglobin	5-18.7
33	Mean Corpuscular volume	69.5-119.6

34	Leukocytes	2.2-13000
35	Platelets	1.71-459000
36	Albumin	1.9-4.9
37	Total Bilirubin	0.3-40.5
38	Alanine transaminase	11-420
39	Aspartate glutamyl transferase	17-553
40	Gamma glutamyl transferase	17-553
41	Alkaline Phosphate	1.28-980
42	Total Proteins	3.9-102
43	Creatinine	0.2-7.6
44	Number of nodules	0-5
45	Major dimension of nodule	1.5-22
46	Direct Bilirubin	0.1-29.3
47	Iron	0-224
48	Oxygen	0-126
49	Ferritin	0-2230
50	Class	0,1

3.2 Preprocessing:

It should be noted that preprocessing is extremely important in machine learning for maximizing the effectiveness and precision of the classifier. The output model can easily become over-fitted and have poor generalization capabilities if suitable preprocessing is not used. The classifier's training and prediction performance are both impacted by preprocessing.

3.2.1 Scaling:

The dataset must be rescaled (i.e., the variable values between 0 and 1) because it has a wide numerical range. StandardScaler scales a feature to unit variance after subtracting the mean to standardize it by dividing all the numbers by the standard deviation to get the unit variance.

$$\text{Scaled value} = \frac{x-u}{s} \qquad (1)$$

where: x - a sample – u - mean of training samples for a single feature – s - a standard deviation of training samples for a single feature.

3.2.2 Cross Validation:

To make sure that relative class frequencies are roughly kept in each training and validation fold, this work use the stratified K-Folds cross-validation (k =10) approach[16].This method keeps the proportions between classes while creating the testing and training sets from separate random selections of HCC data for each class. The complete database that is accessible is cross validated.

3.3 Classification Models:

3.3.1 Logistic Regression (LR)

Applying sigmoid functions causes linear regression to transform into LR [9]. A logistic function uses a wide scale to narrow the range of y values from 0 to 1. Despite making assumptions about the data distribution, maximum-likelihood estimators in logistic regression are reliable for estimating linear-regression coefficients, a machine learning technique.

3.3.2 Naïve Bayesian Classifier (NB)

A Naive Bayes classifier is an uncomplicated probability - based classification method that relies on the application of the Bayes theorem and a strong independence assumption. The self-determining feature model is a more accurate name for the underlying probability model[17].

P(C|X) is the probability of even C if X has occurred, P(X|C) is the probability of even X given that C has occurred, P(C) is the probability of event C, and P(X) is the probability of event X. Where x is ascribed, C is classes. The Bayes formula can be expressed as follows when X is substituted[18].

$$P(C|X) = \frac{P(C)\,P(X|C)}{p(x)} \quad \boldsymbol{Where: x = x_1.x_2.x_3 \ldots x_n} \qquad (2)$$

The conditional probabilities and the class probabilities are used in the training phase to determine the class label of a testing data point. The data point in two categorized data sets is ordered based on whose class probability is higher [19].

$$P(c|x_1.x_2 \ldots x_n) = \frac{P(c)P(x_1.x_2 \ldots x_n|c)}{P(x_1.x_2 \ldots x_n)} \qquad (3)$$

The Naive Bayes classifier has the benefit that it can estimate the means and variances of the classification relevant variables with a relatively modest quantity of training data [17].

3.3.3 Random Forest (RF)

It is a combination of extremely randomized classification trees created by tree induction using bootstrap samples from the training dataset and random feature selection [20]. In contrast to conventional decision tree (DT) algorithms, the new training dataset uses a combination of distinct features to develop the resulting tree to its maximum depth. These fully developed produced trees don't undergo pruning [21]. The RF algorithm's primary advantage over other DT tree methods is this.

Table 2. Grid and Randomized Search list of hyperparameters.

