# Classification of Liver Disease Using Conventional Tree-Based Machine Learning Approaches with Feature Prioritization Using a Heuristic Algorithm

Proloy Kumar Mondal[1], Haewon Byeon[2]*
Department of Electronics and Communication Engineering, Khulna University, Khulna, Bangladesh[1]
Department of Digital Anti-aging Healthcare (BK21), Inje University, Gimhae 50834, South Korea[2]

***Abstract*—Liver disease ranks as one of the leading causes of mortality globally, often going undetected until advanced stages. This study aims to enhance early detection of liver disease by employing machine learning models that utilize key health indicators. Utilizing the Indian Liver Patient Dataset (ILPD) from the UCI repository, we developed a predictive model using the CatBoost algorithm, achieving an initial accuracy of 74%. To improve this, feature selection was performed using the Whale Optimization Algorithm (WOA) and Harris Hawk Optimization (HHO), which increased accuracy to 82% and 85% respectively. The methodology involved preprocessing to correct data imbalances and outlier removal through univariate and bivariate analyses. These optimizations highlight the critical features enhancing the model's predictive capability. The results indicate that integrating metaheuristic algorithms in feature selection significantly improves the accuracy of liver disease prediction models. Future research could explore the integration of additional datasets and machine learning models to further refine predictive capabilities and understand the underlying pathophysiology of liver diseases.**

***Keywords*—Liver disease; classification; prediction; CatBoost algorithm; machine learning; optimization algorithm**

## I. Introduction

The liver is an important part of the body that conducts functions such as gall generation, chemical detoxification, and a supply of critical proteins for blood [1]. A huge increase in different liver illnesses has been observed the world in recent years. About two million people are diagnosed with liver disease each year, with one million dying from cirrhosis complications and one million from viral hepatitis and hepatocellular carcinoma. Because cause-specific death data is scarce in many places where liver disease is common, notably in Africa, accurate figures are not always accessible. Furthermore, nearly one-third of the world's countries lack reliable mortality statistics. Even in industrialized nations, it is impossible to distinguish the burden of liver disease according to the cause and stage of the disease [2]. Cirrhosis is the 11th leading cause of death worldwide, while liver cancer is the 16th, an estimated 1.16 million and 788,000 people die each year. They are responsible for 3.5 percent of all deaths worldwide [3]. Liver-related deaths accounted for 3% of all deaths worldwide in 2000. They are ranked 13th (cirrhosis) and 20th (liver cancer). However, the effects can be even greater if acute hepatitis and alcohol use are considered major factors. According to these Fig. 1, the liver disease dies over two million people worldwide each year. Due to worldwide population pressure, India accounts for one-fifth (18.3%) of all cirrhosis fatalities while China's contribution is 11 In Central Asia and the Russian Federation, mortality is increasing. In the UK, mortality is increasing, but in France and Italy, it is decreasing. Males are affected by cirrhosis at a higher rate than females all over the world [2]. Refer to Fig. 1. Liver disease and cancer are among the leading causes of death worldwide.

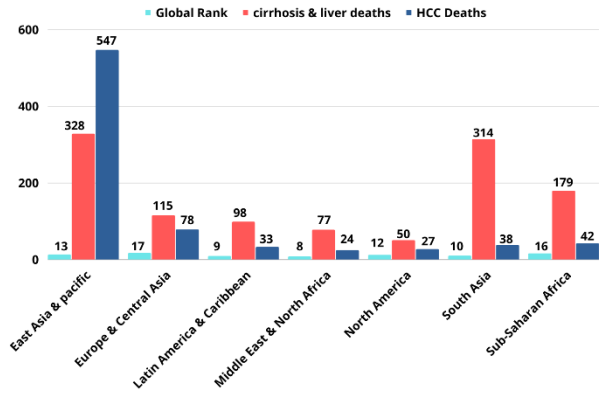


Fig. 1. Global mortality from liver illness and liver cancer.

First, the number of patients with liver disease is increasing every year, however the number of specialist doctors is not increasing. As a result, it has become impossible to diagnose the disease or serve the patient well. It takes a lot of doctors to monitor patients with liver problems which can be very challenging. If we can collect human data in every hospital and every clinic then this process will be much easier for everyone and easy to manage. However, by analyzing the data of these people, we can easily detect the symptoms of the disease if ML is applied. As a result, the number of doctors will be less and the process will be much more comfortable. Artificial Intelligence (AI) includes ML, which allows the system to learn without information. Human inputs and outputs are employed in the training process and prediction accuracy of supervised algorithms, which are used in a variety of classification applications [4]. Fig. 2 shows the five causes of liver failure.

* Corresponding Author

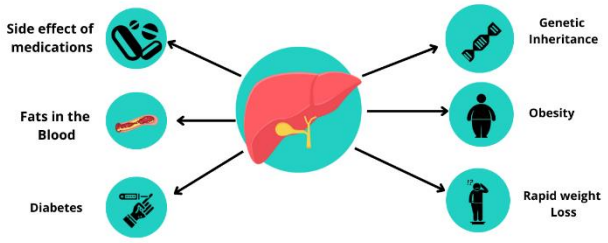


Fig. 2. Five cause of liver failure.

