International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Hybrid Optimized Machine Learning Model for Non-Invasive Procedures Based Early Diagnosis of Hepatocellular Carcinoma Using Novel Biomarker

Babitha Thamby*[1], S Sheeja[2]

Submitted: 29/01/2024 **Revised**: 07/03/2024 **Accepted**: 15/03/2024

Abstract: The primary liver cancer, Hepatocellular Carcinoma (HCC) is found mainly in the alcoholic and non-alcoholic patients. The detection of the disease in the early stage is having a vital role as it is having less or no symptoms till the final stage of the progress. In this paper, a hybrid machine learning model is presented for the detection of HCC in early stage. Here we use a novel biomarker PIVKA II (protein-induced by vitamin K absence). We can use it in combination with traditional routine biomarker Alphafeto protein as well as others. The paper proposes a metaheuristic optimization-based embedded method of feature selection using Least Absolute Shrinkage and Selection Operator (LASSO) blended with Particle Swarm Optimization (PSO). Classification done with three different algorithms called Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), and K Nearest Neighbor (KNN) algorithms. First model emphasizes on the lasso regression and cross validation with the classifiers. Second model proposes PSO optimized LASSO with the classifiers. The proposed second model combines the advantages of both embedded LASSO and PSO algorithms to obtain the best classification results. Using the latter model, among three classification algorithms, LASSO and PSO optimized SVM showed much elevated level of classification results. The proposed LASSO and PSO based SVM showed an elevated accuracy rate of 88.1%, f1 score 91.2% and true positive rate 91.7%.

Keywords: *Hepatocellular carcinoma, detection, machine learning, hybrid, PIVKA II, Lasso, PSO.*

1. Introduction

HCC is one of the most seen and silent cancers around the globe. It is found in both alcoholic and non-alcoholic patients. The studies showing that it mainly affects the male category humans above 60 years of age [1]. The surveillance of HCC provides early detection and there by increases the chance of potentially best treatments. Early-stage detection is amenable to give better clinical remedies like local ablation, surgical resection and sometimes liver transplantation. Mining algorithms have contributed a massive step in the diagnosis of HCC (Most seen cancer of liver) [2]. Potential prediction capabilities provide clinical decisions and a quality life style for the HCC patients. The metastatic condition of tumors could be prevented to a great extent by early diagnosis. Since less or no symptoms are associated with HCC, later diagnosis often leads to economic burden, risky clinical procedures, low level quality of life. Early prediction also helps the healthcare sector since there are some reversible conditions of liver to a great extent.

AlphaFeto Protein (AFP) was the key biomarker used in the diagnosis of HCC [3]. But the problem is that not all the tumors produce elevated level of AFP. So, the biomarker could not be used as a strong and reliable biomarker in the diagnosis [4]. Here comes the importance of novel biomarker DCP (Des-gamma Carboxy Prothrombin) also called PIVKA-II [5]. We can combine the PIVKA with other biomarkers so as to detect HCC in its early stage. The other biomarkers list is provided in this article.

HCC affects both alcoholic and non-alcoholic patients. We have taken non-alcoholic patients for our study since it is the most considerable part of the study as this disease is silent until the last stage of its detection. HCC in Non-Alcoholic Fatty Liver Disease patients (NAFLD-HCC) was our primary consideration [6]. Also, non-alcoholic steatohepatitis (NASH), Hepatitis B or C virus infected persons are there coming under the consideration. Our study mainly focused on the patients in the southern part of India, especially in Kerala region. A retrospective cohort study was done with a total of 400 patients as participants out of which 262 patients were diseased and left were healthy individuals. Majority of the patients considered in our study were above the age of 50.

1.1. Diagnosis

Several methods are used for the early prediction of HCC. Both invasive and non-invasive diagnosis [7][8]. Invasive methods like biopsy, laparoscopy, Hepatic Venous Pressure Gradient (HVPG), Fine Needle Aspiration Cytology (FNAC), Endoscopy etc. are some of the common methods used in the diagnosis of the disease. But as we said earlier, the disease is mainly affected aged human beings the mentioned invasive procedures are risky. Age and physical condition of the patients are the prior consideration while doing the above clinical methods since they are painful and sometimes affects the patient's mental trauma up to an extent. The main disadvantage is that the sample taken from the tumor may or may not contain the malicious cells of carcinoma which contribute more to the active disease. So here we are considering the non-invasive methods since studies are showing that they are having a prominent role in the diagnosis than the invasive methods [9]. The invasive procedures also need a secondary

[1] *Karpagam Academy of Higher Education, Coimbatore-641021, India. ORCID ID: 0000-0002-3033-8243*
[2] *Karpagam Academy of Higher Education, Coimbatore-641021, India. ORCID ID: 0000-0003-4086-4551*
**babitha86@gmail.com*

confirmation by using the non-invasive procedures. Here we are focusing the non-invasive based clinical values into consideration for our study [10]. We have taken a hybrid approach using the blood serum biomarkers mainly, in association with the tumor nature, tumor size from the MRI, CT image findings.

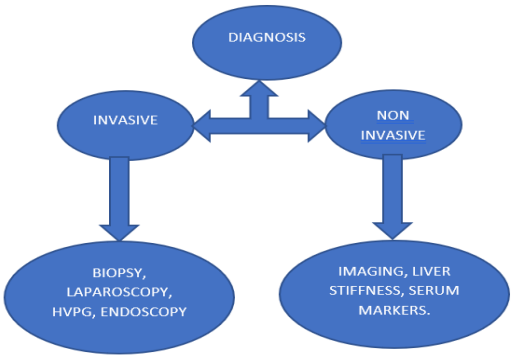

Fig. 1. Invasive and non-invasive methods of diagnosis.