Methods	Hyper Parameters	Value
Random Forest	Bootstrap	False, True
	n estimator	40,50,60,70,80,90,100
	Min sample leaf	1,2
	Min sample split	2,3,4,5
	Max features	auto, sqrt
	Max depth	2,3,4,5,6,7,8

Default sklearn hyperparameters are used for both logistic regression and gaussian naïve bayes. But in order to extract the optimal hyperparameter for Random Forest, we apply Grid and Randomized Search optimization methods. A list of the hyperparameters we used in this study is presented in Table 2, and the results of our Grid search are shown below in Table 3.

3.3.4 Neural Networks

A well-known machine learning technique that simulates the neural networks in the human brain is called neural networks. Artificial neural networks (ANNs) and deep neural networks (DNNs) are the most often utilized neural networks. One of the first neural network models is the ANN [22], which is composed of three layers: an input layer, a hidden layer, and an output layer. A perceptron or a multilayer perceptron (MLP), with or without a hidden layer, can be a component of an ANN model. However, ANNs are unable to handle medical imaging tasks directly. DNNs are frequently utilized in the creation of models as a result of the development of deep learning [23]. One of the most popular DNNs that can automatically recognize, and segment medical images is convolutional neural networks (CNNs)[24][25].

In our work neural network is trained from scratch with 100 epochs (batch size = 100). RMS prop optimizer is used to update network parameters learning rate =0.001 , momentum=0 and epsilon = $1e^{-7}$. Relu activation functions are used in hidden layers and sigmoid activation function is used only in the output layer for the binary classification.

3.4 Feature Selection:

Feature selection is a technique for choosing the data set's best features that are more significant and have a greater impact on classification or prediction. The choice of features is crucial because the accuracy of learning models might be impacted by the inclusion of irrelevant variables in the data. To put it another way, feature

selection is used to limit over fitting, shorten training times, and improve learning model accuracy. We employed the additional chi square and the k-best feature selection approach as two feature selection techniques in our experiment[26].

3.4.1 K-Best :

To determine the best features that have a strong correlation with the goal feature, statistical trials can be carried out. To choose non-negative data, we used the Scikit-Learn library's Select K-Best class. Fourteen top features[26].

3.4.2 Select Percentile:

Another feature selection technique that may be used with Scikit-learn is Select Percentile , which gives percentiles to the features based on their score. The performance of a classifier can therefore be taken into consideration when choosing features for a cutoff percentile.

Feature Selection Score Function (Chi-square):

It has been utilized successfully in applications for medical data analysis. The chi-square test is a technique for testing a hypothesis when counting data. It compares the ratio of two or more sample rates or is used for correlation analysis of two categorical variables, comparing the theoretical frequency and the actual frequency (the degree of fit between the actual frequency). The chi-square test's fundamental formula is as follows:

$$X^2 = \sum \frac{(A*T)^2}{T} \qquad (4)$$

where is X the chi-square value, T is the theoretical frequency, and A represents the actual frequency. Chi square analyses the degree of divergence between actual values and theoretically estimated values to measure the impact of features on classification[27]. This score function is used for both K-Best and Select Percentile methods.

4. Results and Discussion:

Platform Settings (Hardware and Software):

The Pandas and Sklearn libraries were utilized in conjunction with the Python programming language for this work. One core, 8 GB of RAM, and an Intel Core(TM) i5-6200U 2.4 GHz processor were utilized. Without feature selection in the first trial and with feature selection in the second, we employed the four algorithms indicated above.

Using Python 3.7, the machine learning model is created. These main libraries are employed: Data loading with Pandas, Use of based classifiers in Sklearn. Stratified k fold, ensemble, metrics, and model_selection methods are used from Sklearn library. Matplotlib library for plotting model graphs.

An 2*2 matrix used to assess model performance is called a confusion matrix. The machine-learning models and the actual goal values are compared [28], [29]. Performance metrics like Accuracy,Precision,Recall,F1-score, and AUC are calculated based on the confusion matrix to aid in the prediction and identify specific diagnosis.

