Neural Computing and Applications (2021) 33:2783–2792
https://doi.org/10.1007/s00521-020-05157-2

ORIGINAL ARTICLE

LDA–GA–SVM: improved hepatocellular carcinoma prediction through dimensionality reduction and genetically optimized support vector machine

Liaqat Ali[1,2] · Iram Wajahat[3] · Noorbakhsh Amiri Golilarz[4] · Fazel Keshtkar[5] · Syed Ahmad Chan Bukhari[5]

Received: 17 January 2020 / Accepted: 19 June 2020 / Published online: 10 July 2020

Abstract

Hepatocellular carcinoma (HCC) is a common type of liver cancer worldwide. Patients with HCC have rare chances of survival. The chances of survival increase, if the cancer is diagnosed early. Hence, different machine learning-based methods have been developed by researchers for the accurate detection of HCC. However, high dimensionality (curse of dimensionality) and lower prediction accuracy are the problems in the automated detection of HCC. Dimensionality reduction-based methods have shown state-of-the-art performance on many disease detection problems, which motivates the development of machine learning models based on reduced features dimension. This paper proposes a new hybrid intelligent system that hybridizes three algorithms, i.e., linear discriminant analysis (LDA) for dimensionality reduction, support vector machine (SVM) for classification and genetic algorithm (GA) for SVM optimization. Consequently, the three models are hybridized and one black box model, namely LDA–GA–SVM, is constructed. Experimental results on publicly available HCC dataset show improvement in the HCC prediction accuracy. Apart from performance improvement, the proposed method also shows lower complexity from two aspects, i.e., reduced processing time in terms of hyperparameters optimization and training time. The proposed method achieved accuracy of 90.30%, sensitivity of 82.25%, specificity of 96.07% and Matthews Correlation Coefficient (MCC) of 0.804.

Keywords Feature extraction · Genetic algorithm · Hepatocellular carcinoma · Hyperparameter optimization · Support vector machine

✉ Syed Ahmad Chan Bukhari
bukharis@stjohns.edu

Liaqat Ali
engr.liaqat@ustb.edu.pk; engr_liaqat183@yahoo.com

Iram Wajahat
iramwajahat.pharmd@gmail.com

Noorbakhsh Amiri Golilarz
noorbakhsh.amiri@std.uestc.edu.cn

Fazel Keshtkar
keshtkaf@stjohns.edu

[1] School of Information and Communication Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China

[2] Department of Electrical Engineering, University of Science and Technology, Bannu, Pakistan

[3] Allied Institute of Medical Sciences, Ahmad College of Pharmacy, Gujrat, Pakistan

[4] School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China

[5] Division of Computer Science, Mathematics and Science, Collins College of Professional Studies, St. John's University, Jamaica, NY, USA

1 Introduction

As per World Health Organization (WHO) reports, about 14.1 million new cancer patients and 8.2 million deaths are caused by cancer worldwide [36]. Hepatocellular carcinoma (HCC), which is the malignancy of liver and is caused by chronic liver disease and cirrhosis, is one type of cancer. Recent research shows that the deadliest cancer around the world is HCC that is causing around 600,000 deaths each year [14, 17, 34]. Moreover, liver cancer is ranked as the sixth commonly diagnosed cancer all over the world [29]. These facts evidently show the impact of HCC on human life worldwide. Therefore, it is the need of the hour to lower down the deaths caused by HCC which is possible only if HCC is diagnosed at the early stages. In order to meet this objective, we need to exploit different techniques of data mining and machine learning to design an automated diagnostic system for efficient HCC prediction.

Data mining is an interdisciplinary field that uses the computer science and statistical concepts to extract meaningful information (features or rules) from the given data [11]. On the other hand, machine learning is the subfield of computer science that deals with the techniques and methods where machine learns from the experience [21, 25–28, 38, 39]. Nowadays, the machine learning techniques and data mining are rapidly growing and widely applied to address different issues in the field of water resource management such as [2, 3, 12, 22] and in the field of medical diagnostics such as Parkinson's disease [6, 8, 16], heart disease [7, 9–11, 24], breast cancer [40] and Alzheimer's disease [33, 41].

HCC and liver diseases have been exhaustively studied in recent years. A new cluster-based oversampling method has been developed for HCC prediction [29]. For preprocessing of the data, heterogeneous and missing data (HEOM) and *K*-means as clustering technique were used. Dataset was balanced through synthetic minority oversampling technique (SMOTE) method. Furthermore, balance data were input to logistic regression and neural network algorithms. Results of the proposed model suggested the effective detection of the HCC. For timely detection of liver disease, Hasson et al. [19] proposed a new approach based on optimally generated rules using boosted C5.0 through genetic algorithm. The accuracy of C5.0 was increased from 81 to 93% through the proposed approach. For diagnosis of liver disease, a new approach was proposed by Abdar et al. [5]. In their research, classification and regression tree (CART), base C5.0 and Chi-square automatic interaction detector (CHAID) methods were used. A boosted-based decision tree with multilayer perceptron neural network (MLPNN) was used to improve the performance of the methods. For early detection of the liver disease, the proposed approach based on MLPNN with boosted-based decision trees was found efficient.