1.2. Biomarkers

Among the conventional blood biomarkers like protein biomarkers, AFP was the only widely used biomarker for HCC detection and further monitoring; however, AFP is not considered to be a strong biomarker nowadays. Most of the patients with HCC do not show elevated levels in AFP. So, it should not be considered as a prominent marker. Instead of that we can use PIVKA II. Studies in the area of hepatology showing that PIVKA-II can be considered as an excellent biomarker for the sole detection of HCC worldwide. So, we add that biomarker to the routine biomarker list of HCC predictors [11]. Other features were also taken into consideration.

Age and Gender are taken into consideration for the biomarkers list as most of the patients were of males above 60 years.

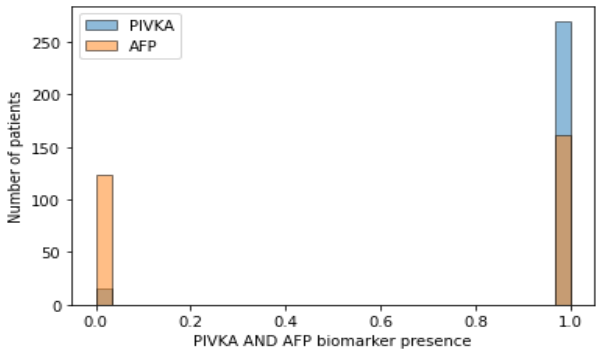

Fig. 2. PIVKA and AFP biomarker statistics comparison among patients.

The statistics are showing that the PIVKA II is more crucially positively detected among the patients than AFP. That means PIVKA II should be considered as a strong prevalent biomarker for the diagnosis of HCC.

2. Related works

Here we are mentioning some related works in the field of HCC early diagnosis using noninvasive methods. In a study of Sato, Masaya, et al., they used machine learning method gradient boosting with optimized parameter, provided the high score of accuracy in the occurrence of HCC (>87%) and developed an area under the curve (AUC) as 0.94 by the usage of cut-off value of 200 ng/mL and 40 mAu/mL for AFP and PIVKA respectively with other biomarkers [12]. Another work done by [13] integrated three diverse algorithms. Linear Discriminant Analysis (LDA) is used to lessen the dimensions. SVM was chosen for classification and Genetic Algorithm (GA) to optimize SVM. The three models were joined and one black box model, and achieved accuracy > 90%, sensitivity >82%, specificity >96%. The study done by Wu, Chieh-Chen, et al [14] with a total of 577 patients to detect fatty liver was done with for algorithms. The area under the receiver operating characteristic (AUROC) was 0.93 for Random Forest, 0.89 for Naïve Nayes, 0.89 for Artificial neural network, and 0.85 for linear regression respectively and accuracy was 87.48, 82.65, 81.85, and 76.96% respectively. Another implementation study was done by Saraswathi, V., S. Anitha Jebamani, and D. Dev [15]. In the primary phase, they took six classification algorithms Logistic Regression, K Nearest Neighbor, Support Vector Machine, Naïve Bayes, decision tree and Random Forest are compared by classification performance metrics. Then, Logistic Regression reached a higher accuracy of approximately 72%. In second phase, they took a sample with reduced accuracy by applying Random Forest (RF) and Decision Tree (DT). Tuning was done. The results of accuracy showed 4% and 12% hike by RF and DT respectively. Another experiment was done by Fathi, Mohammad, et al., [16] to classify the patients to fatty liver positive and negative. Linear SVM, Quadratic SVM and Gaussian SVM were applied. The results showed 91%accuracy, 89%sensitivity, and 94%F1-score with first dataset while it was 92%,89% and 94% with second dataset BUPA.

3. Methodology

The primary collection of the dataset was done from the Kerala region of South India. The dataset contains both HCC positive as well as HCC negative patients. It mainly contains most participants/patients who are above 50 years of age, and majority among them are of above 60 years of aged males. The features in the dataset contains blood serum biomarkers that are relevant for HCC prediction as well as another pathological value like tumor size are taken into consideration. The tumor size greater than 2cm are taken into consideration.

The dataset has undergone bootstrapping. It can create multiple subsets of the data, thereby helps to lower overfitting and improve the accuracy of final results. This study's primary objective is to hybridize the Embedded method Lasso with PSO optimization to find out the best feature out of the dataset attributes. Then we use a classifier for predicting whether the patient will be having HCC or not. By integrating PSO into the feature selection process of embedded model, the hybrid approach is aiming to overcome the limitations of feature selection techniques and explore the full potential of predicting nature by learning methods. Here we are comparing the results getting from LASSO feature selection with different classifiers, and LASSO-PSO feature selection with the same classifiers. Finally, we designed a novel hybrid prediction with LASSO-PSO-SVM. The core contributions of our article are:

(1) First Embedded method, LASSO regression applied for feature selection with cross validation, and classification algorithms were applied for prediction. Hyperparameter tuning done.

(2) Second method of LASSO-PSO hybrid feature selection applied. LASSO is trained for filtering of features to obtain

the prominent elements of HCC prediction. PSO is used for global hunting of features. The selected features undergone for further classification and done hyperparameter tuning.

