Journal of Ambient Intelligence and Humanized Computing (2022) 13:5763–5777
https://doi.org/10.1007/s12652-021-03256-z

ORIGINAL RESEARCH

Improved hepatocellular carcinoma fatality prognosis using ensemble learning approach

Manoj Sharma[1] · Naresh Kumar[2]

Received: 19 September 2020 / Accepted: 29 March 2021 / Published online: 10 April 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract

Hepatocellular Carcinoma (HCC) is the most common type of liver cancer which accounts for around 75% of all liver cancer cases. From statistical data, it has been found that fatality due to liver cancer is higher regardless of improved screening and discoveries in medicines, HCC escalate fatality rate. This paper presents an ensemble learning model for HCC survival prediction. The input predictors for the proposed model consist of geographical information, risk factors and clinical trial information of HCC patients. Fifteen different models are presented to evaluate the prediction. These models present data pre-processing, feature reduction/elimination and survival classification phase. For feature evaluation, LASSO Regression (L-1 penalization), Ridge Regression (L-2 penalization), Genetic Algorithm (GA) Optimization and Random Forest (RF) are proposed for weight valuation of features wherein features with significant weights are selected for prediction. With the aid of feature evaluators, L-1 penalized Nu-Support Vector Classification (Nu-SVC) model, L-2 penalized Nu-SVC model, GA optimized Nu-SVC model, RF-NuSVC model, L-1 penalized RidgeCV (RCV) model, L-2 penalized RCV model, GA optimized RCV model, RF-RCV model, L-1 penalized Gradient Boosting Ensemble Learning (GBEL) model, L-2 penalized GBEL model, GA optimized GBEL model and RFGBEL model are presented for survival prediction. The prediction performances of models were measured in terms of accuracy, recall/sensitivity, F-1 score, Log-Loss score, Jaccard score and Area Under Receiver Operating Curves (AUROC). The results indicate that RFGBEL model shows excellent performance in contrast to other proposed models. The proposed RFGBEL model achieves an accuracy of 93.92%, sensitivity of 94.73%, F-1 score of 0.93, Log-Loss/Cross entropy score of 5.89 and Jaccard score of 0.72. RFGBEL estimates value of area under the curve as 0.932. Comparison of RFGBEL model with other existing state of the art models are presented for performance assessment. Overall, the RFGBEL model has a capability to predict the result with more accuracy and sensitivity by means of machine learning and data mining approach.

Keywords Hepatocellular carcinoma · Tree based ensemble learning · Supervised learning and liver cancer

1 Introduction

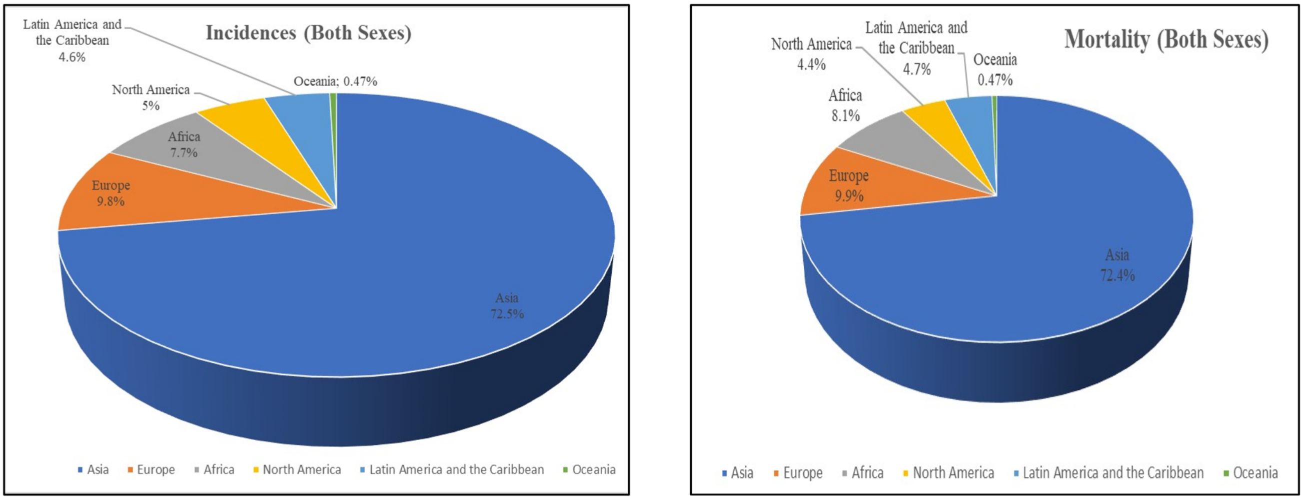
HCC is the second prominent reason for cancer related fatality globally and is fifth most common cancer type (Galle et al. 2018; Njei et al. 2015). It is also categorized as fifth most common cause of cancer in men and the seventh most common cause in women (Fitzmaurice et al. 2018). In studies, it was found that probability of occurring of HCC is more frequent in males as compared to females (2.4:1) (Ferlay et al. 2010). HCC is typical category of primary liver cancer and accounts for around 75% of all liver cancer cases. Intrahepatic Cholangiocarcinoma (ICC) is another classification of liver cancer and comprises around 12–15% of all liver cancer cases (Petrick and McGlynn 2019). The liver cancer incidence and mortality vary from Europe to Eastern Asia due to its wide geographical variations. Figure 1 shows the incidence and mortality (both sexes) for liver cancer globally (Ferlay et al. 2018). In 2018, total 841,080 new incidences and 781,631 mortality was reported globally. Its incidence varies from Asia (609,596 new cases) to Europe (82,466 new

✉ Manoj Sharma
neelmanoj@gmail.com

Naresh Kumar
naresh_uiet@yahoo.com

[1] Department of Electronics and Communication Engineering, Giani Zail Singh Campus College of Engineering and Technology, MRSPTU, Bathinda, India

[2] Department of Electronics and Communication Engineering, UIET, Panjab University, Chandigarh, India



(a)

(b)

Fig. 1 **a** Liver cancer incidences globally in 2018 and **b** Mortality due to liver cancer in 2018

cases) and mortality also varies from Asia (566,269 fresh cases) to Europe (77,375 fresh cases) (Ferlay et al. 2018). HCC primarily appears in the individuals with chronic liver diseases, such as cirrhosis. Cirrhosis is caused mainly by Hepatitis B and Hepatitis C infections. Chronic infections with Hepatitis B virus and/or Hepatitis C Viruses are major risk factors for HCC.

Approximately 60% of HCC cases are caused by viral hepatitis (de Martel et al. 2015). Alcohol, heavy exposure to aflatoxin and metabolic syndrome are the other important risk factors of HCC. From statistical data, it has been found that number of fatality due to HCC is high and is likely to be continued. Regardless of improved screening and discoveries in medicines, HCC shows escalated mortality rate. The 5-year overall survival of HCC patients is 3–5% across all countries (Dhanasekaran et al. 2012).

With the recent developments in data mining, soft computing and machine learning techniques many researchers have taken keen interest in medical and clinical analysis from the available data sources. Recently, researchers and medical practitioner have been applying machine learning and statistical methods to develop prediction models for clinical analysis and treatments. Many state-of-art literatures have been presented for prediction of HCC. Masaya et al. (2019) proposed Gradient Boosting (GB) based prediction model. Clinical information was collected from 1582 patients (539 HCC patients; 1043 non-HCC patients) at University of Tokyo Hospital from January 1997 to May 2016. Using Gradient Boosting classifier, they obtain an accuracy of 87.34%, for the data used in their study, among all the proposed classifiers. Decision Tree Algorithm based HCC prediction model was proposed by Omran et al. (2015). The data was collected from Endemic Medicine Department, Cairo University Hospital, Egypt. Total 315 patients with Hepatitis C Virus (HCV) related chronic liver disease were registered for the study. 135 patients were suffered from HCC, 116 Cirrhosis of the liver patients without HCC and 64 patients with chronic hepatitis C.

The dataset comprises 29 features that encompass demographic features, haematological features, biological features, viral markers with additional clinical features. The proposed decision tree algorithm was able to predict HCC instances with an accuracy of 82.2%, sensitivity (recall) of 83.5% and specificity of 83.3%. The decision tree model predicts serum AFP as the foremost feature for HCC prediction. Their study reveals that male patients are 2.9 times more prone to develop HCC as compared to female patients. Liang et al. (2016) proposed biomarkers for early prediction of HCC. They proposed metabolic profiling, multivariate data exploration, machine learning method, pathway examination and ROC for the analysis of HCC. The proposed model with identified biomarkers achieves an accuracy of 83%, sensitivity of 96.50% and specificity of 83%. Comparison between Artificial Neural Network (ANN) and Logistic Regression (LR) based model was proposed by Chiu et al. (2013) for prediction of significant mortality attributes for HCC. Their ANN model shows better performance than LR model. The clinical information consisting of 21 features, that includes demographics and hepatic biochemical parameters of patients, was collected from 434 patients at Kaohsiung Medical University Hospital and Yuan's Hospital, Taiwan. Comorbidity, liver cirrhosis, α-Fetoprotein, platelet, ASA classification, and TNM stage were predicted as highest significant features by ANN model. Their ANN model predicts an accuracy of 85.10%. Another ANN based HCC prediction model was proposed by Liu et al. (2020). Their model comprises of 39 features (10 patient related features, 03 HBV-related features, 19 laboratory data related features, 03 tumour-related features, 04 subsets of BCLC staging) and 3 target features. Their model predicts AUROC of 87.70%.