According to Equation(5), the percentage of cases that are successfully classified acts as an indicator for the classification's correctness.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \qquad (5)$$

Divide the true-positive findings by the total of true positives and false positives in order to calculate precision, according to Equation (6).

$$\text{Precision(Pr)} = \frac{TP}{TP+FP} \qquad (6)$$

Similar to equation (6), equation (7) divides the true positives by the true positives and false negatives to determine the recall.

$$\text{Recall (Re)} = \frac{TP}{TP+FN} \qquad (7)$$

Equation(8) illustrates how the F1-score combines the precision and recall into a single metric that encompasses both features.

$$\text{F1} = \frac{2(Pr*Re)}{Pr+Re} \qquad (8)$$

First experiment (without feature selection) :

Table 3. Grid search and Randomized Search CV Results

Methods		**Result of Grid search**		**Result of Randomized search**		**Default Hyperparameters**	
Random Forest	Hyper parameters results	Bootstrap n estimator Min sample leaf Min sample split Max features Max depth	True 60 2 2 Sqrt 10	Bootstrap n estimator Min sample leaf Min sample split Max features Max depth	False 80 2 5 Sqrt 8	Bootstrap n estimator Min sample leaf Min sample split Max features Max depth	False 100 2 5 Sqrt 8
	Average Test Accuracy	76%		78%		80.65	

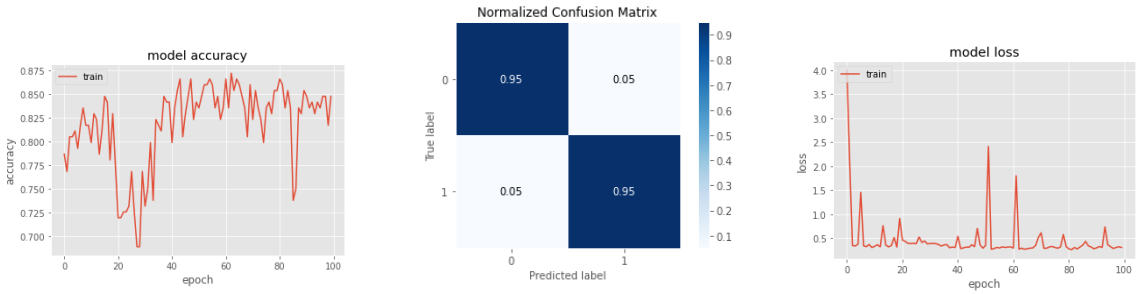
From Table 3, it is shown that randomized search method is better than grid search method for random forest classifier.

Table 4. Grid search and Randomized Search CV Results.

Classifiers	**Accuracy (Random data split)**	**Accuracy(K fold cross validation)**
Logistic Regression	70.97%	74%
Gaussian Naïve Bayes	67.74%	72%
Random Forest	80.65%	79%
ANN	76%	95%

Fig.2 The accuracies of 4 classifiers in case of random data split and k-fold cross validation.

Table 2 and Fig.2 show that K-fold cross validation owe a high accuracy compared with random data split. Which has an impressive effect in preprocessing step. Also it is clear that ANN classifier has a great capability to detect HCC with accuracy of 95%.



a) b) c)

Fig.3 ANN Model a) Accuracy b) Confusion Matrix c) Loss.

Fig.3 Represents ANN model accuracy, loss performance through 100 epochs after preprocessing step and confusion matrix respectively.

Second experiment(with feature selection) :

The strategy of feature selection was used to advance the model accrual. Based on the two feature selection processes, a subset of relevant features has been chosen for this view, including chi-square and k-best methods.

Table 5. Attribute selections and selected features.

Attribute Selection	Selected Features
Percentile 20 (10 Features)	Leucocytes, Platelets, Albumin, Total_Bil, GGT, ALP, Dir_Bil, Iron, Sat, and Ferritin.
Selected k-best method (Top13 Features)	AFP, Total_Bil, AST, ALP, Hemoglobin, platelets, Age, Direct bilirubin, creatine, total protein, ferritin, performance status, and major dimension of nodule.