(3) The proposed second method, which can be used in the early diagnosis of HCC since it is having high accuracy rate than the first method in terms of performance metrics.

(4) The proposed novel method can be deployed for the early diagnosis of HCC.

(5) Moreover, the novel biomarker PIVKA II's contribution in detection could be an advantage in the steps of diagnosis.

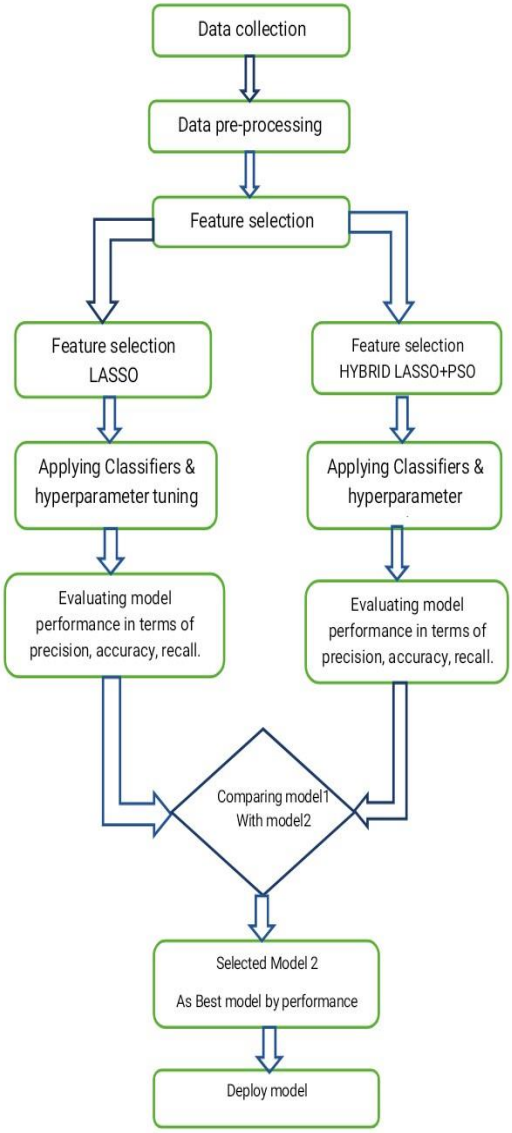

Fig. 3. The flowchart for the model comparison and selection.

4. Materials and Methods

4.1. Dataset

Our work was based on machine learning algorithms using the collected HCC dataset. The data collected retrospectively from Amala Institute of Medical Sciences, Thrissur and from other regional clinics. We focused on the HCC patients above 45 years. There were 365 participants from Kerala region in which 233 participants were HCC patients. The features (26 as input to ML models and 1 for target class) are detailed as follows:

Table 1. Attributes for the detection of HCC.

Attributes	Description
AGE	Age of the participant
GENDER	Sex category of the participant
VIRUS	Hepatitis B or C virus (Infected-1, non-infected-0)
TUMOUR	Tumour in the liver (normally taken those with >2cm)
PVT	Portal Vein Thrombosis (if yes-1, No-0)
PHTN	Portal Hypertension (if yes-1, No-0)
CIRRHOSIS	Whether the person is having cirrhosis or not (if yes-1, No-0)
NASH	Non-alcoholic Steatohepatitis (if yes-1, No-0)
LIV_STIFF	Liver Stiffness (in kPa) more than 15(if yes-1, No-0)
PIVKA	Protein Induced in presence of Vitamin K Antagonist, normally >40 mAU/mL in HCC cases
AFP	Alphafeto protein normally > 40 ng/mL in HCC cases
Hb	Haemoglobin, considerable if, males <14g/dl, females <12 g/dl
RBC	Red Blood Cells, normal range
PLATELET	Platelet Count
NEUTRO	Neutrophil
LUMPHO	Lymphocyte
CREAT	Creatinine
TOT_BIL	Total Bilirubin
DIR_BIL	Direct Bilirubin
SGOT	Serum glutamic oxaloacetic transaminase, an enzyme produced by the liver
SGPT	Serum glutamate pyruvate transaminase, an enzyme produced by the liver
ALP	Alkaline Phosphate
A/G	Albumin-Globulin ratio
NA	Sodium
K	Potassium
INR	International normalized ratio (indicated the clotting time of blood)
CLASS (Target variable)	Whether the patient is having HCC (1) or not (0)

4.2. Data Preprocessing

The data collected in a retrospective mode of study, contain 27 predictive attributes or input attributes. The final attribute named 'CLASS' is the target variable which defines whether the patient

will be having HCC or not. The dataset contains both categorical and numeric data. Initially the dataset had albumin and globulin attributes separately, but we merged the two columns together to make Albumin-Globulin ratio (A/G) as a single attribute. The non-important blood serum values independent on the disease were omitted manually at the initial stage. All the operation operations were performed in Anaconda Jupyter notebook 6.4.8 using python as our language. The attribute 'GENDER' contained the data whether the patient is male or female. The data is converted using a Labelencoder module with the help of scikit-learn library.

The missing value imputation is done by doing Mean imputation. We fill the missing values of a particular variable with the calculated mean of non-missing cases of that variable.