Adar et al. [4] used two decision trees for the detection of liver disease. They used an Indian patient dataset (ILPD) and C5.0 and CHAID algorithms for the experiments. The performance of C5.0 was optimized by using C5.0 boosting method. Moreover, Adar et al. introduced various simple rules in C5.0 and CHAID techniques. The accuracy of liver disease detection was 93.75% through boosted C5.0. Abajian et al. carried out a study on 36 patients of HCC [1] which were treated with transarterial chemoembolization through using machine learning technique. Through linear regression and random forest, 78% overall accuracy was achieved by Abajian et al. A regression model was proposed by Wasyluk et al. [37] to analyze the liver disorder. In their research, 200 cirrhotic patients were studied. Furthermore, clinical examination consisting of various types of factors based on histopathological data and laboratory tests was carried out. A comparative study by Shi et al. [32] suggested that ANN has better performance than LR for predicting in-hospital mortality after preliminary liver cancer surgical operation.

In recent years, various studies proposed LDA-based automated systems and obtained state-of-the-art prediction accuracies on different disease detection problems. A brief survey of these methods is as follows: Dogantekin et al. [18] developed a diagnostic system for hepatitis disease detection using LDA for dimensionality reduction and adaptive neuro-fuzzy inference system (ANFIS) for classification and obtained 94.16%. Abdulkadir Sengur proposed LDA-ANFIS system for heart valve disease detection and obtained 94% specificity and 95.9% of sensitivity rates [31]. Subasi and Gursoy proposed a hybrid method by hybridizing LDA with SVM model for improved epilepsy prediction [35]. Calisir and Dogantekin [15] proposed hybridization of LDA with wavelet SVM, i.e., WSVM for improved detection of diabetes disease.

Motivated by the development of different diagnostic systems based on linear discriminant analysis and machine learning models to improve precision of decision making about HCC diagnosis, we also develop a hybrid intelligent system. The proposed system hybridizes three algorithms, i.e., linear discriminant analysis (LDA), SVM and GA. The LDA model is used for reducing dimensionality of feature vectors, SVM is used for classification purposes, and GA is used for efficient optimization of the SVM model. The performance of the proposed hybrid model named LDA–GA–SVM is compared with conventional SVM models, other state-of-the-art ensemble models and previously proposed methods. Experimental results validated the effectiveness of the proposed model.

2 Materials and methods

2.1 Hepatocellular carcinoma dataset

The dataset used in this paper for HCC prediction is adopted from UCI machine learning repository. The dataset was collected at Coimbra's Hospital and Universitary Centre (CHUC), Portugal, and contains samples collected from 165 subjects [29]. The dataset contains 49 features in total, which can be subdivided into two groups, i.e., quantitative features and qualitative features. The number of quantitative features is equal to 23, and the number of qualitative features is equal to 26. Details about the features of the dataset are given in Table 1. The label of the dataset denotes survival at one year and can assume a value of 0 (dies) or 1 (lives/survives). It is important to discuss that each feature has missing values and there are only eight samples in the dataset that have the information of all features. In the literature, two types of methodologies are used for dealing with missing values. First, delete all those samples that contains missing values. Second, impute the missing values. The first method cannot be utilized as it will result in only eight samples. Therefore, we used imputation of missing values. In this study, we utilized univariate feature imputation method using SimpleImputer class of the scikit-learn library of python [13]. The SimpleImputer class has different strategies for imputing the missing values. However, we used statistical method of imputing missing values by mean value of the column or feature where the missing value is located.

2.2 Proposed method

The main objective of development of an automated diagnostic system based on supervised machine learning methods is to come up with a hypothesis (a fitting function) that can better fit the training data (producing high training accuracy) as well as unseen testing data (i.e., also shows better testing accuracy). To improve the diagnostic performance (i.e., disease classification accuracy), different data mining algorithms are utilized for features preprocessing. This features preprocessing is broadly divided into two categories, i.e., features selection and features extraction. In features selection, different statistical or search-based strategies are utilized in order to explore a subset of features having high dependence on the label of the data. Hence, in features selection subset of original features is selected. On the other hand in features extraction, the original set of features are transformed and new features are extracted from the original features. In this paper, we exploit a feature extraction method (also known as dimensionality reduction method). The method is known as linear discriminant analysis (LDA) in literature. LDA is used to transform the original features set into a reduced dimension in order to improve the predictive capabilities of machine learning-based predictive models. LDA-reduced feature vector(s) are selected using two criteria. First, such vector(s) are selected that ensure maximum class separation (distance between the two classes). Second, the transformed feature vector(s) might ensure minimum within-class scatter (i.e., distance of within-class samples). This objective is met by maximizing the fisher ratio which is given as follows:

$$\frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \tag{1}$$

where the variance of the first class has been denoted by σ_1 and the variance of the second class has been denoted by σ_2. In (1), $(\mu_1 - \mu_2)$ denotes the difference between center points or means of the two classes or distributions. To make the concepts of (1) more clear, $(\mu_1 - \mu_2)^2$ can be denoted by S_B, i.e., between-class scatter and $\sigma_1^2 + \sigma_2^2$ can be denoted by S_W, i.e., within-class scatter. Hence, (1) can also be formulated as follows:

$$\frac{S_B}{S_W} \tag{2}$$

The job of LDA is to maximize the fisher ratio which will result in minimum within-class scatter and maximum between-class scatter. To achieve this goal, we need a transformation matrix w.