Percentile method:

Table 6. The performance evaluation of the classification operators using the Percentile 20% Feature Selection.

Classifiers	Accuracy	Average Precision	Average Recall	F1-Score
Logistic Regression	67.74%	68%	67%	67%
Gaussian Naïve Bayes	67.74%	77%	66%	63%
Random Forest	77.4%	77.5%	78%	77.5%

From Table 6 it is shown that Random Forest has the best performance among other classifiers with accuracy 77.4% and f1-score 77.5%

Select K-best Model:

Below features with their scores resulted using Select K-best technique after dropping Gender and class columns from the main dataset. Therefor 48 attributes are remained starting from feature 1 which is Symptoms and ending with feature 47 which is Ferritin.

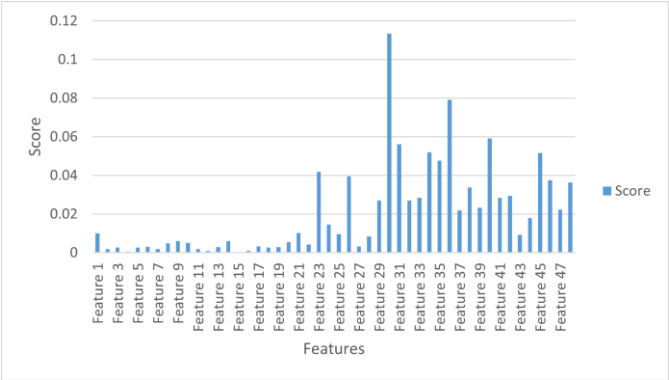


Fig 4. Select K-best method feature selection result.

From fig 4 it shows that top 3 attributes are no.29 (Alpha Feto-protein), no.35(Total Bilirubin) and no.39(Alkaline Phosphate) respectively.

Table 7. The performance evaluation of the classification operators using k-best Features.

Classifiers	**Accuracy**	**Average Precision**	**Average Recall**	**F1-Score**
Logistic Regression	66.13%	66%	66%	66%
Gaussian Naïve Bayes	67.74%	77%	66%	63%
Random Forest	66.13%	67%	67%	66%

From Table 7. it is shown that Random Forest has the best performance among other classifiers with accuracy 67.73%. since ANN is proven to have an effective performance as mentioned above. We also used ANN after feature selection methods.

Table 8. Accuracy comparison of different classification operators after feature selection.

Classifiers	**Accuracy after feature selection using k-best**	**Accuracy after feature selection using percentile 20%**
Logistic Regression	66.13%	67.74%
Gaussian Naïve Bayes	67.74%	67.74%
Random Forest	66.13%	77.4%
ANN	77.5%	92.5%

Table.8 shows that percentile 20% method is better than k-best for feature selection by accuracy 92.5%.By comparing Table.4 with Table.8 using all features, preprocess it and train the model using ANN gives the best results as shown in fig.5.

Table 9. Accuracies of 4 classification operators.

Classifiers	**Accuracy (Random data split)**	**Accuracy(K fold cross validation)**	**Accuracy after feature selection using k-best**	**Accuracy after feature selection using percentile 20%**
Logistic Regression	70.97%	74%	66.13%	67.74%
Gaussian Naïve Bayes	67.74%	72%	67.74%	67.74%
Random Forest	80.65%	79%	66.13%	77.4%
ANN	76%	95%	77.5%	92.5%