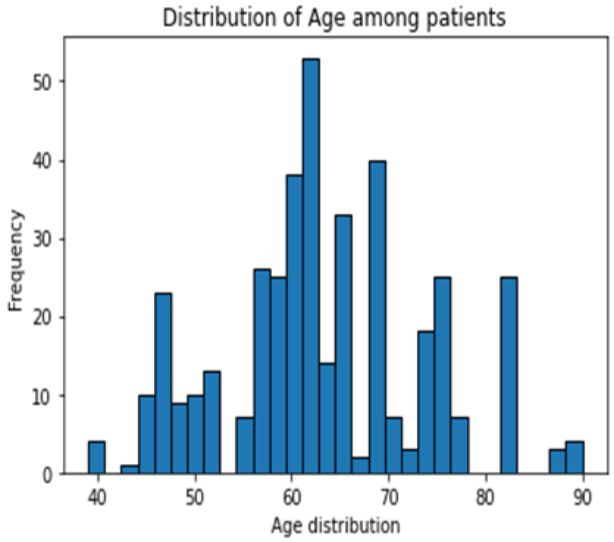

Fig. 4. Frequency of Age among various patients of HCC.

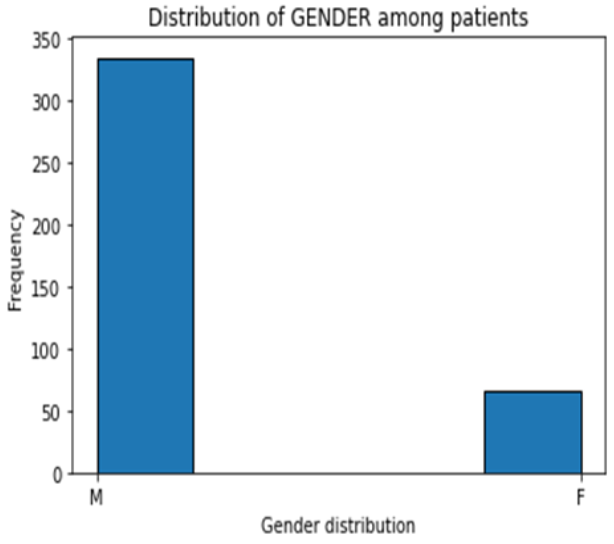

Fig. 5. Frequency of gender among various patients of HCC.

The above two statistics are showing that most of the patients are males, and above 60 years of age.

4.3. Feature selection

Feature selection is a process in machine learning by which the more relevant features are selected to improve model performance and for reducing complexity. We can avoid overfitting and thereby stimulate the model interpretability. We use filter methods, wrapper methods, as well as embedded methods, according to our choice.

KBest is a feature selection method used in machine learning. It's part of filter methods and involves selecting the top k features based on statistical tests like f_classif. This helps to focus on the most informative features, improving model efficiency and reducing dimensionality. The k-best feature selection method typically involves ranking features based on a certain criterion and selecting the top k features for their selection.

The process generally follows these steps:

1. Score calculation: Use a statistical test (like F-statistic or chi-square) to calculate a score for each feature.
2. Feature ranking: Rank the features based on their scores, which is calculated above using the score function.
3. Select top ranked features: Choose the top ranked k features with the highest scores by giving specified k values.

The dataset is firstly sliced into training set, to train the model and testing set, for evaluating the performance of the model, as 80% of data is taken for training and left were taken for testing with random state.

4.3.1 Feature Selection-Methodology I (LASSO)

After dividing the dataset into training and testing sets, we applied LASSO feature selection method. LASSO regression is applied as a regularization method to prevent overfitting by adding term of penalty, dependent on the coefficients' absolute values. The parameter alpha of the regression can be given a value depends on the regularization will enable feature selection by pushing some of the coefficients to zero. As it will be making some coefficients to zero, we can neglect them. We will select the coefficients that are not equal to zero. We can call them as our selected features.

Let linear regression models predict the outcome based on linear combination of predictable variables.

$$Y= \beta 0 + \beta 1\ x1 + \beta 2\ x2 + \beta 3\ x3 + \ldots\ldots + \beta n\ Xn \qquad (1)$$

y is called the target variable.
$\beta_0, \beta_1, \beta_2, \ldots, \beta n$ are the coefficients or parameters to estimate.
$x_1, x_2, \ldots, xn$ are the independent variables called features.
ε is representing the term of error.

The values of the linear regression coefficients are generally calculated by minimizing the squared difference between the original and predicted value of y.

$$\text{minimize}\left\{\sum_{j=1}^{N}(yj - \beta 0 - \sum_{i-1}^{n}\beta i\ \ xij)2\right\} \qquad (2)$$

this is known as least squared loss.

If the features in the dataset are highly-dimensional then, linear models have a tendency to overfit the data. To cancel the same, the search for the optimal coefficients can be done with regularization method. Mainly there are two regularization procedures: the Ridge and the Lasso regularization. With the Lasso regression, estimation of coefficients can be done by minimizing the following equation. LASSO has an additional term of penalty which depends on the coefficients' absolute values. The L1 regularization term is equal to the sum of coefficients' absolute values, multiplied by a tuning parameter λ.

(3)

$$minimize\{\sum_{j=1}^{N}(y_j - \beta_0 - \sum_{i=1}^{n}\beta_i x_{ij})^2 + \lambda\sum_{i=1}^{n}|\beta_i|\}$$

- λ is the regularization parameter, controls the regularization amount. Also known as alpha ά.
- $\beta_1, \beta_2, ..., \beta n$ are the coefficients.