We can formulate S_W and S_B as follows:

$$S_B = w^T S_B w \tag{3}$$

$$S_W = w^T S_W w \tag{4}$$

Hence, (2) becomes

$$\frac{w^T S_B w}{w^T S_W w} \tag{5}$$

LDA maximizes (5) by evaluating the transformation matrix, i.e., w. This is done by calculating the eigenvectors of $S_W^{-1} S_B$. Thus, LDA uses the transformation matrix to transform data having Q dimensions into L dimension(s) where $L <= (C-1)$, C stands for the number of classes in the dataset. During the transformation through w, LDA ensures maximization of (2).

The feature extraction process through LDA offers two advantages. It reduces the time complexity of the machine learning models by reducing the original features set. Second, it improves the disease prediction accuracy by increasing the class separability. After the dimensionality reduction by LDA model, the reduced feature vector is applied to SVM model for classification. The SVM model formulation is briefly discussed as follows:

Table 1 Details of the features of the HCC dataset

Feature no.	Features	Rang	Type/scale
1	Gender	0/1	Qualitative/dichotomous
2	Symptoms	0/1	Qualitative/dichotomous
3	Alcohol	0/1	Qualitative/dichotomous
4	Hepatitis B surface antigen: HBsAg	0/1	Qualitative/dichotomous
5	Hepatitis B e-antigen: HBeAg	0/1	Qualitative/dichotomous
6	Hepatitis B core antibody: HBcAb	0/1	Qualitative/dichotomous
7	Hepatitis C virus antibody: HCVAb	0/1	Qualitative/dichotomous
8	Cirrhosis	0/1	Qualitative/dichotomous
9	Endemic countries	0/1	Qualitative/dichotomous
10	Smoking	0/1	Qualitative/dichotomous
11	Diabetes	0/1	Qualitative/dichotomous
12	Obesity	0/1	Qualitative/dichotomous
13	Hemochromatosis	0/1	Qualitative/dichotomous
14	Arterial hypertension: AHT	0/1	Qualitative/dichotomous
15	Chronic renal insufficiency: CRT	0/1	Qualitative/dichotomous
16	Human immunodeficiency virus: HIV	0/1	Qualitative/dichotomous
17	Nonalcoholic steatohepatitis: NASH	0/1	Qualitative/dichotomous
18	Esophageal varices	0/1	Qualitative/dichotomous
19	Splenomegaly	0/1	Qualitative/dichotomous
20	Portal hypertension	0/1	Qualitative/dichotomous
21	Portal vein thrombosis	0/1	Qualitative/dichotomous
22	Liver metastasis	0/1	Qualitative/dichotomous
23	Radiological hallmark	0/1	Qualitative/dichotomous
24	Age at diagnosis	20–93	Quantitative/ratio
25	Grams of alcohol per day: Grams/day	0–500	Quantitative/ratio
26	Packs of cigarettes per year: Packs/year	0–510	Quantitative/ratio
27	Performance status	0, 1, 2, 3, 4	Qualitative/ordinal
28	Encephalopathy degree	1, 2, 3	Qualitative/ordinal
29	Ascites degree	1, 2, 3	Qualitative/ordinal
30	International normalized ratio: INR	0.84–4.82	Quantitative/ratio
31	Alpha-fetoprotein (ng/mL): AFP	1.2–1,810,346	Quantitative/ratio
32	Hemoglobin (g/dL)	5–18.7	Quantitative/ratio
33	Mean corpuscular volume (fl): MCV	69.5–119.6	Quantitative/ratio
34	Leukocytes (G/L)	2.2–13,000	Quantitative/ratio
35	Platelets (G/L)	1.71–459,000	Quantitative/ratio
36	Albumin (mg/dL)	1.9–4.9	Quantitative/ratio
37	Total bilirubin (mg/dL): Total Bil	0.3–40.5	Quantitative/ratio
38	Alanine transaminase (U/L): ALT	11–420	Quantitative/ratio
39	Aspartate transaminase (U/L): AST	17–553	Quantitative/ratio
40	Gamma glutamyl transferase (U/L): GGT	23–1575	Quantitative/ratio
41	Alkaline phosphatase (U/L): ALP	1.28–980	Quantitative/ratio
42	Total proteins (g/dL): TP	3.9–102	Quantitative/ratio
43	Creatinine (mg/dL)	0.2–7.6	Quantitative/ratio
44	Number of nodules	0–5	Quantitative/ratio
45	Major dimension of nodule (cm)	1.5–22	Quantitative/ratio
46	Direct bilirubin (mg/dL)	0.1–29.3	Quantitative/ratio
47	Iron (mcg/dL)	0–224	Quantitative/ratio
48	Oxygen saturation (%)	0–126	Quantitative/ratio
49	Ferritin (ng/mL)	0–2230	Quantitative/ratio
Label	Class	0, 1	Quantitative/ratio

SVM is a machine learning (supervised) model that can be used for classification and regression problems. SVM tries to construct a hyperplane having as large margin as possible. The hyperplane $g(x) = \theta^T * x + \delta$, where δ represents the bias and θ denotes the weight vector, is constructed based on training data and it works like a decision boundary for deciding the class of a data point (a multi-dimensional feature vector) in case of classification problem. In order to generate a margin, SVM finds out the closest vectors (data points) of two classes in case of binary classification. These vectors are known as support vectors. Margin is calculated as the perpendicular distance between the lines passing through the support vectors and is given by $\frac{2}{\|\theta\|_2^2}$. The main objective is to construct an optimized SVM model that will yield an optimal hyperplane with maximum margin.