Machine learning based HCC survival prediction models were proposed by different authors. Dong et al. (2019) presents model for survival of HCC patients based on DNA methylation and machine learning. Cox regression as well as Support Vector Machine (SVM)-Recursive Feature Elimination (RFE) algorithm and forward-SVM algorithms were proposed to screen differently methylated sites. The proposed SVM-RFE model obtain tenfold cross-validation score of 0.50 and FW-SVM obtain tenfold cross-validation score of 0.95. The model predicted the best score with 134 best selected features. Shi et al. (2012) introduces prediction model for mortality after liver cancer surgery. Comparative analysis of Artificial Neural Network (ANN) and Logistic Regression (LR) Models were presented and the result shows that ANN performs better than LR in terms of accuracy, Hosmer–Lemeshow (H–L) statistics and AUROC curves. ANN model attains accuracy of 97.28%, H–L Statistics of 41.18% and AUROC curve value of 84.67% as compared to LR model with accuracy of 88.29%, H–L statistics of 54.53% and AUROC curve value of 76%. An unsupervised cluster-based survival prediction model of HCC patients was used by Santos et al. (2015). Neural Network (NN) and Logistic Regression (LR) classifiers were used for prediction in terms of accuracy, AUC and F-Measure. NN demonstrates better performance than LR classifier. NN achieves an accuracy of 75.2%, AUC of 70% and F- Measure score as 0.665. Inclusion of penalty function to the existing firefly algorithm for HCC prediction was projected by Sawhney et al. (2018). Firefly algorithm with penalty function was used to evaluate the most optimal subset of features. The optimal subset of features Random Forest classifier with optimum subset of features was used HCC classification. The number of subset of features were reduced to eight and accuracy of 83% was attained with RF classifier. Tuncer and Ertam (2019) used Neighbourhood Component Analysis (NCA) and reliefF methods for feature reduction. Twenty-three traditional machine learning classifiers were used for HCC prediction. The result was predicted in terms of accuracy, precision, recall and F-1 score. An accuracy of 92.12% and 83.03% was obtained for NCA and reliefF based methods.

A gene-based study for discriminating HCC from cirrhosis tissues was proposed by Zhang et al. (2020). Machine learning approach was applied on microarray data having 1091 HCC samples and 242 without HCC samples. Within-sample relative expression ordering (RECs) technique was implemented to draw out numerical descriptors. Maximum redundancy minimum relevance (mRMR) feature extraction was implemented to obtain significant gene pairs. Classification of gene pairs was obtained using Support Vector Machine classifier. Their proposed model obtained eleven most significant gene pairs with excellent classification result. These investigated gene pairs can be expressed as signature for HCC. The obtained signature gene pairs were:—TRMT112-SF3B1; MFSD5-COLEC10; FDXR-APC2; LAMC1-CHST4; UBE4B-HGF; NCAPH2-APC2; HSPH1-MTHFD2; TMEM38B-AGO3; PLGRKT-COLEC10; HNF1A-APC2; ARPC2-SF3B1. Their model had significant advantage that it can discriminate HCC and non-HCC samples for minimum biopsy specimens and even for inaccurately samples specimens.

In the present work, novel hybrid models using, LASSO Regression, Ridge Regression, Genetic Algorithm optimization and Random Forest with three machine learning classifiers are proposed for HCC prediction. The proposed method consists of data pre-processing, feature selection/optimization and classification. The main contributions of this paper are:

- For performance evaluation, LASSO Regression, Ridge Regression, Genetic Algorithm optimization and Random Forest based feature evaluators with machine learning classifiers are presented.
- Excellent performance results in terms of accuracy, recall, F-1 score, Jaccard score and AUROC.
- The proposed methodology is compared with existing methodology and proposed methodology shows improved performance existing methods.

2 Material and proposed methodology

The proposed methodology is shown in Fig. 2. In first step, HCC liver cancer survival dataset, available at UCI data repository (UCI 2020), was selected for analysis. The data was collected at University Hospital in Portugal (Santos et al. 2015). The dataset contains 49 feature values obtained from 165 patients diagnosed with HCC. Step 2 performs the data pre-processing of HCC survival dataset. Step 3 performs feature weight assignment for assessing the feature importance. Step 4 performs the model implementation with significant features. Step 5 measures the performance of proposed model in terms of Accuracy, Recall, F1 Score, Log-Loss Score, Jaccard Score and AUROC.

2.1 Dataset

The dataset encompasses analysis of demographic, risk factor, laboratory and overall survival features from 165 patients diagnosed with HCC. The dataset covers 49 features for prediction of survival of HCC patients. The dataset consists of clinical attributes that are considered to be notable for clinical decision process. The clinical attributes

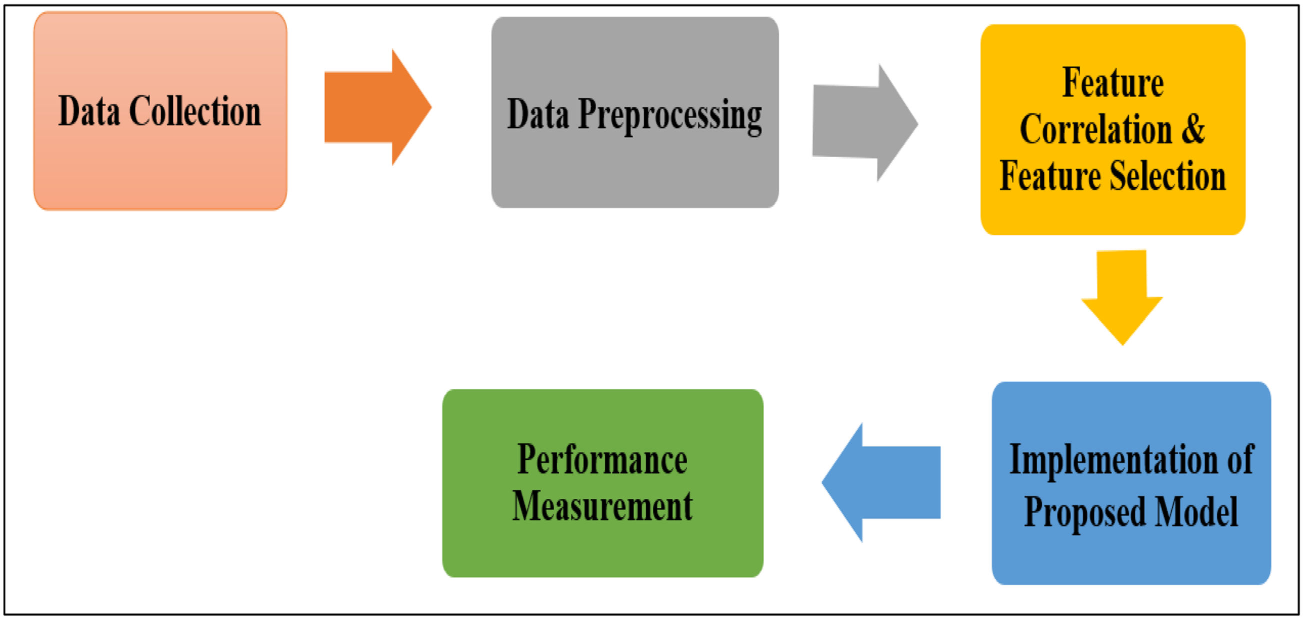

Data Collection
Data Preprocessing
Feature Correlation & Feature Selection
Implementation of Proposed Model
Performance Measurement

Fig. 2 Proposed methodology

considered for analysis are: Gender, Symptoms, Alcohol, HBsAg (Hepatitis B Surface Antigen), HBeAg (Hepatitis B e-Antigen), HBcAb (Hepatitis B Core Antibody), HCVAb (Hepatitis C Virus Antibody), Cirrhosis, Endemic Countries, Smoking, Diabetes, Obesity, Hemochromatosis, Arterial Hypertension, Chronic Renal Insufficiency, Human Immunodeficiency Virus, Non Alcoholic Steatohepatitis, Esophageal Varices, Splenomegaly, Portal Hypertension, Portal Vein Thrombosis, Liver Metastasis, Radiological Hallmark, Age at diagnosis, Grams of Alcohol per day, Packs of cigarettes per year, Performance Status, Encefalopathy degree, Ascites degree, International Normalised Ratio, Alpha-Fetoprotein (ng/mL), Haemoglobin (g/dL), Mean Corpuscular Volume (fl), Leukocytes(G/L), Platelets (G/L), Albumin (mg/dL), Total Bilirubin(mg/dL), Alanine transaminase (U/L), Aspartate transaminase (U/L), Gamma glutamyl transferase (U/L), Alkaline phosphatase (U/L), Total Proteins (g/dL), Creatinine (mg/dL), Number of Nodules, Major dimension of nodule (cm), Direct Bilirubin (mg/dL), Iron (mcg/dL), Oxygen Saturation, Ferritin(ng/mL). The target variable is encoded with value of 0 (patient did not survive) and 1 (patient survived). The description of qualitative input variables and quantitative input variables are presented in Tables 1 and 2 respectively.