Hyperparameter Tuning: The value of the hyperparameter α will control trade-off between model building and accuracy. The optimal value of α is found through cross-validation.
The below code explains the same:

```
# Number of Folds
kv=KFold(n_splits=5,shuffle=True, random_state=42)
# Initialization of the Model
lasso_m = Lasso ()
# GridSearch Cross Validation
lasso_cv=GridSearchCV(lasso_m, param_grid=params, cv=kv)
lasso_cv.fit(X, y)
```

Here we have given the alpha value as 0.00001 to get the selected features. We applied Grid search cross validation to find the optimal solution values for LASSO, with 5-fold cross validation criteria by assigning number of folds as 5.
The best parameter was found to be {'alpha': 1e-05}. That is why we have given the alpha parameter value as 0.00001. The features selected were considered as our feature subset for further binary classification.

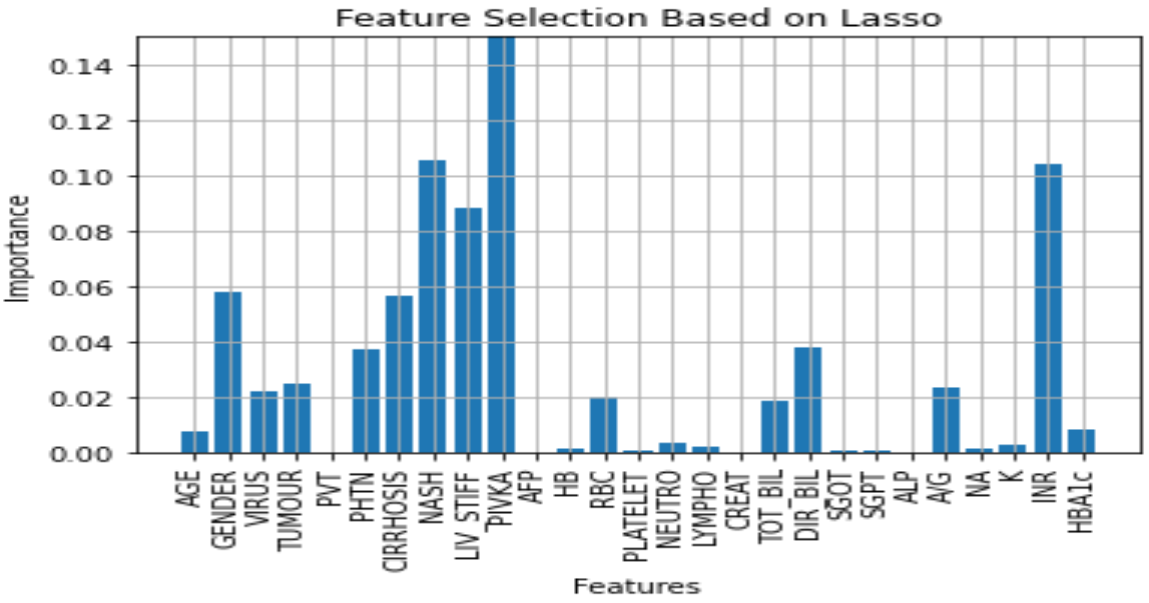

Fig. 6. Feature selection based on LASSO with cross validation.

4.3.2 Feature Selection-Methodology II (LASSO+PSO)

We had tried a second method of hybrid feature selection in order to check whether we can improve the accuracy of classifiers by making a hybrid selection of features. Not only that, our aim is to develop a model for lowering the false positive rate and to improve the true positive rate. We made many combinations with LASSO and finally Particle Swarm Optimization (PSO) combination helped us to reach our goal [17].
Compared to LASSO, Optimal Feature Selection using PSO can reduce falsely chosen features by a half by maintaining the constant level of true positive rates [18]. So, it is more efficient for selecting the right variables, and the final output will be more interpretable, and more accurate. It is a metaheuristic algorithm contributed by the concept of the social behavior of fishes and birds. PSO is tuned with a few parameters, helps to get optimal solution, and also effective in terms of computational cost [19].
The feature selection based upon PSO has so many advantages. Mainly it has a capability of global search. It can explore the space by maintaining a population of possible potential solutions called particles. It can reach up to global optimum. That means it can give an optimal feature combination for a better performance. Its adoptability to nonlinear relationship can make it versatile. That means if the relationship between features and the target variables are not adequately captured, even then the PSO can do the best. PSO can maintain a balance between the search space and exploitation of the solution regions. It is very flexible with different optimization scenarios, even in the high dimensional spaces. It is computationally efficient and faster exploring of solution space [20].
The steps in PSO are:
1. Initialization: - we initialize a population containing particles that are having random positions and different velocities.
2. Evaluating the fitness of every particle according to the current position.
3. Update the local or personal best position of each and every particle by the value of the current fitness.
4. Updating the global best: - find out the particle that is having high fitness among all particles which can give the best solution in the swarm.
5. Update velocities and positions: - Based on the current velocity, position, global best and local best update, the velocities and positions of each particle is updated.
6. Repeat the steps 2-5 by iterations till a convergence criterion is met.
7. Retrieve optimized features.