SVM uses a set of slack variables denoted by ξ_i, $i = 1, \ldots, S$ and a penalty parameter, i.e., λ, and attempts the maximization of $\| \theta \|_2^2$ and minimization of the misclassification errors. This fact is formulated as follows:

$$\min_{\theta,\delta,\xi} \underbrace{\frac{1}{2} \| \theta \|_2^2}_{\text{Regularizer}} + \lambda \underbrace{\sum_{i=1}^{S} \xi_i}_{\text{Error or Loss}} \quad \text{s.t} \begin{cases} y_i(\theta x_i + \delta) \geq 1 - \xi_i \\ \xi_i \geq 0, \quad i = 1, \ldots, S \end{cases} \tag{6}$$

where ξ is slack variable that calibrate the degree of misclassification and Euclidean norm or L_2-norm is the penalty term.

The main problem is that most of the times, linear hyperplane cannot separate the data points of the two classes efficiently (i.e., with minimum classification error). In such a case, SVM exploits kernel trick in which the SVM model transform the regional data points into a higher dimensional points with an aim of transforming non-separable data points into a separable form. For this job, different types of kernels are used, namely radial basis function (RBF) kernel, sigmoid kernel and polynomial kernel. These kernels acts as hyperparameters of a SVM model which needs to be optimized for a given problem.

In order to obtain an SVM model that would show better performance on a certain problem, we need to tune or optimize its hyperparameters carefully. The commonly used method to meet this objective is grid search. However, gird search is computationally very expensive as it evaluates every possible combination of hyperparameters; thus, to avoid this problem in this paper, we propose to use genetic algorithm (GA). In the proposed LDA–GA–SVM method, the hyperparameters of the SVM model are dynamically optimized through GA evolutionary algorithm. GA randomly generates initial population consisting of chromosomes. The values of the three important hyperparameters of the SVM model, i.e., type of kernel, λ and gamma, are directly coded in the chromosomes. To assess the performance of each chromosome, a fitness function is designed.

In this paper, we design the fitness function as the generalization loss achieved over stratified k-fold cross-validation so that the optimization process yields both generalized and efficient models in terms of disease classification accuracy. The developed GA algorithm works in three main stages, i.e., selection, crossover and mutation operators to generate the offspring of the existing population. The literature shows that two different methods are usually utilized in the selection phase. The first one is known as roulette wheel method, and the second method is tournament selection. This study uses tournament selection for the selection process. During selection process, the tournament method selects those individuals or chromosomes which have the best fitness value. The selected individuals contribute to the population of the next generation. During the crossover process, parents are combined to produce children for next generation. During the mutation process, a chromosome is mutated based on a predefined probability, i.e., mutation probability. Finally, the mutated chromosomes/individuals constitute the population for next generation. And the same process is repeated in the next generation. Number of generations, population size and mutation probability are the parameters of GA algorithms. In this paper, we have used 20 number of generations, population size of 50, mutation probability of 0.10 and tournament size of 5.

3 Experimental results and discussion

In this section, we discuss the experimental setting and the obtained results. Initially, the proposed LDA–GA–SVM method is implemented and the results are analyzed. Next, to validate the effectiveness of the proposed method, we compare its performance with other state-of-the-art machine learning ensemble models. And to further validate its effectiveness, we also performed a comparative study with other methods presented in the literature for HCC detection. In all the experiments, we utilized stratified fivefold cross-validation and for evaluation purposes we used six evaluation metrics namely ROC curve, accuracy, area under the curve (AUC), sensitivity and specificity and MCC. All the experiments were performed using Python software package and scikit-learn library [23]. Moreover, all the simulations were carried out using Intel(R) Core (TM) i5 CPU with 8GB RAM and 64-bit operating system.

In the first set of experiments, we developed four models based on SVM. The first two models use commonly used

traditional SVMs, i.e., SVM with linear kernel and SVM with RBF kernels. Both the models are optimized using genetic algorithm. The other two models exploit the proposed hybrid framework, i.e., LDA–GA–SVM is developed for SVM with linear kernel and another LDA–GA–SVM model is developed using SVM model with RBF kernel. For the traditional SVM model with linear kernel, we obtained classification accuracy of 78.18%, specificity of 82.35%, sensitivity of 72.58% and MCC of 0.547. For the traditional SVM model with RBF kernel, we obtained accuracy of 78.78%, specificity of 83.33% and sensitivity of 72.58% and MCC of 0.559. In the same way, the proposed method LDA–GA–SVM produced 89.69% of accuracy, 95.09% of specificity, 82.25% of sensitivity and 0.791 of MCC using linear kernel for SVM model. The proposed model, i.e., LDA–GA–SVM, was also simulated for RBF kernel, and the results were 90.30% accuracy, 96.07% specificity, 82.25% sensitivity and 0.804 MCC. These results are summarized and tabulated in Table 2. From the results, it is cleared that the proposed method not only brings down the complexity of the machine learning models by reducing the dimensionality of the feature vectors but also improves the classification accuracy.