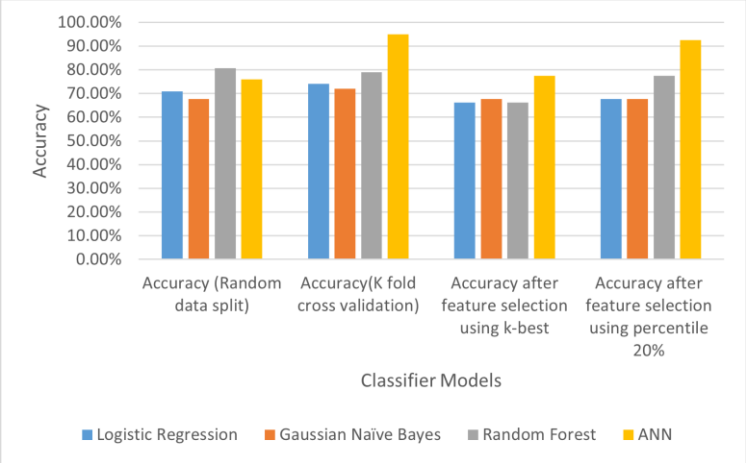


Fig 5. Comparison of overall accuracies for 4 classifiers.

Comparison with other results:

Table 10. Accuracies of 4 classification operators.

No	Method	Accuracy	Dataset	Publication Year	Reference
1	NN + augmented set approach	75.19%	Same dataset CHUC	2017	[19]
2	BFA + RF	83.5%	CHUC	2018	[9]
3	SVC with GA optimizer	88.49%	CHUC	2019	[10]
4	LASSO + SVM RFE + LASSO + SVM	89.14%	CHUC	2019	[10]
5	LDA-GA-SVM	90.30%	CHUC	2019	[30]
6	K-means + SMOTE + SVM	84.90%	CHUC	2020	[31]
7	Relief + LDA NCA + FGSVM	92.12%	CHUC	2020	[10]
8	GA	90.30%	Not the same data	2020	[32]
9	SMOTE + XGBOOST	87%	Not the same data	2021	[33]
10	NCA + GA + SVM	87.4%	Not the same data	2022	[34]
11	**This Study**	**95%**	**CHUC**	**2022**	

The identification of hepatocellular carcinoma in the CHUC dataset has been described in a few articles published in the literature. Table 10 lists the statistics that were gathered. The scientific literature contains a variety of approaches to the challenge of hepatocellular carcinoma detection.

5. Conclusion and Future Work:

The present study discusses four classification models and two feature selection methods with respect to the HCC dataset. ANN 95% and random forest (80.65%) are the two best performance classifiers for the 48 features, respectively. Also, ANN shows a high performance among other classifiers in case of feature selection with accuracy 92.5 .This result represents a better improvement over earlier studies on the HCC survival data Set. The average accuracy differences between the 49 and 7 selected features reveal a slight fluctuation, and the prediction and construction of a reliable model can benefit from having the fewest characteristics picked up. Along with this, a further drop of the chosen features will produce an under fitting effect. However, it depends on the dataset, methods, settings, and equipment that were employed.

We intend to expand our dataset and apply another machine and deep learning algorithms to solve challenges in hepatocellular carcinoma survival prediction in the future.

6. References:

[1] J. Balogh, "Hepatocellular carcinoma: a review - PubMed," no. July 2018, 2016, doi: 10.2147/JHC.S61146.

[2] R. Etzioni *et al.*, "The case for early detection," *Nat. Rev. Cancer*, vol. 3, no. 4, pp. 243–252, 2003, doi: 10.1038/nrc1041.

[3] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA. Cancer J. Clin.*, vol. 68, no. 6, pp. 394–424, 2018, doi: 10.3322/caac.21492.

[4] K. Saigo *et al.*, "Integration of hepatitis B virus DNA into the myeloid/lymphoid or mixed-lineage leukemia (MLL4) gene and rearrangements of MLL4 in human hepatocellular carcinoma," *Hum. Mutat.*, vol. 29, no. 5, pp. 703–708, 2008, doi: 10.1002/humu.20701.

[5] S. Almotairi, G. Kareem, M. Aouf, B. Almutairi, and M. A. M. Salem, "Liver tumor segmentation in CT scans using modified segnet," *Sensors (Switzerland)*, vol. 20, no. 5, 2020, doi: 10.3390/s20051516.