PSO for LASSO regression uses a fitness function or objective function, aiming to minimize the Mean Squared Error (MSE) by considering the L1 regularization penalize the absolute values of coefficients. We can encourage the sparsity of the model by

doing so. We can calculate the same as

mse_with_lasso=mean_squared_error(y,y_pred)+alpha * sum(abs(lasso_model_coef_) (4)

The primary aim is to minimize the sum of MSE and the L1 term of regression.
Objective function = LASSO+ Penalty

The criteria of LASSO regression in model1 are incorporated with the PSO optimization. An approach of metaheuristic algorithm optimized feature selection is done. The three binary classification algorithms are compared like we have done in model1 and calculated the accuracy, precision, recall, f1 score and true positive rate by doing a five-fold cross validation using grid search.
The selected features were 'AGE','GENDER', 'VIRUS', 'TUMOUR','PVT','PHTN', 'CIRRHOSIS','NASH','LIV_STIFF' ,'PIVKA', 'AFP', 'HB' ,'RBC', 'PLATELET', 'NEUTRO', 'LYMPHO', 'CREAT', 'TOT_BIL' ,'DIR_BIL', 'SGOT', 'SGPT', 'ALP', 'A/G' ,'NA', 'K', 'INR', and 'HBA1c'. RMSE was reduced from 0.1 to 0.06 when calculated.

4.4 Binary Classification & prediction

Based on the selected features using methodology I and methodology II, we have applied binary classification to predict the patient is having HCC or not. We have applied three algorithms Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Linear Discriminant Analysis (LDA)[21][22][23]. We have two models named Model1 and Model2. The results of the two models are described below in detail.

5. Experimental analysis & Results

The results of both model1 and model2 are estimated in terms of accuracy, precision, recall, and true positive(TP) rate. They were calculated as:

$$accuracy = \frac{true\ positives + true\ negatives}{true\ positives + true\ negatives + false\ negatives + false\ positives} \quad (5)$$

$$precision = \frac{true\ positives}{true\ positives + false\ positives} \quad (6)$$

$$recall = \frac{true\ positives}{true\ positives + false\ negatives} \quad (7)$$

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (8)$$

In model1, we are applying the three above mentioned binary classifiers for prediction, based on the feature selection relying on simple LASSO with cross validation.
The results showing that SVM detects 74.5% of the participants are having truly identified that they are having HCC while LDA has 97.8%, and KNN has 97.9%. The accuracy of SVM, LDA, KNN were 82.5, 95 and 98.1 respectively. The results were shown in the below table.

Table 2. Results of binary classification using algorithms in model1

Algorithm	Accuracy	Precision	Recall	F1 score	TP
SVM	82.5	94.5	74.5	83.3	74.5
LDA	95.0	93.8	97.8	95.8	97.8
KNN	98.1	98.9	97.8	98.4	97.9

The bar chart for the above results is below:

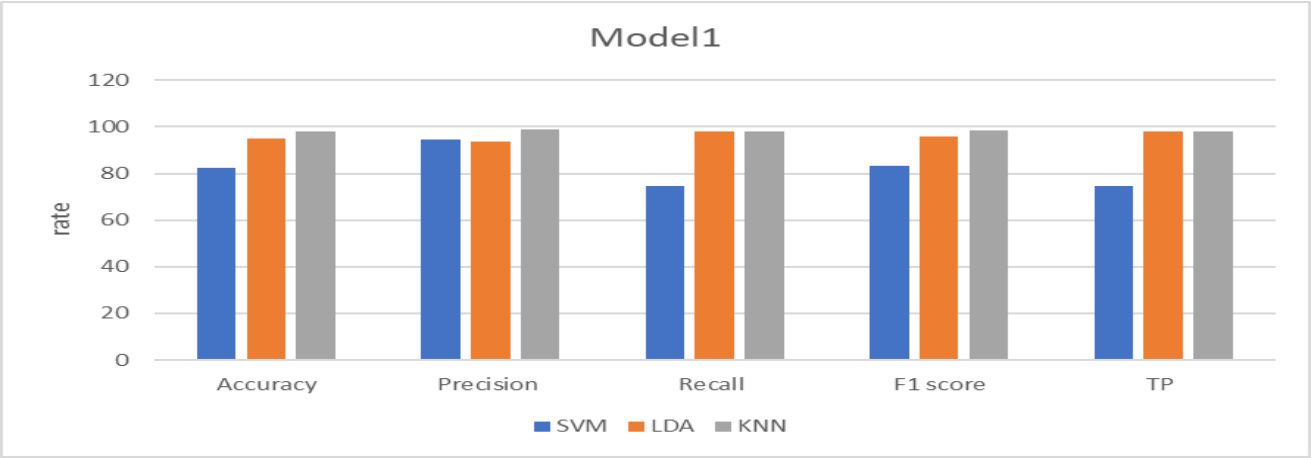

Fig. 7. Model 1- Performance matrices of different algorithms using feature selection methodology I

The findings from the first model showed that SVM has a low performance rate when compared with others. So, our aim was to improve the accuracy as well as producing a high true positivity rate mainly in the case of SVM. Results of other algorithms were also computed. We tried many feature selection methods and found that we can achieve the target by doing a hybrid method for feature selection. Then we opted PSO optimized LASSO as our next method to reach the goal.

Here we achieved improved accuracy and true positive rate compared to the first model. SVM has the large variance in accuracy, as it showed 88.1% of accuracy and a high elevated TP rate of 91.7%. LDA has also improved in its results of performance. There was 1.9% increase in the LDA accuracy. But KNN showed no more elevation in its performance metrices. The model 2 details are given below.