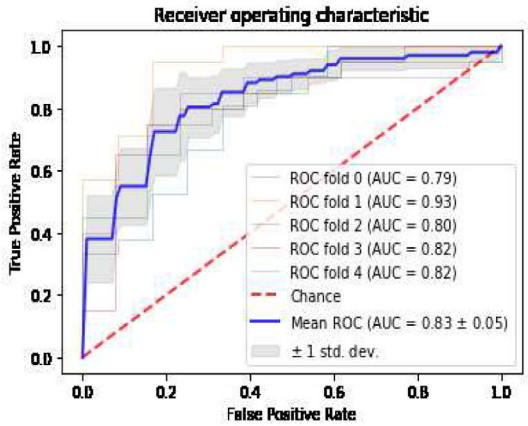
In machine learning, it is a common practice to use receiver operating curves (ROC) for evaluating the quality of output of a constructed model. Thus, we further evaluate the constructed models by using ROC charts for the above discussed four models. The ROC charts for the SVM model with linear kernel and the proposed LDA–SVM–GA model with linear SVM are given in Fig. 1a, b, respectively. It is evidently clear from these figures that the area under curve (AUC) for traditional SVM with linear kernel is 0.83 and the AUC for the proposed method using SVM linear models is 0.96. Similarly, the ROC charts for the SVM model with linear kernel and the proposed LDA–GA–SVM model with RBF SVM are given in Fig. 2a, b, respectively. The AUC for the traditional SVM model with RBF kernel is 0.82, and the AUC for the proposed model, i.e., LDA–GA–SVM with SVM having RBF kernel, is 0.96. Hence, the improvement in AUC due to the proposed method is also evidently clear. Thus, based on the performance reflected by accuracy and AUC, we can opt the LDA–GA–SVM model with SVM having RBF kernel as optimal model owing to its higher accuracy. The results of the ROC curves are also tabulated in the table. These results are summarized and tabulated in Table 2.

In the literature, ensemble models are well known in machine learning for their enhanced performance. Based on this fact, in this study we also developed three different ensemble models, namely random forest ensemble model, extra tree ensemble model and Adaboost ensemble model. While developing these ensemble models, the hyperparameters of these models were optimized using grid search algorithm. Simulation results show that the optimal performance under Adaboost model was obtained with 74.39% of accuracy, 60.31% of specificity, 83.16% sensitivity and 0.448 MCC. Similarly, for the extra tree ensemble model, optimal performance was obtained with 74.39% of accuracy, 52.38% of specificity, 88.11% sensitivity and 0.441 MCC. Finally, for the random forest ensemble model, optimal performance was obtained with 75.60% of accuracy, 52.38% of specificity, 90.09% sensitivity and 0.469 MCC. To evaluate the quality of output of these ensemble models, we further evaluated the AUC and the ROC curves for these models. The ROC curves are depicted in Fig. 3. From the ROC curves, it is clear that the AUC of 0.72, 0.77 and 0.78 is produced for the Adaboost, RF and ET models, respectively.

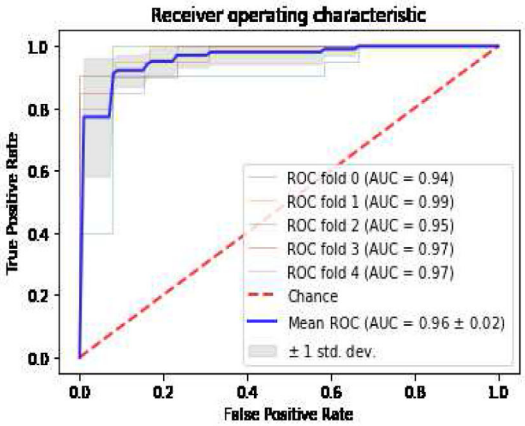
From the above experiments, it is evident that the proposed LDA–GA–SVM model shows better performance than traditional SVMs and three different state-of-the-art machine learning ensemble methods. In this part of the study, we further validate the effectiveness of the proposed model by showing that it also offers lower complexity in terms of processing or training time. To meet this objective, comparative analysis is performed from processing time aspect. In this experiment, the hyperparameters optimization through genetic algorithm is compared with the conventional method of hyperparameters optimization through grid search method. The results are reported in Table 3. It can be seen in the table that the optimization of SVM model through GA is performed in 1.15 s, while the optimization through baseline grid search algorithm is done in 41.56 s using SVM model with RBF kernel. Similarly,

Table 2 Performance evaluation of different models

Method	ACC (%)	Sen. (%)	Spec. (%)	MCC	AUC
SVM (linear)	78.18	72.58	82.35	0.547	0.83
SVM (RBF)	78.78	72.58	83.33	0.559	0.82
LDA–GA–SVM (linear)	89.69	82.25	95.09	0.791	0.96
LDA–GA–SVM (RBF)	90.30	82.25	96.07	0.804	0.96
Random forest	74.54	58.06	85.29	0.454	0.77
Extra tree	76.36	67.47	82.35	0.504	0.81
Adaboost	73.93	62.90	81.37	0.449	0.78