Table 1 Description of qualitative (nominal) input attributes

Sr. no.	Input variables	Range	Mean/mode	Sr. no.	Input variables	Range	Mean/mode
1	Gender	0/1	1	13	Hemochromatosis	0/1	0
2	Symptoms	0/1	1	14	Arterial Hypertension	0/1	0
3	Alcohol	0/1	1	15	Chronic Renal Insufficiency	0/1	0
4	HBsAg (Hepatitis B Surface Antigen)	0/1	0	16	Human Immunodeficiency Virus	0/1	0
5	HBeAg (Hepatitis B e Antigen)	0/1	0	17	Non-Alcoholic Steatohepatitis	0/1	0
6	HBcAb (Hepatitis B Core Antibody)	0/1	0	18	Esophageal Varices	0/1	1
7	HCVAb (Hepatitis C Virus Antibody)	0/1	0	19	Splenomegaly	0/1	1
8	Cirrhosis	0/1	1	20	Portal Hypertension	0/1	1
9	Endemic Countries	0/1	0	21	Portal Vein Thrombosis	0/1	0
10	Smoking	0/1	1	22	Liver Metastasis	0/1	0
11	Diabetes	0/1	0	23	Radiological Hallmark	0/1	1
12	Obesity	0/1	0				

Table 2 Description of quantitative (numeric) input attributes

Sr. no.	Input variables	Range	Mean/mode	Sr. no.	Input variables	Range	Mean/mode
1	Age at diagnosis	20–93	64.69	14	Total Bilirubin (mg/dL)	0.3–40.5	3.09
2	Grams of alcohol per day	0–500	71.01	15	Alanine transaminase (U/L)	11–420	67.09
3	Packs of cigarettes per year	0–510	20.46	16	Aspartate transaminase (U/L)	17–553	69.38
4	Performance status	0,1,2,3,4	0	17	Gamma glutamyl transferase (U/L)	23–1575	268.03
5	Encefalopathy degree	1,2,3	1	18	Alkaline phosphatase (U/L)	1.28–980	212.21
6	Ascites degree	1,2,3	1	19	Total Proteins (g/dL)	3.9–102	8.96
7	International normalised ratio	0.84–4.82	1.42	20	Creatinine (mg/dL)	0.2–7.6	1.13
8	Alpha-fetoprotein (ng/mL)	1.2–1,810,346	19,299.95	21	Number of nodules	0–5	2.74
9	Haemoglobin (g/dL)	5–18.7	12.88	22	Major dimension of nodule (cm)	1.5–22	6.85
10	Mean corpuscular volume (fl)	69.5–119.6	95.12	23	Direct bilirubin (mg/dL)	0.1–29.3	1.93
11	Leukocytes(G/L)	2.2–13,000	1473.96	24	Iron (mcg/dL)	0–224	85.6
12	Platelets (G/L)	1.71–459,000	113,206.44	25	Oxygen saturation	0–126	37.03
13	Albumin (mg/dL)	1.9–4.9	3.45	26	Ferritin (ng/mL)	0–2230	439

2.2 Data pre-processing

The dataset contains 49 attributes with 23 attributes having nominal value and 26 attributes with continuous values as presented in Tables 1 and 2 respectively. Initially, 10.22% data is missing in the entire dataset. Santos MS et al. in 2015 observed that missing value imputation can be carried out using KNN with different values of k and with k = 1 gives the best fit for the missing values for the given dataset. It was further evaluated by Beretta and Santaniello (2016) that missing value imputation using KNN for any value of k > 1, standard deviations are significantly affected and inflated, hence KNN with k = 1 outperformed. So, KNN with k = 1 using Heterogeneous Euclidean Overlap Metric (HEOM) Distance is used for missing value imputation.

HEOM was described by Wilson and Martinez (1997, 2000) as an example of a heterogeneous distance measure. Suppose, we wish to find distances between some subset of n objects and that for each object we have measured the values of R predictors, Let J = {1, 2,..., n} be an index set for each of the n objects. For each i, j ∈ J, the HEOM defines the distance between the ith object and the jth object as

$$HEOM(P_i, P_j) = \sum_{r=1}^{R} d_r(P_{i,r}, P_{j,r}) \tag{1}$$

where

$$d_r\left(P_{i,r}, P_{j,r}\right) = \begin{cases} \frac{\left|P_{i,r} - P_{j,r}\right|}{range_r} \; if\ r\ indexes\ a\ continuous\ attribute \\ \delta_{i,j}\ if\ r\ indexes\ a\ categorical\ attribute \end{cases} \tag{2}$$

And $\delta_{i,j} = 1$ if $P_{i,r} \neq P_{j,r}$ and $\delta_{i,j} = 0$ if $P_{i,r} = P_{j,r}$. Here, $d_r(P_{i,r}, P_{j,r})$ can be thought of as the contribution of the rth attribute to the overall distance and $range_r = \max_{j \in J}\left\{P_{j,r}\right\} - \min_{j \in J}\left\{P_{j,r}\right\}$. Notice that a continuous attribute's contribution to the HEOM distance is bounded above by 1.

The HCC dataset containing clinical attributes of 63 patients (dead) with target value encoded as 0 and clinical attributes of 102 patients (alive) with target value encoded as 1. The number of instances of dead and alive cases illustrates certain grade of class disproportion. Synthetic Minority Over-sampling Technique (SMOTE) proposed by Nitesh et al. (2002) is pertain to remove the class disproportionate. SMOTE is an oversampling approach used to attain quasi samples from minority class. This oversampling technique had varied applications in different areas (Sharma 2019; Fallahi and Jafari 2011; Liu et al. 2006; MacIsaac et al. 2006) with k = 3 as nearest neighbour value, SMOTE generate 204 pseudo samples with 102 instances each for target value.

2.3 Feature correlation and feature importance

The HCC dataset implemented in proposed machine learning model has 49 attributes with one target class. Certain features could be strongly correlated with other features. So, it is worth to eliminate one feature from these highly correlated features. Figure 3 shows the correlation amongst features and correlation of features with target class. The features with dark colour in figure show strong positive correlation and features with low colour shows negative correlation. One of the features from two features having high correlation can be eliminated as they have same consequence on the target class.

LASSO Regression, Ridge Regression, Genetic Algorithm optimization and Random Forest are proposed for feature evaluation and feature elimination.