Table 3. Results of binary classification using algorithms in model2

Algorithm	Accuracy	Precision	Recall	F1 score	TP
SVM	88.1	90.8	91.7	91.2	91.7
LDA	96.9	98.1	97.2	97.7	97.2
KNN	98.1	98.9	97.8	98.4	97.9

The bar chart for the above results is below:

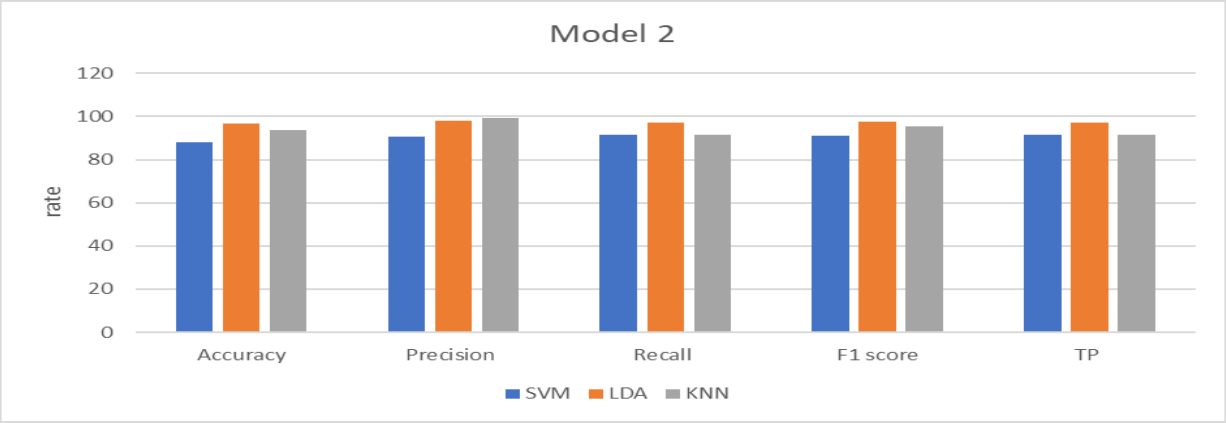

Fig. 8. Model 2- Performance matrices of different algorithms using feature selection methodology II.

We derived two models as model1 and model2. As we discussed earlier Model1 is having a scenario of low accuracy and True positive rate in the case of SVM, when compared with other models. So, we derived another hybrid method as Method 2, and we can improve the performance matrices of algorithms. In Model1 for the case of SVM, we had an accuracy, precision, recall, f1 score and tp rate as 82.5%, 94.5%, 74.5%, 83.3%, and 74.5% respectively. So, we tried to implement another boosted hybrid feature selection named as Model 2 (LASSO-PSO-SVM) and found that SVM achieved a high-performance rate difference compared to LDA and KNN. It achieved accuracy 88.1%, precision 90.8%, recall 91.7%, f1 score 91.2% and tp rate 91.7%. The other algorithms also showed improvement, but not a

6. Proposed Model

In the proposed model, at first stage, PSO algorithm is applied to do global search in order to reduce dimensions as well as to reduce the time of computation for the second stage. During the second stage, LASSO will be appointed to do the feature selection in order to avoid overfitting and for the better performance. PSO can find optimal parameters for LASSO model of regression. It can control the amount of shrinkage of the coefficients. The results shows that the proposed novel method LASSO-PSO-SVM shows an increase in the accuracy, f1 score and tp rate.

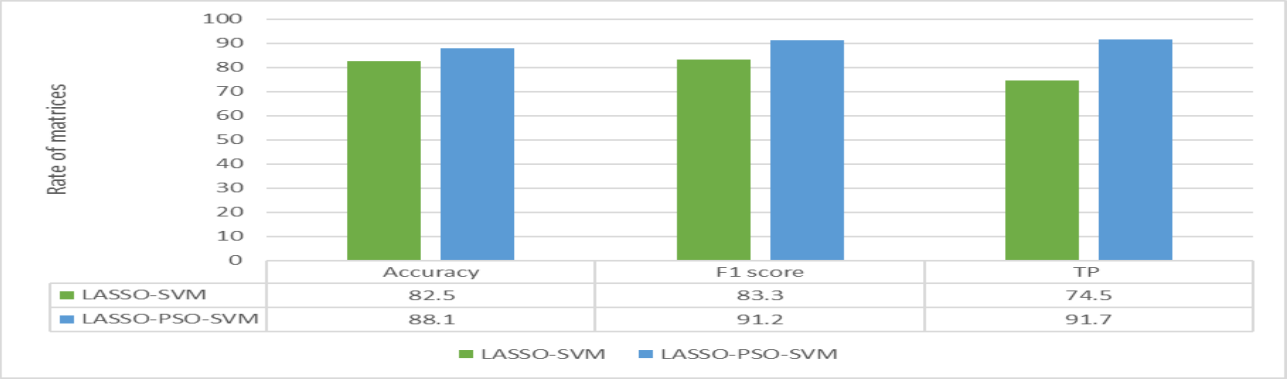

Fig. 9. Graph showing performance increase with the proposed model.

7. Conclusion

The system proposed in the paper can give a novel hybrid optimization method in clinical decision-making capability in healthcare diagnostic area to identify HCC in non-alcoholic patients from certain data at an early stage that will allow to commence the remedial precautions and solutions for a quality

life style and increase in life span. Also in some cases, the early findings are early decision makers, as the physical condition of the patient is often an important factor for some procedures like liver transplantation and liver resection. Here we used a novel biomarker PIVKA II, which will be helpful for detection in most of the patients since traditional biomarker AFP is still exist as normal, in them. The study shows that HCC is more prevalent in males more than in females, who are between 50 and 70 years of age. The proposed algorithm LASSO-PSO-SVM performs well with the data and provide a high accuracy (88.1%), f1 score (91.2%) and True positivity (91.7%) with a hike of +5.6% in accuracy as well as more than 15% increment in the rate of true positivity.so that we can use the algorithm as an efficient and convenient method in the diagnosis of HCC as well as for other carcinoma detection.