[6] J. Wang, S. Jain, D. Chen, W. Song, C. T. Hu, and Y. H. Su, "Development and Evaluation of Novel Statistical Methods in Urine Biomarker-Based Hepatocellular Carcinoma Screening," *Sci. Rep.*, vol. 8, no. 1, pp. 1–9, 2018, doi: 10.1038/s41598-018-21922-9.

[7] M. S. Santos, P. H. Abreu, P. J. García-Laencina, A. Simão, and A. Carvalho, "A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients," *J. Biomed. Inform.*, vol. 58, pp. 49–59, 2015, doi: 10.1016/j.jbi.2015.09.012.

[8] G. Scheuner, C. P. Mitzscherling, C. Pfister, A. Pöge, and E. Seidler, "Functional morphology of the human placenta," *Zentralbl. Allg. Pathol.*, vol. 135, no. 4, pp. 307–328, 1989.

[9] L. Ali, I. Wajahat, N. Amiri Golilarz, F. Keshtkar, and S. A. C. Bukhari, "LDA–GA–SVM: improved hepatocellular carcinoma prediction through dimensionality reduction and genetically optimized support vector machine," *Neural Comput. Appl.*, vol. 33, no. 7, pp. 2783–2792, 2021, doi: 10.1007/s00521-020-05157-2.

[10] W. Książek, M. Hammad, P. Pławiak, U. R. Acharya, and R. Tadeusiewicz, "Development of novel ensemble model using stacking learning and evolutionary computation techniques for automated hepatocellular carcinoma detection," *Biocybern. Biomed. Eng.*, vol. 40, no. 4, pp. 1512–1524, 2020, doi: 10.1016/j.bbe.2020.08.007.

[11] P. Aonpong *et al.*, "Comparison of Machine Learning-Based Radiomics Models for Early Recurrence Prediction of Hepatocellular Carcinoma," *J. Image Graph.*, vol. 7, no. 4, pp. 117–125, 2019, doi: 10.18178/joig.7.4.117-125.

[12] P. Pal, B. Singh, and M. Kaur, "Prediction of Accuracy for Hepatocellular Carcinoma Patients using Cluster based Feature Ranking," *Int. J. Med. Res. Heal. Sci.*, vol. 7, no. 8, pp. 130–140, 2018.

[13] M. Chen *et al.*, "Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning," *npj Precis. Oncol.*, vol. 4, no. 1, pp. 1–7, 2020, doi: 10.1038/s41698-020-0120-3.

[14] "UCI Machine Learning Repository: HCC Survival Data Set." http://archive.ics.uci.edu/ml/datasets/HCC+Survival (accessed Sep. 22, 2022).

[15] "HCC dataset | Kaggle." https://www.kaggle.com/datasets/mrsantos/hcc-dataset/code (accessed Sep. 22, 2022).

[16] X. Zeng and T. R. Martinez, "Distribution-balanced stratified cross-validation for accuracy estimation," *J. Exp. Theor. Artif. Intell.*, vol. 12, no. 1, pp. 1–12, 2000, doi: 10.1080/095281300146272.

[17] D. Vijayarani, "Liver Disease Prediction using SVM and Naïve Bayes Algorithms," *Int. J. Sci. Eng. Technol. Res.*, vol. 4, no. 4, pp. 816–820, 2015.

[18] C. Güzel and F. Engineering, "Breast Cancer Diagnosis Based on Naïve Bayes Machine Learning Classifier with KNN Missing Data Imputation," *AWERProcedia Inf. Technol. Comput. Sci.*, vol. 04, no.

May, pp. 401–407, 2013.

[19] W. Książek, M. Abdar, U. R. Acharya, and P. Pławiak, "A novel machine learning approach for early detection of hepatocellular carcinoma patients," *Cogn. Syst. Res.*, vol. 54, no. May, pp. 116–127, 2019, doi: 10.1016/j.cogsys.2018.12.001.