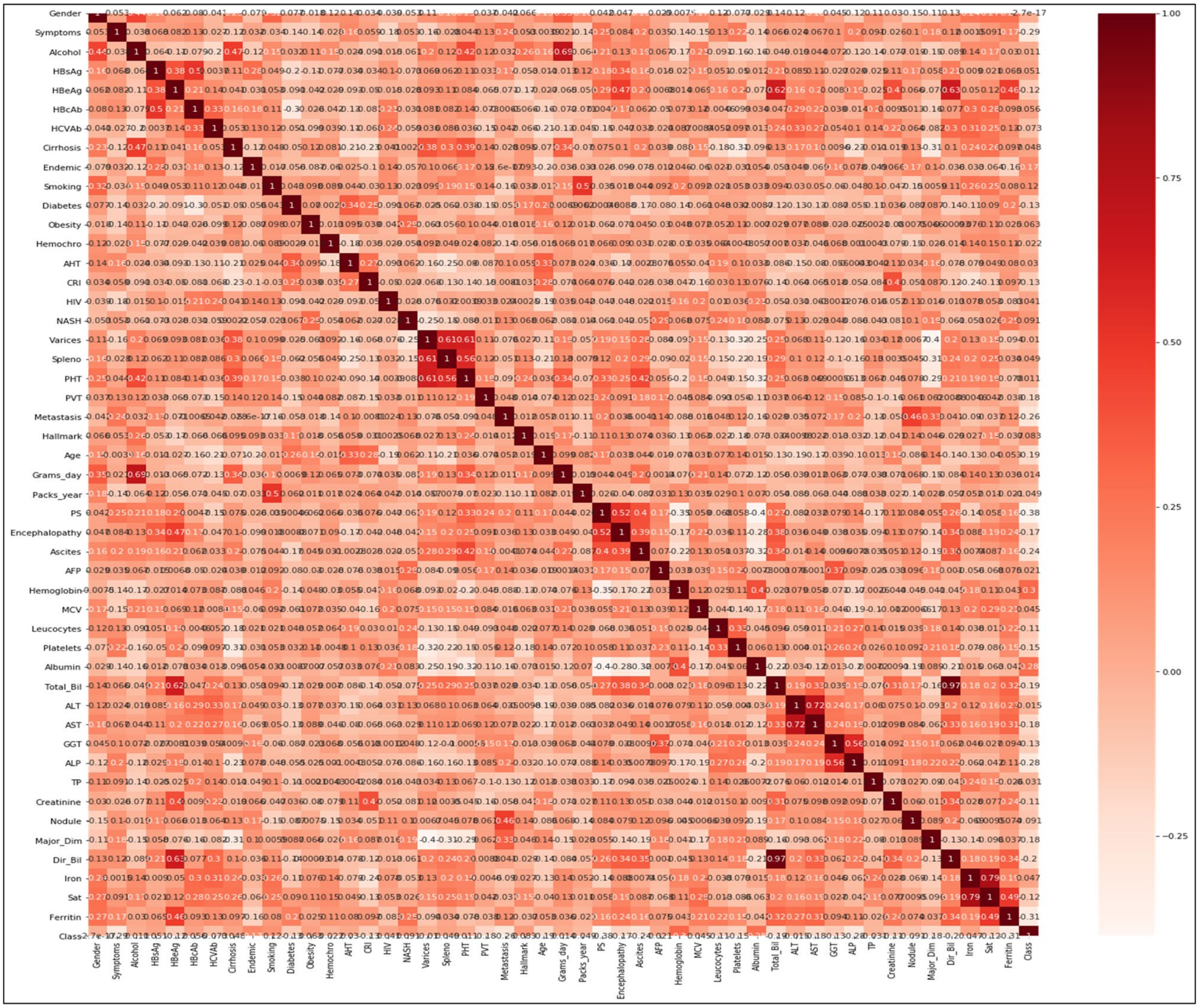


Fig. 3 Heatmap showing the correlation of feature variables and target class

2.4 Model implementation using random forest and gradient boosting hybrid approach

In this paper, Random Forest and Gradient Boosting hybrid approach is proposed for survival prediction of HCC. Weight is assigned to individual feature using proposed feature evaluation technique and significant features having high weight value are selected for prediction. Initially, the dataset has 49 features for survival prediction of HCC.

2.4.1 Random forest approach

Random Forest utilize construction of multiple trees (Breiman 2001). While constructing the tree, RF explore random subset of input variables at each division of node and the tree matures fully without pruning. Due to random selection of variables at each node, the correlation among the tree in forest decreases and hence the forest rate decreases (Hideko and Hiroaki 2012). Tree progression in RF can be given as:

- At node N, randomly sample R from the given the independent variables Q.
- For every random sampled variable (D = 1,2,3……R), estimate the best split A_D amongst all the probable splits for Dth variable.
- Select the optimum split A_O among D = 1,2,3……R, best splits A_D.
- This Jth variable at its recognized cut point CA_O is used to divide the node N.
- Now, split the data at this node by sending the P = 1,2,3…. T observations with $Y_{TJ} < CA_O$ to the left

descendant and all the observations with $\mathrm{Y_{TJ}} > \mathrm{CA_O}$ to the right descendant.

- Repeat the steps till the tree matures.

Gini Importance approach is implemented to select the split with lowest impurity at each node. For each node N in decision tree, the split is estimated by the decrease in Gini impurity $\Delta GI(N)$. Whereas, Gini impurity is given as

$$\Delta GI(N) = S_T \Delta I(N) - S_L \Delta I\left(N_L\right) - S_R \Delta I\left(N_R\right) \tag{3}$$

where $\Delta I(N)$ is known as Gini Index and can be given as

$$\Delta I(N) = 1 - \sum_k r(k|N)^2 \tag{4}$$

where, $r(k|N)$ is the rate at which target class k is discriminated correctly at node N;

$\Delta I\left(N_L\right)$ and $\Delta I\left(N_R\right)$ are the Gini Index on the left side and right side of the node respectively; S_T is the number of samples before split; S_L and S_R are the number of samples on left and right side of node after split.

Gini Importance can be obtained from average of all the decrease in Gini Impurity. Simulation parameters for RF approach are given in Table 3.

Figure 4 shows the relative feature importance of all the input attributes. It has been observed that 19 features have significant impact on result prediction. The significant features are:—'Age at diagnosis', 'Performance Status', 'Alpha-Fetoprotein', 'Haemoglobin', 'Mean Corpuscular Volume', 'Leukocytes', 'Platelets', 'Albumin', 'Total Bilirubin', 'Aspartate transaminase', 'Gamma glutamyl transferase', 'Alkaline phosphatase', 'Total Proteins', 'Creatinine', 'Major dimension of nodule', 'Direct Bilirubin', 'Iron', 'Oxygen Saturation', and 'Ferritin'.

The proposed Random Forest approach predicts 'Alpha-Fetoprotein', 'Hemoglobin', 'Ferritin' and 'Alkaline phosphatase' as most significant factors.

Figure 5 shows the heatmap of features selected by proposed Random Forest feature selector. Simulation results shows that higher accuracy can be achieved with selected features.

Table 3 Simulation parameters for regression and optimisation models

Model	Simulation parameters
LASSO regression (L-1 Penalization)	Alpha = 1.0, number of iterations = 1000, selection = cyclic, tolerance for optimization = 0.0001
Ridge regression (L-2 Penalization)	Alpha = 1.0, solver = auto, number of iterations = 1000, tolerance for optimization = 0.001
GA optimized models	Maximum features = 20, number of populations = 120, crossover probability = 0.5, mutation probability = 0.2, no. of generations = 50, crossover independent probability = 0.5, mutation independent probability = 0.05
RF feature selector model	Cost-Complexity Pruning = 0, class weight = 1, criterion = Gini, max no. features = sqrt (no. of features), maximum no. of leaf nodes = unlimited, minimum threshold for early stopping = 0, minimum no. of samples for leaf node = 1, minimum no. of samples essential to fragmented an internal node = 2, no. of estimators = 100