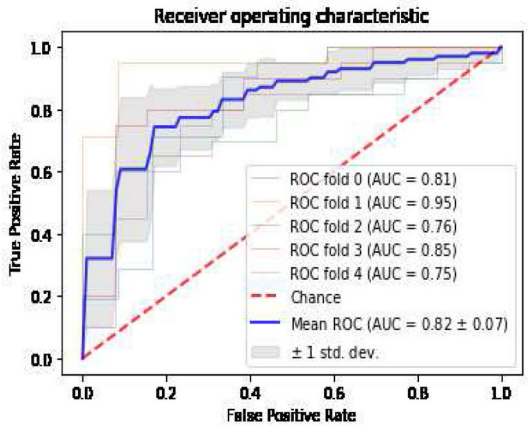


(a) ROC chart of SVM linear model with traditional optimization

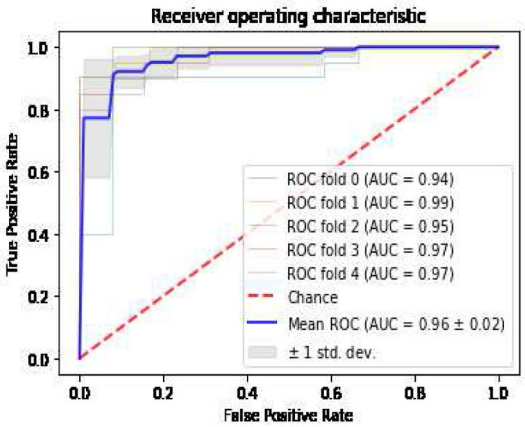


(b) ROC chart of the proposed LDA-GA-SVM model with SVM having linear kernel

Fig. 1 ROC charts using traditional SVM linear model and the proposed LDA–GA–SVM model



(a) ROC chart of SVM RBF model with traditional optimization

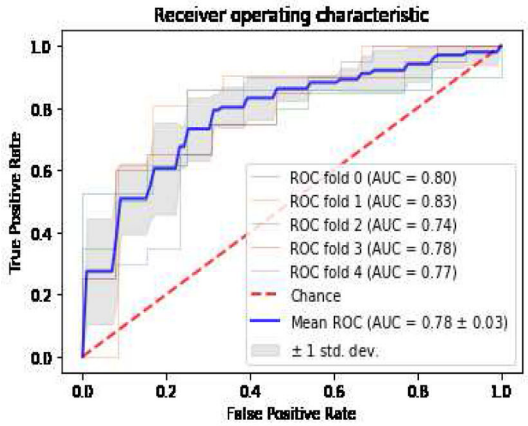


(b) ROC chart of the proposed LDA-GA-SVM model with SVM having RBF kernel

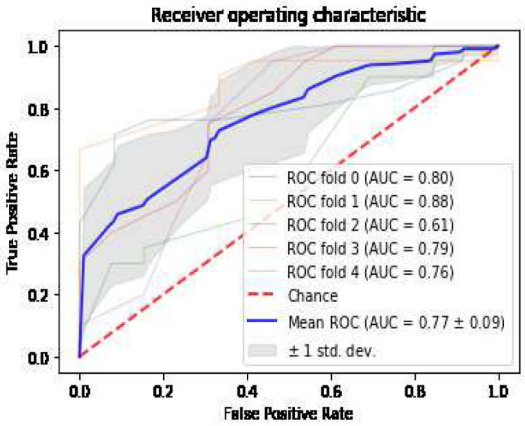
Fig. 2 ROC charts using traditional SVM RBF model and the proposed LDA–GA–SVM model

for SVM model with linear kernel, the hyperparameters optimization through genetic algorithm took 0.48 s, while the hyperparameters optimization through grid search algorithm took 1.66 s. Apart from time complexity reduction in terms of hyperparameters optimization, similar reduction in the time complexity of the developed intelligent system is also observed. This is due to the fact that LDA reduces the higher-dimensional feature vector to lower-dimensional space. Thus, it is evidently clear that the application of the proposed method not only enhances the predictive capabilities of the SVM models but also reduces their complexity tremendously.

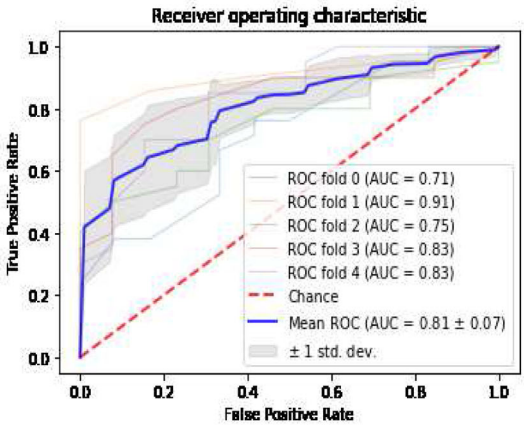
To further validate effectiveness of the proposed method, we review the methods proposed for the HCC prediction based on the same dataset that has been used in this study. The dataset was collected by Santos et al. in [29] and analyzed using augmented set approach and neural network. They achieved 75.19% of accuracy. Another study was conducted by Sawhney et al. [30] in which they proposed binary firefly algorithm (BFA)-based feature selection and random forest-based classification by using a penalty function to the existing fitness function of BFA. They achieved HCC prediction accuracy of 83.5%. Recently, Ksiazek et al. [21] proposed a novel method by