Author contributions

All authors contributed in writing and reviewing the main manuscript.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] 1T. C. Yip et al., "Impact of age and gender on risk of hepatocellular carcinoma after hepatitis B surface antigen seroclearance," J. Hepatol., vol. 67, no. 5, pp. 902-908, 2017.

[2] K. A. McGlynn et al., "Epidemiology of hepatocellular carcinoma," Hepatology, vol. 73 suppl. 1, pp. 4-13, 2021.

[3] F. Özdemir and A. Baskiran, "The importance of AFP in liver transplantation for HCC," J. Gastrointest. Cancer, vol. 51, no. 4, pp. 1127-1132, 2020

[4] T. Inoue and Y. Tanaka, "Novel biomarkers for the management of chronic hepatitis B," Clin. Mol. Hepatol., vol. 26, no. 3, p. 261-279, 2020.

[5] H. Hadi et al., "Utility of Pivka-II and AFP in Differentiating Hepatocellular Carcinoma from Non-malignant High-risk patients," Medicina (Kaunas), vol. 58, no. 8, p. 1015, 2022

[6] G. N. Ioannou, "Epidemiology and risk-stratification of NAFLD-associated HCC," J. Hepatol., vol. 75, no. 6, pp. 1476-1484, 2021

[7] L. T. de Lima Franca, Daniel Broszczak, Xi, "Zhang, Kim Bridle," Darrell Crawford, and Chamindie Punyadeera. "The use of minimally invasive biomarkers for the diagnosis and prognosis of hepatocellular carcinoma." Biochimica et Biophysica Acta (BBA)-Reviews on Cancer 1874, vol. 2, 2020, p. 188451.

[8] T. H. Kim et al., "Comparison of international guidelines for noninvasive diagnosis of hepatocellular carcinoma: 2018 update," Clin. Mol. Hepatol., vol. 25, no. 3, p. 245-263, 2019.

[9] G. Marasco et al., "Non-invasive tests for the prediction of primary hepatocellular carcinoma," World J. Gastroenterol., vol. 26, no. 24, p. 3326-3343, 2020.

[10] T. H. Kim et al., "Comparison of international guidelines for noninvasive diagnosis of hepatocellular carcinoma: 2018 update," Clin. Mol. Hepatol., vol. 25, no. 3, p. 245-263, 2019.

[11] F. Qi et al., "The diagnostic value of Pivka‐II, AFP, AFP‐L3, CEA, and their combinations in primary and metastatic hepatocellular carcinoma," J. Clin. Lab. Anal., vol. 34, no. 5, p. e23158, 2020.

[12] M. Sato et al., "Machine-learning approach for the development of a novel predictive model for the diagnosis of hepatocellular carcinoma," Sci. Rep., vol. 9, no. 1, p. 7704, 2019.

[13] L. Ali et al., "LDA–GA–SVM: Improved hepatocellular carcinoma prediction through dimensionality reduction and genetically optimized support vector machine," Neural Comput. Appl., vol. 33, no. 7, pp. 2783-2792, 2021.

[14] C. C. Wu et al., "Prediction of fatty liver disease using machine learning algorithms," Comput. Methods Programs Biomed., vol. 170, pp. 23-29, 2019.

[15] V. Saraswathi et al., "Implementation of hyper parameter optimization in liver disease prediction" in International Conference on Power, Energy, Control and Transmission Systems (ICPECTS). IEEE, 2022, pp. 1-6.

[16] M. Fathi et al., "A machine learning approach based on SVM for classification of liver diseases," Biomed. Eng. Appl. Basis Commun., vol. 32, no. 3, p. 2050018, 2020.

[17] R. K. Huda and H. Banka, "Efficient feature selection and classification algorithm based on PSO and rough sets," Neural Comput. Appl., vol. 31, no. 8, pp. 4287-4303, 2019.

[18] Y. Xiong et al., "An efficient gene selection method for microarray data based on lasso and BPSO," BMC Bioinformatics, vol. 20 suppl. 22, pp. 715, 2019.

[19] M. Amoozegar and B. Minaei-Bidgoli, "Optimizing multi-objective PSO based feature selection method using a feature elitism mechanism," Expert Syst. Appl., vol. 113, pp. 499-514, 2018.

[20] S. Binti et al. "Particle swarm optimization feature selection for breast cancer recurrence prediction.", "Sakri," IEEE Access, vol. 6, pp. 29637-29647, 2018.

[21] K. Wang et al., "Software defect prediction model based on lasso–SVM," Neural Comput. Appl., vol. 33, no. 14, pp. 8249-8259, 2021.

[22] S. Afrin et al., "Mst Fahmida Muntasim", Md Shakil Moharram, M. M. Imran, and Md Abdulla. "Supervised machine learning based liver disease prediction approach with LASSO feature selection." Bulletin of Electrical Engineering and Informatics 10, vol. 6, 2021, pp. 3369-3376.

[23] G. R. Nitta et al., "Lasso-based feature selection and naïve Bayes classifier for crime prediction and its type," Serv. Oriented Comput. Appl., vol. 13, no. 3, pp. 187-197, 2019.