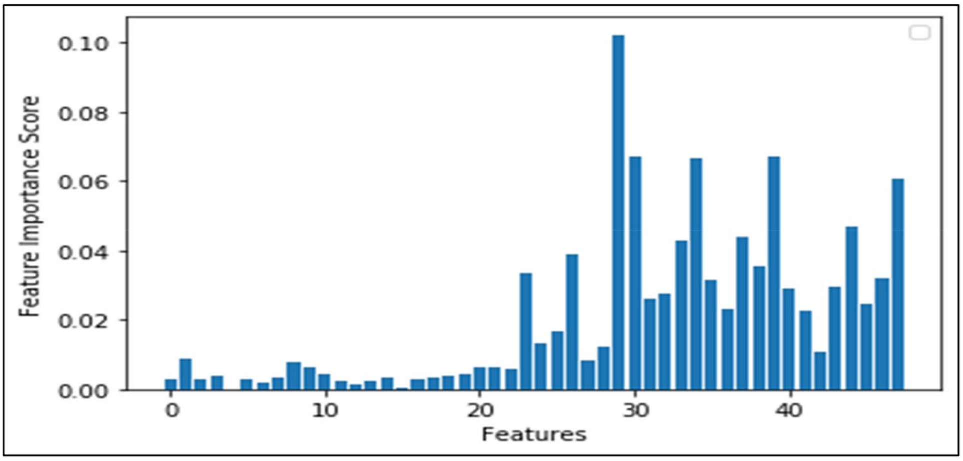


Fig. 4 Relative feature importance score using Random Forest feature selector

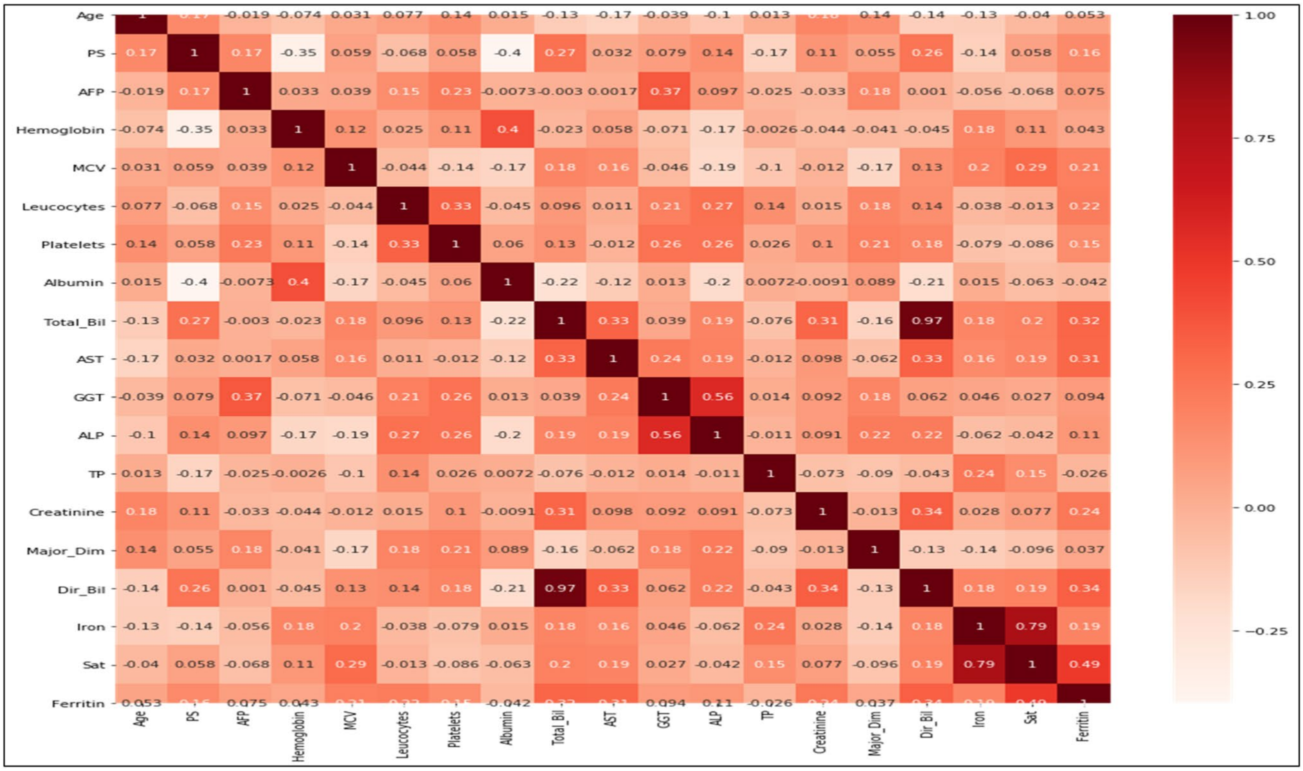


Fig. 5 Heatmap of variables selected by Random Forest feature selector

2.4.2 Gradient boosting approach

Gradient Boosting is non-parametric algorithm proposed by Friedman (2001). The objective of GB algorithm is to sequentially build each decision tree model on the gradient descent direction of a loss function. Each supplement base model is intended to correct the errors made by its preceding base models. The loss function defines the accuracy of the models. Greater is the loss function, worst is the prediction accuracy of model. Prediction accuracy can be increased if the loss function decreases with each supplement of new base model. The probable method is to let the value of the loss function deteriorate in the direction of its gradient descent. The pseudo code for GB is as follows:

```
Inputs:
➢   Training Data $(x, y)_{i=1}^{N}$
➢   No. of iterations: Boosting Stages;
➢   Loss-Function: $\mathcal{L}(y_i, \rho)$;(proposed: Deviance);
➢   Base-Prediction Model $h\,(x, \vartheta)$;
Training:
i) Initialize the base model $f_0(x)$;
  $f_0(x) = argmin_{\rho} \sum_{i=1}^{N} \mathcal{L}\,(y_i, \rho)$;        (5)
ii) for k=1 to Boosting Stages do
for i=1 to N do
estimate the negative gradient
$\mathbb{Z}_i = -\left[\frac{\partial \mathcal{L}\,(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{k-1}(x)}$   $i = 1,2,3 \ldots .N$        (6)
end;
iii)   construct a new prediction model $h(x; \Phi)$;
iv)   evaluate the gradient descent step size as
$\rho_t = argmin_{\rho} \sum_{i=1}^{N} \mathcal{L}\,(y_i, f_{k-1}(x_i) + \rho\, h(x_i; \Phi))$        (7)
v) Update the function estimate as
 $f_k(x) = f_{k-1}(x) + \rho_t h(x; \Phi)$        (8)
end for;
```

In step *i),* base model $F_0(x)$is initialized by GB. In step *ii),* for *k* boosting stages it trains *k* models using *for* loop. Increasing the boosting stages *k* diminishes the error on the training set, but very high values of *k* leads to the problem of overfitting. Using Loss-Function, for every trained/imperfect model *k,* the value of negative gradient $\mathbb{Z}_i$ is calculated according to already trained *k-1* models. *Step iii)* constructs new prediction model $h(x;\Phi)$ and attain its parameter Φ by fitting it to the $\mathbb{Z}_i$. Mean square method is used to achieve the minimum value in gradient direction. *Step iv)* estimates the gradient descent step size of the new model using the loss function. *Step v)* updates the model using prediction model $h(x;\Phi)$. The proposed GB model use deviance as loss function and 100 boosting stages.

3 Performance evaluators

For the proposed model, prediction performance is measured in terms of Accuracy (%), Recall (%), F1 Score, Log-Loss Score and Jaccard Score. The performance evaluators can be defined as:

$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}}$; it defines the accurately predicted number of test events from the total number of test events.

$\boldsymbol{R}ecall = Sensitivity = TruePositiveRate(TPR) = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$; it signifies the number (%age) of correct positive prediction from total number of positives. Value of 1 (100%) indicate as best sensitivity and 0 (0%) indicate worst sensitivity.

where, True Positive (TP) signifies correct positive prediction; False Positive (FP) indicate incorrect positive prediction; False Negative (FN) indicate incorrect negative prediction and True Negative (TN) signifies correct negative.

3.1 F1 Score

It can be defined as weighted average of the precision (PPV) and recall (TPR). A model with F1 score of 1 is assumed to be its best value and 0 to be its worst value. Mathematically, F1 score is given as:

$$\text{F1 Score} = 2 * [\text{Precision (PPV)} * \text{Recall (TPR)}] / [\text{Precision (PPV)} + \text{Recall (TPR)}].$$

3.2 Log-loss score

This score is demarcated on probability approximations. This is also known as cross-entropy loss score. Instead of defining discrete predictions, this score is used to evaluate probability outputs. Mathematically, Log-Loss for a binary classifier can be defined as:

$$\text{Log Loss} = \sum_{j=1}^{M} y_j \log(p_j) + (1 - y_j)\log(1 - p_j) \tag{9}$$

where p_j is the likelihood/probability that the jth data point fits to class "1" as forecast by the classifier and y_j is the actual class can be either "0" or "1". It evaluates the uncertainty of the likelihoods of proposed model by equating them with true label. The accuracy of classifier can be maximized by minimizing the log loss score. A good model should have a small log loss score. The lower is the log loss better is the prediction.

3.3 Jaccard score

Jaccard score calculates the average value of Jaccard Similarity Coefficients (JSC) amongst pairs of label sets.

Mathematically, JSC can be calculated as

$$J(T_i, P_i) = \frac{|T_i \cap P_i|}{|T_i \bigcup P_i|} = \frac{|T_i \cap P_i|}{|T_i| + |P_i| + |T_i \cap P_i|} \tag{10}$$

where, T_i is the actual truth label set and P_i is the predicted label set. A good model should have a high Jaccard Score.

4 Performance evaluation of models

Serval experiments have been conducted using distinct penalized and optimization techniques on proposed machine learning algorithms for prediction of HCC survival. To evaluate the accuracy and other prediction parameters, the data set is divided randomly into training dataset (80%) which is used to build the model and test dataset (20%) to test the model. L-1 Penalized Nu-SVC Model, L-1 Penalized GBEL Model, L-1 Penalized RidgeCV Model, L-2 Penalized Nu-SVC Model, L-2 Penalized GBEL Model, L-2 Penalized RidgeCV Model, GA Optimized Nu-SVC Model, GA Optimized GBEL Model, GA Optimized RidgeCV Model, RF-Nu-SVC Model, RFGBEL Model and RF-RidgeCV Model are tested for envisaging the result. The experiments were simulated using Python 3.8 on an IBM PC with Intel Core i-7–6700 CPU @ 3.40 GHz processor with 8 GB RAM. Table 3 shows simulation parameters for L-1 Penalized, L-2 Penalized, Genetic Algorithm Optimized and RF models. The simulation parameters have their usual meanings. The performance of each classifier is measured in terms of Accuracy (%), Recall (%), Precision (%), F1 Score, Log-Loss Score and Jaccard Score.