(a) ROC chart of the Adaboost model



(b) ROC chart of the RF model



(c) ROC chart of the ET model

Fig. 3 ROC charts of ensemble models

Table 3 Time complexity analysis

Method	Processing time (s)	Model	Type
Proposed	1.150	LDA–GA–SVM (RBF)	HPO
Baseline	41.56	LDA–GA–SVM (RBF)	HPO
Proposed	0.480	LDA–GA–SVM (Lin)	HPO
Baseline	1.660	LDA–GA–SVM (Lin)	HPO

HPO hyperparameters optimization

utilizing two-stage genetic algorithm base machine learning framework. The first genetic algorithm was used as feature selector, while the second one was coupled for hyperparameters optimization. Their results showed HCC prediction accuracy of 88.49%. From Table 4, it can be seen that the proposed model has shown better result as compared to the recently proposed methods. In this paper, we proposed a hybrid machine learning framework, i.e., LDA–GA–SVM and further improved the HCC prediction to 90.24%.

4 Conclusion

In this paper, the problem of hepatocellular carcinoma (HCC) automated detection or prediction through machine learning was considered. In order to improve the HCC prediction accuracy, we proposed a learning method namely LDA–GA–SVM. The proposed method uses linear discriminant analysis model for reducing the dimensionality of the HCC feature vector, while genetic algorithm (GA) was utilized to construct an optimized version of support vector machines (SVMs) which were used for classification purposes. Two types of SVM models were developed, i.e., SVM with linear kernel and SVM with RBF kernel. It was observed that the proposed method shows lower complexity in terms of processing time. The lower complexity was observed from two aspects, i.e., hyperparameters optimization and training time. Apart from reducing the complexity, the proposed method also showed better HCC prediction. The performance of the

Table 4 HCC prediction performance obtained by other methods proposed in the literature for HCC prediction

Study	Method	Acc.	F-score
Santos et al. [29]	NN + augmented set approach	0.7519 ± 0.0105	0.6650 ± 0.0182
Sawhney et al. [30]	BFA + RF	0.835	N. A
Ksiazek et al. [21]	SVC with new 2-level genetic optimizer approach	0.8849	0.8762
Kayal et al. [20]	DNN	0.78	0.80
Proposed method (2019)	LDA–GA–SVM (with linear and RBF kernel)	0.9030	N.A

proposed method was compared with state-of-the-art ensemble machine learning models (which are known for their enhanced performance) and previously proposed methods. The findings of the study suggest that the proposed method can be proved helpful to oncologists for improving quality of decision making during diagnosis of HCC patients.

Compliance with ethical standards

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.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Abajian A, Murali N, Savic LJ, Laage-Gaupp FM, Nezami N, Duncan JS, Schlachter T, Lin M, Geschwind JF, Chapiro J (2018) Predicting treatment response to intra-arterial therapies for hepatocellular carcinoma with the use of supervised machine learning-an artificial intelligence concept. J Vasc Interv Radiol 29(6):850–857
2. Abbaszadeh P (2016) Improving hydrological process modeling using optimized threshold-based wavelet de-noising technique. Water Resour Manag 30(5):1701–1721
3. Abbaszadeh P, Alipour A, Asadi S (2018) Development of a coupled wavelet transform and evolutionary l evenberg-m arquardt neural networks for hydrological process modeling. Comput Intell 34(1):175–199
4. Abdar M, Zomorodi-Moghadam M (2018) Impact of patients' gender on parkinson's disease using classification algorithms. J AI Data Min 6(2):277–285
5. Abdar M, Yen NY, Hung JCS (2018a) Improving the diagnosis of liver disease using multilayer perceptron neural network and boosted decision trees. J Med Biol Eng 38(6):953–965
6. Abdar M, Zomorodi-Moghadam M, Zhou X, Gururajan R, Tao X, Barua PD, Gururajan R (2018b) A new nested ensemble technique for automated diagnosis of breast cancer. Pattern Recognit Lett 132:123–131
7. Ali L, Khan SU, Anwar M, Asif M (2019) Early detection of heart failure by reducing the time complexity of the machine learning based predictive model. In: 2019 International conference on electrical, communication, and computer engineering (ICECCE), IEEE, pp 1–5
8. Ali L, Khan SU, Arshad M, Ali S, Anwar M (2019) A multi-model framework for evaluating type of speech samples having complementary information about parkinson's disease. In: 2019 International conference on electrical, communication, and computer engineering (ICECCE), IEEE, pp 1–5
9. Ali L, Khan SU, Golilarz NA, Yakubu I, Qasim I, Noor A, Nour R (2019) A feature-driven decision support system for heart failure prediction based on statistical model and gaussian naive bayes. Comput Math Methods Med 2019:1–8
10. Ali L, Niamat A, Khan JA, Golilarz NA, Xingzhong X, Noor A, Nour R, Bukhari SAC (2019) An optimized stacked support vector machines based expert system for the effective prediction of heart failure. IEEE Access 7:54007–54014
11. Ali L, Bukhari S (2020) An approach based on mutually informed neural networks to optimize the generalization capabilities of decision support systems developed for heart failure prediction. IRBM
12. Alipour A, Ahmadalipour A, Abbaszadeh P, Moradkhani H (2020) Leveraging machine learning for predicting flash flood damage in the southeast us. Environ Res Lett 15(2):024011
13. Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, Niculae V, Prettenhofer P, Gramfort A, Grobler J, Layton R, VanderPlas J, Joly A, Holt B, Varoquaux G (2019) Imputation of missing values using scikit-learn. https://scikit-learn.org/stable/modules/impute.html
14. Cabibbo G, Latteri F, Antonucci M, Craxì A (2009) Multimodal approaches to the treatment of hepatocellular carcinoma. Nat Rev Gastroenterol Hepatol 6(3):159
15. Çalişir D, Doğantekin E (2011) An automatic diabetes diagnosis system based on lda-wavelet support vector machine classifier. Expert Syst Appl 38(7):8311–8315
16. Das R (2010) A comparison of multiple classification methods for diagnosis of parkinson disease. Expert Syst Appl 37(2):1568–1572
17. DeWaal D, Nogueira V, Terry AR, Patra KC, Jeon SM, Guzman G, Au J, Long CP, Antoniewicz MR, Hay N (2018) Hexokinase-2 depletion inhibits glycolysis and induces oxidative phosphorylation in hepatocellular carcinoma and sensitizes to metformin. Nat Commun 9(1):446
18. Dogantekin E, Dogantekin A, Avci D (2009) Automatic hepatitis diagnosis system based on linear discriminant analysis and adaptive network based on fuzzy inference system. Expert Syst Appl 36(8):11282–11286
19. Hassoon M, Kouhi MS, Zomorodi-Moghadam M, Abdar M (2017) Rule optimization of boosted c5.0 classification using genetic algorithm for liver disease prediction. In: 2017 International conference on computer and applications (ICCA), IEEE, pp 299–305