[20] H. Dureja, S. Gupta, and A. K. Madan, "Topological models for prediction of pharmacokinetic parameters of cephalosporins using random forest, decision tree and moving average analysis," *Sci. Pharm.*, vol. 76, no. 3, pp. 377–394, 2008, doi: 10.3797/scipharm.0803-30.

[21] M. Pal, "Random forest classifier for remote sensing classification," *Int. J. Remote Sens.*, vol. 26, no. 1, pp. 217–222, 2005, doi: 10.1080/01431160412331269698.

[22] A. E. Karnga, "Neural networks," *Proc. Int. Jt. Conf. Neural Networks*, vol. 6, pp. 4419–4421, 1999.

[23] J. Lee *et al.*, "2017-Lee-Deep Learning in Medical Imaging_ Gen," *Korean J. Radiol.*, vol. 18, no. 4, pp. 570–584, 2017.

[24] S. Muhammad, A. Muhammad, M. Adnan, Q. Muhammad, A. Majdi, and M. K. Khan, "Medical Image Analysis using Convolutional Neural Networks A Review," *J. Med. Syst.*, pp. 1–13, 2018.

[25] R. Brehar *et al.*, "Comparison of deep-learning and conventional machine-learning methods for the automatic recognition of the hepatocellular carcinoma areas from ultrasound images," *Sensors (Switzerland)*, vol. 20, no. 11, pp. 1–22, 2020, doi: 10.3390/s20113085.

[26] O. P. Samantray and S. Narayan Tripathy, "A Knowledge-Domain Analyser for Malware Classification," *2020 Int. Conf. Comput. Sci. Eng. Appl. ICCSEA 2020*, no. March, 2020, doi: 10.1109/ICCSEA49143.2020.9132916.

[27] C. F. Selection, A. Overfitting, and R. T. Time, "Table of Contents Chi-Square Feature Selection," pp. 4–7.

[28] D. H. Kim, B. Kim, S. Y. Youn, H. Kim, and J. Il Choi, "Diagnostic performance of klca-ncc 2018 criteria for hepatocellular carcinoma using magnetic resonance imaging: A systematic review and meta-analysis," *Diagnostics*, vol. 11, no. 10, 2021, doi: 10.3390/diagnostics11101763.

[29] E. Y. Dessie *et al.*, "Construction and validation of a prognostic gene-based model for overall survival prediction in hepatocellular carcinoma using an integrated statistical and bioinformatic approach," *Int. J. Mol. Sci.*, vol. 22, no. 4, pp. 1–18, 2021, doi: 10.3390/ijms22041632.

[30] M. A. S. Ali *et al.*, "A Novel Method for Survival Prediction of Hepatocellular Carcinoma Using Feature-Selection Techniques," *Appl. Sci.*, vol. 12, no. 13, 2022, doi: 10.3390/app12136427.

[31] W. Książek, M. Gandor, and P. Pławiak, "Comparison of various approaches to combine logistic regression with genetic algorithms in survival prediction of hepatocellular carcinoma," *Comput. Biol. Med.*, vol. 134, no. June, 2021, doi: 10.1016/j.compbiomed.2021.104431.

[32] L. Akter, Ferdib-Al-Islam, M. M. Islam, M. S. Al-Rakhami, and M. R. Haque, "Prediction of Cervical Cancer from Behavior Risk Using Machine Learning Techniques," *SN Comput. Sci.*, vol. 2, no. 3, 2021, doi: 10.1007/s42979-021-00551-6.

[33] M. Mroweh, T. Decaens, P. N. Marche, Z. M. Jilkova, and F. Clément, "Modulating the crosstalk between the tumor and its microenvironment using rna interference: a treatment strategy for hepatocellular carcinoma," *Int. J. Mol. Sci.*, vol. 21, no. 15, pp. 1–26, 2020, doi: 10.3390/ijms21155250.

[34] Z. Liu *et al.*, "Using embedded feature selection and cnn for classification on ccd-inid-v1—a new iot dataset," *Sensors*, vol. 21, no. 14, pp. 1–34, 2021, doi: 10.3390/s21144834.