4.1 Nu-support vector classification (Nu-SVC) hybrid model

Nu-SVC is analogous to SVC except that in Nu-SVC the number of support vectors can be specified. The Nu denotes the numeral values of samples that act as support vectors but lie on the wrong side of the hyperplane. It represents the limit of higher bound on the segment of training errors and lower bound of segment support vectors. Nu-SVC model with simulation parameters: Nu = 0.5; class weight = 1; degree of the polynomial kernel = 3; kernel = RBF, hard limit on iterations = -1 and tolerance for stopping criteria = 0.001 is proposed for prediction of HCC. Table 4 presents performance analysis of Nu-SVC classifier with different penalization and optimization techniques. It was found that, L-2 penalize and GA optimized Nu-SVC models attain lowest Log-Loss score as 15.16 and highest accuracy as 56.09%. Except GA optimized model, all models evaluate F-1 score as 0.63. Nu-SVC, L-1 penalized Nu-SVC, and RF-Nu-SVC predicts 100% recall/sensitivity value which signifies that these models are more efficient in correct positive prediction from the total number of positive values.

Nu-SVC, L-1 Penalized Nu-SVC Model and RF-Nu-SVC Model predict same accuracy, recall, F-1 score, Log-Loss score and Jaccard score. GA optimized Nu-SVC demonstrates better performance in contrast to other Nu-SVC models in terms of accuracy, F-1 score, Log-Loss score and Jaccard score. Nu-SVC Model, L-1 Penalized Nu-SVC Model and RF-Nu-SVC Model demonstrates same outcome for all the performance matrices. As the Nu-SVC model acquires high value of Log-Loss which indicates the poor performance in terms of accuracy.

4.2 RidgeCV hybrid model

RidgeCV classifiers are based upon Ridge regression classifiers that perform cross validation. The proposed model uses tenfold cross validation. RidgeCV Model with simulation parameters Alpha array = ([0.1, 1., 10.]), cv = 10, class weight = 1 and maximum number of iterations = 1000 is proposed for result prediction. Table 5 shows the predicted results for Hybrid RidgeCV models. L-1 and L-2 penalized RidgeCV models predict same result in terms of accuracy (63.41%) and log loss (12.63). RidgeCV and RF-RidgeCV models predict the same value of accuracy (65.85%) and Log-Loss (11.79). Similar F-1 score (0.63) was evaluated by L-1 Penalized RidgeCV Model and RF-RidgeCV Model. GA optimized RidgeCV model shows superior performance as compared to other RidgeCV hybrid models in terms of accuracy, recall, F-1 score, Log-Loss score and Jaccard score.

It predicts the result with highest accuracy of 68.29%, recall/sensitivity of 73.68, Jaccard score of 0.51, F-1 score of 0.68 with minimum Log-Loss score of 10.95 amongst all RidgeCV hybrid models. All proposed RidgeCV models show improved performance as compared to Nu-SVC models and they obtain low Log-Loss score as compared to Nu-SVC models which indicated its better performance than Nu-SVC models.

Table 4 Prediction performance measure of Nu-SVC models

Model	Accuracy (%)	Recall/ sensitivity (%)	F-1 score	Log-loss score	Jaccard score
Nu-SVC model	46.34	100	0.63	18.53	0.46
L-1 penalized Nu-SVC model	46.34	100	0.63	18.53	0.46
L-2 penalized Nu-SVC model	56.09	84.21	0.63	15.16	0.47
GA optimized Nu-SVC model	56.09	94.73	0.66	15.16	0.50
RF-Nu-SVC model	46.34	100	0.63	18.53	0.46

Table 5 Prediction performance measure of RidgeCV models

Model	Accuracy (%)	Recall/ sensitivity (%)	F-1 score	Log-Loss score	Jaccard Score
RidgeCV Model	65.85	73.68	0.66	11.79	0.50
L-1 penalized RidgeCV model	63.41	68.42	0.63	12.63	0.46
L-2 penalized RidgeCV model	63.41	63.15	0.61	12.63	0.44
GA optimized RidgeCV model	68.29	73.68	0.68	10.95	0.51
RF-RidgeCV model	65.85	63.15	0.63	11.79	0.46

Table 6 Prediction performance measure of GBEL models

Model	Accuracy (%)	Recall/sensitivity (%)	F-1 score	Log-loss score	Jaccard score
GBEL model	70.73	73.68	0.70	10.10	0.53
L-1 penalized GBEL model	75.60	89.47	0.77	8.42	0.62
L-2 Penalized GBEL Model	75.60	78.94	0.75	8.42	0.60
GA optimized GBEL model	73.17	78.94	0.73	9.26	0.57
RFGBEL model	93.92	94.73	0.93	5.89	0.72

4.3 Gradient boosting ensemble learning (GBEL) hybrid model

Gradient Boosting Ensemble Learning Model with simulation parameters: Cost-Complexity Pruning = 0; evaluation Criterion = Friedman Mean Square Error; Learning Rate = 0.1; Loss Function = Deviance; Validation Fraction = 0.1 and Boosting Stages = 100 is proposed for HCC survival result prediction.

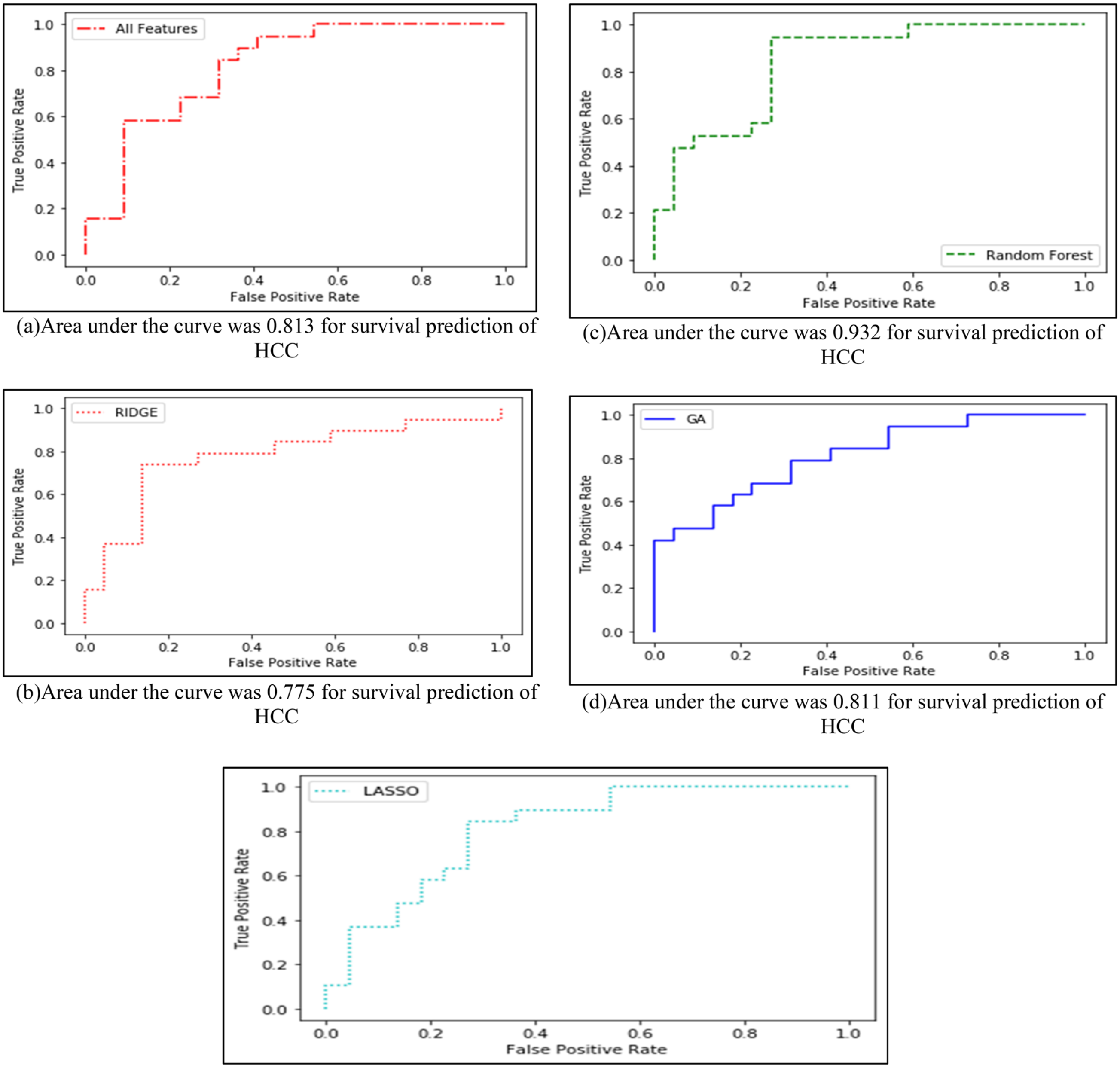
The result presented in Table 6 shows that RFGBEL model predicts excellent results for all performance metrices. RFGBEL model predicts the HCC results with an accuracy of 93.92%. The RFGBEL model shows a significant improvement of 14.83–23.19% in accuracy as compared to other Gradient Boosting models. The RFGBEL model attain minimum cross entropy loss score. It obtains a cross entropy loss score of 5.89 as compared to highest cross entropy loss score of 10.10 obtained by GBEL model. The RFGBEL model shows a significant improvement in cross entropy loss of 2.53–4.21 amongst other proposed GBEL models. RFGBEL model shows excellent result in terms recall, F-1 score and Jaccard score. It estimates 94.73% recall, F-1 score of 0.93 and Jaccard score of 0.72. It shows significant improvement of 21.05% in recall, 0.19 improvement in Jaccard score and improvement of 0.16 in F-1 score as compared to other GBEL models.