20. 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 VE, Mandal JK (eds.) Proceedings of international ethical hacking conference 2018, Springer, Singapore, pp 349–358
21. Ksi W (2019) A novel machine learning approach for early detection of hepatocellular carcinoma patients. Cognit Syst Res 54:116–127
22. Nourani V, Tahershamsi A, Abbaszadeh P, Shahrabi J, Hadavandi E (2014) A new hybrid algorithm for rainfall-runoff process modeling based on the wavelet transform and genetic fuzzy system. J Hydroinform 16(5):1004–1024
23. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
24. Pławiak P (2018) Novel methodology of cardiac health recognition based on ecg signals and evolutionary-neural system. Expert Syst Appl 92:334–349
25. Pławiak P, Maziarz W (2014) Classification of tea specimens using novel hybrid artificial intelligence methods. Sens Actuators B Chem 192:117–125
26. Pławiak P, Rzecki K (2014) Approximation of phenol concentration using computational intelligence methods based on signals from the metal-oxide sensor array. IEEE Sens J 15(3):1770–1783
27. Rzecki K, Pławiak P, Niedźwiecki M, Sośnicki T, Leśkow J, Ciesielski M (2017) Person recognition based on touch screen gestures using computational intelligence methods. Inf Sci 415:70–84
28. Rzecki K, Sośnicki T, Baran M, Niedźwiecki M, Król M, Łojewski T, Acharya U, Yildirim Ö, Pławiak P (2018) Application of computational intelligence methods for the automated identification of paper-ink samples based on libs. Sensors 18(11):3670
29. 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
30. Sawhney R, Mathur P, Shankar R (2018) A firefly algorithm based wrapper-penalty feature selection method for cancer diagnosis. In: Gervasi O, Murgante B, Misra S, Stankova E, Torre CM, Rocha AMA, Taniar D, Apduhan BO, Tarantino E, Ryu Y (eds) Computational science and its applications-ICCSA 2018. Springer, Cham, pp 438–449
31. Sengur A (2008) An expert system based on linear discriminant analysis and adaptive neuro-fuzzy inference system to diagnosis heart valve diseases. Expert Syst Appl 35(1–2):214–222
32. Shi HY, Lee KT, Lee HH, Ho WH, Sun DP, Wang JJ, Chiu CC (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
33. Shi J, Zheng X, Li Y, Zhang Q, Ying S (2017) Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of alzheimer's disease. IEEE J Biomed Health Inform 22(1):173–183
34. Singh S, Singh PP, Roberts LR, Sanchez W (2014) Chemopreventive strategies in hepatocellular carcinoma. Nat Rev Gastroenterol Hepatol 11(1):45
35. Subasi A, Gursoy MI (2010) EEG signal classification using PCA, ICA, LDA and support vector machines. Expert Syst Appl 37(12):8659–8666
36. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A (2015) Global cancer statistics, 2012. CA Cancer J Clin 65(2):87–108
37. Wasyluk HA, Cianciara J, Bobrowski L, Drapato A (2010) Founding of database for cirrhotic patients for early detection of hepatocellular carcinoma. Hepatology 6(3):13–16
38. Yildirim Ö (2018) A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification. Comput Biol Med 96:189–202
39. Yildirim O, San Tan R, Acharya UR (2018) An efficient compression of ECG signals using deep convolutional autoencoders. Cognit Syst Res 52:198–211
40. Zheng B, Yoon SW, Lam SS (2014) Breast cancer diagnosis based on feature extraction using a hybrid of *k*-means and support vector machine algorithms. Expert Syst Appl 41(4):1476–1482
41. Zhi X, Yan H, Fan J, Zheng S (2018) Efficient discriminative clustering via QR decomposition-based linear discriminant analysis. Knowl Based Syst 153:117–132