5 Discussion

GBEL models (Table 6) particularly RFGBEL model demonstrates excellent performance for HCC prediction in contrast to Nu-SVC models (Table 4) and RidgeCV models (Table 5). Nu-SVC hybrid models shows unfavourable performance in terms of accuracy, F-1 score, Log-Loss score and Jacard score. It shows minimal performance, in contrast to RidgeCV and GBEL models, with average accuracy, F-1 score, Log-Loss score and Jacard score of 50.24%, 0.63, 17.18 and 0.47 respectively. However, RidgeCV hybrid models achieve improved performance in comparison to Nu-SVC hybrid models. RidgeCV hybrid models predict the HCC with an average accuracy, F-1 score, Log-Loss score and Jacard score of 65.36%, 0.64, 11.95 and 0.47 respectively. RidgeCV models shows an average improvement of 15.12%, and 5.23 (decrease) in accuracy and Log-Loss score respectively in contrast to Nu-SVC model. Nu-SVC and RidgeCV models foresees identical average jacard score. GBEL models accomplish excellent results for HCC prediction in terms of accuracy, recall, F-1 score, Log-Loss score and Jaccard score. The RFGBEL model outperform in terms of all performance measurement matrices. It classifies the survival and non-survival HCC samples with an accuracy and sensitivity of 93.92% and 94.73% respectively. The RFGBEL model also achieves excellent result in terms of F-1 score, Log-Loss score and Jaccard score with their values as 0.93, 5.89 and 0.72, respectively. In contrast to the average score of Nu-SVC and RidgeCV models, RFGBEL model shows significant improvement of 28.56 to 43.63% in accuracy, 0.29–0.30 in F-1 Score, 6.06 to 16.25 decrease in Log-Loss/Cross entropy loss and 0.25 improvement in Jaccard score.

The prediction performance of RFGBEL model is also tested in terms of AUROC. Figure 6 shows the AUC for Gradient Bosting Model, L-1 Penalized GBEL model, L-2 Penalized GBEL Model, GA optimized GBEL model and RFGBEL Model. AUROC curve is a plot between TPR and False Positive Rate (FPR) i.e. sensitivity against specificity. Area Under Curve can be computed by aggregating the area under the ROC curve. The larger is the area, the more accurate is the prediction (Bowers and Zhou 2019). The RFGBEL model computes value of AUC as 0.93. The RFGBEL model computes highest value of area under the curve amongst other proposed method. The highest value of AUC for RFGBEL model validate accurate result prediction by RFGBEL model. The high value of sensitivity predicted by RFGBEL model indicates its ability to correctly predict the positive cases.

The comparison result of proposed method with existing methods are presented in Table 7. Our proposed RFGBEL model predicts the result with an accuracy of 93.92%, F-1 score-0.93 and AUROC-0.93. The RFGBEL model predict Alpha-Fetoprotein as most significant factor. It can be clearly seen from the Table 7 that, our proposed RFGBEL method obtains much better performance than other methods. Using different machine learning algorithms, Tuncer et al. (2019)



(a)Area under the curve was 0.813 for survival prediction of HCC

(c)Area under the curve was 0.932 for survival prediction of HCC

(b)Area under the curve was 0.775 for survival prediction of HCC

(d)Area under the curve was 0.811 for survival prediction of HCC

(e)Area under the curve was 0.809 for survival prediction of HCC

Fig. 6 Receiver operating curves (ROC) for **a** GBEL model, **b** L-2 penalized GBEL model, **c** RFGBEL model, **d** GA optimized GBEL model, **e** L-1 penalized GBEL model

and Książek et al. (2019) proposed HCC prediction model in 2019. An Accuracy of 92.12%, recall of 91.20% and F-1 score 0.91 was obtained by Tuncer et al. (2019) and an accuracy of 88.49% and F-1 score of 0.87 was obtained by Książek et al. (2019). Chen et al. (2020) proposed HCC classification using three well known Support Vector Machine (SVM), Logistic Regression (LR) and Decision Tree (DT) classifiers. The missing values were handled by replacing them with median values. They obtained best prediction results with LR model with an accuracy of 73%, recall of 75%, F-1 score of 0.85.

Santos et al. (2015) proposed ANN with cluster based oversampling method for improving HCC survival prediction. Accuracy of 75.20%, F-1 score of 0.66 and AUROC of 0.70 was achieved with the proposed method. The proposed methodology was complex and time consuming as eleven different configurations (5–55 number of neurons in hidden layer in step of 5) with 30 runs, were performed to obtain optimal parameter. Elgin Christo et al. (2020) proposed co-operative coevolution approach was implemented for selection of relevant features and attributes. Random

Table 7 Comparison of proposed method with existing methods

Authors	Method	Accuracy (%)	Recall (%)	F-1 Score
Tuncer et al. (2019)	Feature weights were assigned using neighbourhood component analysis and reliefF for dimensionality reduction. The missing values were replaced by their mean values	92.12	91.20	0.91
Książek et al. (2019)	SVC classifier with 2 level genetic algorithm optimization approach and feature selection. Missing categorial attributes were replaced by modal values and missing numeric attributes were replaced by mean values	88.49	NA	0.87
Chen et al. (2020)	SVM, LR and Decision Tree classifiers for HCC prediction. Attributes with missing values were replaced with median values	73	75	0.85
Santos et al. (2015)	ANN + Augmented Sets Approach. Missing values were replaced with KNN imputation method with k = 1	75.20	NA	0.66
Elgin Christo et al. (2020)	Co-operative coevolution approach was used for attribute and instance selection. Attributes missing values were replaced using KNN imputation approach with k = 5 Random Forest classifier was used for result prediction	72.20	70	0.73
Rajesh et al. (2020)	KNN, Naive Bayes, Decision Tree, Random Forest and SVM classifiers Missing values were replaced by attributes median values	80.64	85	0.82
Kayal et al. (2019)	Deep Neural Network with four hidden layers for HCC prediction. The missing values were replaced by mean values and most frequent values	78	81.25	0.80
Proposed method	RFGBEL	93.92	94.73	0.93

Forest classifier was implemented for HCC prediction. Their obtained an accuracy of 72.20%, recall of 70% and F-1 score of 0.73. S. Rajesh et al. (2020) proposed five different machine learning algorithms for HCC classification. Using Random Forest classifier, they achieved highest results amongst all proposed classifier with an accuracy of 80.64%, recall of 85% and F-1 score of 0.82. HCC dataset used by Tuncer and Ertam 2019; Książek 2019; Chen et al. 2020; Elgin Christo et al. 2020; Rajesh et al. 2020) contains feature values of 102 alive patients and 63 dead patients. The dataset represents some degree of imbalance in the size of patient's profile and classification using imbalance data leads to result in favour of majority class (He and Garcia 2009; Daskalaki et al. 2006; Blagus and Lusa 2010; Hulse et al. 2007). HCC classification using Deep Neural Network (DNN) was proposed by Kayal et al. (2019). DNN with four hidden layers having 1024, 512, 256 and 128 neurons respectively in each layer was implemented. Their model obtained an accuracy of 78%, recall of 81.25% and F-1 score of 0.80. However, their model has disadvantage of selecting random number of neurons for each layer that was based on trial-and-error approach.

6 Conclusion

Distinguishing significant risk features/factors for HCC survival is of great importance in clinical medicine for screening and targeting the patients. The correct identification will help in development of better models to predict the results with more accuracy. Machine learning algorithms are capable of predicting performance parameters using static/dynamic data and the value of performance parameters can be maximized by using analytics and probabilistic models. In this study, HCC survival prediction model was developed using geographical factors, risk factors and clinical trial attributes. The suitable selection of input attributes/features advocate significant role in success of prediction model. The proposed model identifies 19 significant clinical features from 49 clinical attributes employing Random Forest approach. Proposed RFGBEL model demonstrates significant improvement in accuracy, recall, F-1 Score, Log-Loss/Cross entropy loss and Jaccard score as compared to NU-SVC, RidgeCV models. RFGBEL obtains more value of Area under curve (AUROC) as compared to other existing models. The study was compared with existing methods and it was suggested that RFGBEL model could be used as tool for predicting HCC survival using clinical attributes. It is also suggested that data mining techniques can be used as supplement tool for prognostic evaluation and clinical decision making.

Funding There is no funding source.

Availability of data Dataset is available at UCI data repository.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Beretta L, Santaniello A (2016) Nearest neighbor imputation algorithms: a critical evaluation. BMC Med Inform Decis Mak. https://doi.org/10.1186/s12911-016-0318-z

Blagus R, Lusa L (2010) Class prediction for high-dimensional class-imbalanced data. BMC Bioinform 11:523

Bowers AJ, Zhou X (2019) Receiver operating characteristic (ROC) area under the curve (AUC): a diagnostic measure for evaluating the accuracy of predictors of education outcomes. J Educ Stud Placed Risk (JESPAR) 24(1):20–46. https://doi.org/10.1080/10824669.2018.1523734

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Chen K-H, Wang H-W, Liu -M (2020) Applying artificial intelligence to survival prediction of Hepatocellular Carcinoma patients. In: Proceedings of the 2020 4th International Conference on Deep Learning Technologies (ICDLT). Pp: 135–139. https://doi.org/10.1145/3417188.3417197

Chiu H-C, Ho T-W, Lee K-T, Chen H-Y, Ho W-H (2013) Mortality predicted accuracy for Hepatocellular Carcinoma patients with hepatic resection using artificial neural network. Sci World J. https://doi.org/10.1155/2013/201976

Daskalaki S, Kopanas I, Avouris N (2006) Evaluation of classifiers for an uneven class distribution problem. ApplArtifIntell 20(5):381–417

Dhanasekaran R, Limaye A, Cabrera R (2012) Hepatocellular carcinoma: current trends in worldwide epidemiology, risk factors, diagnosis, and therapeutics. Hepat Med 4:19.

Dong R-Z, Yang X, Zhang X-Y et al (2019) Predicting overall survival of patients with hepatocellular carcinoma using a three-category method based on DNA methylation and machine learning. J Cell Mol Med 23:3369–3374. https://doi.org/10.1111/jcmm.14231

Elgin Christo VR, Khanna Nehemiah H, Brighty J, Kannan A (2020) Feature selection and instance selection from clinical datasets using Co-operative Co-evolution and classification using Random Forest. IETE J Res. https://doi.org/10.1080/03772063.2020.1713917

Fallahi A, Jafari S (2011) An expert system for detection of breast cancer using data pre-processing and bayesian network. Int J AdvSciTechnol 34:65–70

Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM (2010) Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer 127(12):2893–2971

Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, et al (2018) GLOBOCAN 2018, global and regional estimates of the incidence and mortality for 36 cancers global cancer observatory: cancer today. Lyon: International Agency for Research on Cancer. 2018; Available online at: https://gco.iarc.fr/today/fact-sheets-cancers. (Accessed on 23–05–2020)

Fitzmaurice C, Fitzmaurice C, Akinyemiju TF et al (2018) Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability adjusted life-years for 29 cancer groups, 1990 to 2016: a systematic analysis for the global burden of disease study. JAMA Oncol 4(11):1553–1568

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 10(29):1189–1232. https://doi.org/10.1214/aos/1013203451

Galle PR, Forner A, Llovet JM, Mazzaferro V, Piscaglia F, Raoul J et al (2018) EASL clinical practice guidelines: management of hepatocellular carcinoma. J Hepatol 69(1):182–236

He H, Garcia EA (2009) Learning from imbalanced data. IEEE Trans Knowledge Data Eng 21(9):1263–1284

Hideko K, Hiroaki Y (2012) Rapid feature selection based on random forests for high-dimensional data. IPSJ SIG Technical Reports. MPS 89(3): 1–7.

Hulse JV, Khoshgoftaar TM, Napolitano A (2007) Experimental perspectives on learning from imbalanced data. In: Proceedings of the 24th international conference on Machine learning. Corvallis, Oregon. Oregon State University. pp. 935–942.

Kayal CK, Bagchi S, Dhar D, Maitra T, Chatterjee S (2019) Hepatocellular carcinoma survival prediction using deep neural network. In: Chakraborty M, Chakrabarti S, Balas V, Mandal J (eds) Proceedings of International Ethical Hacking Conference 2018. Advances in Intelligent Systems and Computing, vol 811. Springer, Singapore. https://doi.org/10.1007/978-981-13-1544-2_28

Książek W, Abdar M, Acharya UR, Pławiak P (2019) A novel machine learning approach for early detection of hepatocellular carcinoma patients. CognitSyst Res 54:116–127

Liang Q, Liu H, Wang C, et al (2016) Phenotypic characterization analysis of human hepatocarcinoma by urine metabolomics approach. Sci Rep. 6: Article no. 19763.

Liu Y, Chawla NV, Harper MP, Shriberg E, Stolcke A (2006) A Study in Machine Learning from imbalanced data for sentence boundary detection in speech. Comput Speech Lang 20(4):468–494

Liu X, Hou Y, Wang X, Yu L, Wang X, Jiang L, Yang Z (2020) Machine learning-based development and validation of a scoring system for progression-free survival in liver cancer. HepatolInt 14(4):567–576. https://doi.org/10.1007/s12072-020-10046-w

MacIsaac KD, Gordon DB, Nekludova L, Odom DT, Schreiber J, Gifford DK, Young RA, Fraenkel E (2006) A hypothesis-based approach for identifying the binding specificity of regulatory proteins from Chromatin Immuno precipitation data. Bioinformatics 22(4):423–429

de Martel C, Maucort-Boulch D, Plummer M, Franceschi S (2015) World-wide relative contribution of hepatitis B and C viruses in hepatocellular carcinoma. Hepatology 62:1190–1200. https://doi.org/10.1002/hep.27969

Masaya S, Kentaro M, Shigeki K, Ryosuke T, Shuichiro S, Kazuhiko K et al (2019) Machine-learning approach for the development of a novel predictive model for the diagnosis of Hepatocellular Carcinoma. Sci Rep 1:7704. https://doi.org/10.1038/s41598-019-44022-8

Nitesh VC, Kevin WB, Lawrence OH, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J ArtifIntell Res 16:321–357

Njei B, Rotman Y, Ditah I, Lim JK (2015) Emerging trends in hepatocellular carcinoma incidence and mortality. Hepatology 61(1):191–199

Omran DA, Awad AH, Mabrouk MA et al (2015) Application of data mining techniques to explore predictors of HCC in Egyptian patients with HCV related chronic liver disease. Asian Pac J Cancer Prevent 16(1):381–385

Petrick JL, McGlynn KA (2019) The changing epidemiology of primary liver cancer. CurrEpidemiol Rep 6:104–111. https://doi.org/10.1007/s40471-019-00188-3

Rajesh S, Choudhury NA, Moulik S (2020) Hepatocellular Carcinoma (HCC) liver cancer prediction using machine learning algorithms. In: IEEE 17th India Council International Conference (INDICON), New Delhi, India. pp. 1–5, https://doi.org/10.1109/INDICON49873.2020.9342443.

Santos MS, Abreu PH, García-Laencina PJ, Simão A, Carvalho A (2015) A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. J Biomed Inform 58:49–59

Sawhney R, Mathur P, Shankar R (2018) A firefly algorithm-based wrapper-penalty feature selection method for cancer diagnosis. In: International Conference on Computational Science and Its Applications. Springer. pp. 438–49.

Sharma M (2019) Cervical cancer prognosis using genetic algorithm and adaptive boosting approach. Heal Technol 9(5):877–886

Shi H-Y, Lee K-T, Lee H-H, Ho W-H, Sun D-P, Wang J-J, et al (2012) Comparison of artificial neural network and logistic regression models for predicting in-hospital mortality after primary liver cancer surgery. PLoS One.7(4): e35781. https://doi.org/10.1371/journal.pone.0035781

Tuncer T, Ertam F (2019) Neighborhood component analysis and reliefF based survival recognition methods for Hepatocellular carcinoma. Phys A. https://doi.org/10.1016/j.physa.2019.123143

UCI Machine learning repository (2020) https://archive.ics.uci.edu/ml/index.php Accessed on 22–02–2020.

Wilson DR, Martinez TR (1997) Improved heterogeneous distance functions. J ArtifIntell Res 6:1–34

Wilson DR, Martinez TR (2000) An integrated instance-based learning algorithm. ComputIntell 16(1):1–28

Zhang Z-M, Tan J-X, Wang F, Dao F-Y, Zhang Z-Y, Lin H (2020) Early diagnosis of hepatocellular carcinoma using machine learning method. Front BioengBiotechnol 8:254. https://doi.org/10.3389/fbioe.2020.00254

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.