ML is making a significant contribution to healthcare and is expanding day by day. One of the most important problems in healthcare is the growing number of liver patients. The incidence of fatty liver in liver disease is early stage and cirrhosis is the final stage of chronic liver disease which later leads to liver cancer. Many data mining techniques and medical data mining techniques help to present and predict liver disease first and foremost. As a result, the use of this technique greatly reduces the doctor's work.

This paper is organized as follows: Section II provides an overview of related work and highlights the main differences between our work and other existing studies. Section III presents the research methodology, experimental details, configuration and system flowchart. The test results and analysis are discussed in detail in Section IV. Finally, the paper is concluded in Section V.

## II. Related Work

Disease prediction has become possible by uncovering hidden features in medical datasets using machine learning algorithms. Different types of datasets, such as blood panels with liver function tests, histologically stained slide images, and the presence of specific molecular markers in blood or tissue samples, have been used to train classifier algorithms to predict liver disease, which provided good accuracy. Machine learning methods described in previous studies have been evaluated for accuracy using a combination of confusion matrix, area under the receiver operating characteristic curve, and k-fold cross-validation. In study [2], the authors studied the prognosis of liver disease and used genetic algorithm combined with XGBoost to predict liver disease and analyzed from the test that the algorithm helped to predict the disease efficiently.

In recent studies, various machine learning algorithms have been applied to improve the diagnosis and prediction of liver-related diseases. In study [3], four machine learning algorithms were tested on ILDP datasets with the Pearson Correlation Coefficient (PCC-FS) optimization technique, resulting in the AdaBoost algorithm achieving a maximum accuracy of 92.19%. Similarly, the study in [4] explored the use of Support Vector Machine (SVM) and Logistic Regression (LR) for diagnosing liver disease, achieving an accuracy of 96%. Additionally, the study in [5] implemented a Random Forest (RF) algorithm to predict liver disease with notable accuracy. Furthermore, the research in [6] focused on the prediction of hepatocellular carcinoma (HCC) using the RF algorithm, achieving an accuracy of 80.86%. These studies collectively highlight the potential of machine learning techniques in enhancing the accuracy of liver disease diagnosis and prediction.

In study [1], the authors in this paper help to identify the patient's liver disease from the data and contribute to the field of medical science so that treatment can be started and the disease can be cured before it becomes severe. To do this they first used the classifier model decision tree (DT) algorithm and achieved the highest accuracy. Then they use seven more classifier algorithms: RF, LR, SVM, K-nearest neighbors (KNN), linear discriminant analysis, AdaBoost, and gradient boosting. Then they used the least absolute shrinkage and selection operator (LASSO) feature selection technique to achieve better accuracy.

Furthermore, recent research has delved into various machine learning techniques to enhance the diagnosis and prediction of liver-related diseases. In study [7], a diagnostic system for chronic liver infections was developed using six classifiers: Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Naive Bayes (NB), with LR achieving the highest accuracy at 75%. Similarly, the study in [8] utilized LR, SVM, and KNN for liver disease prediction, identifying LR as the most effective. Dhamodharan et al. [9] focused on predicting cirrhosis, liver cancer, and hepatitis, employing Naive Bayes and the FT Tree algorithm, with Naive Bayes providing the highest accuracy. Rosalina et al. [10] used SVM and the Wrapper method for hepatitis prognosis, effectively removing noise features before classification and achieving optimal results by combining these methods. Soliman et al. [11] introduced a hybrid classification system for HCV detection, utilizing Least Squares Support Vector Machine (LS-SVM) and Modified Particle Swarm Optimization (PSO). With Principal Component Analysis (PCA) for feature extraction and a modified PSO for parameter optimization, their method outperformed other systems in accuracy using HCV benchmark data from the UCI repository. Lastly, in study [12], NB and SVM algorithms were applied for liver disease prediction, with SVM achieving the highest accuracy. Collectively, these studies highlight the significant role of machine learning in enhancing the precision and effectiveness of diagnostics for liver diseases.

## III. Methodology

This suggested model uses data from machine learning Indian Liver Patient Dataset (ILPD) that has been taken from the UCI Repository to predict the disease in multiple patients with liver disease. To begin, pre-processed data is used to create "clean" data. The feature extraction approach selects the relevant data from all of the dataset's attributes in order to improve accuracy by using only relevant data. After then, the algorithms and data used to classify the objects were examined. CatBoost classifier Algorithm is used to classify the data throughout the analysis process. Performance is evaluated based on the classification findings. We then employed optimization methods, such as the whale Optimization algorithm, in order to improve our results even further. Algorithms are compared on accuracy, sensitivity, precision and f1-scores in order to determine the best performing algorithm for the system's performance. Fig. 3 depicts the system's overall working procedure.

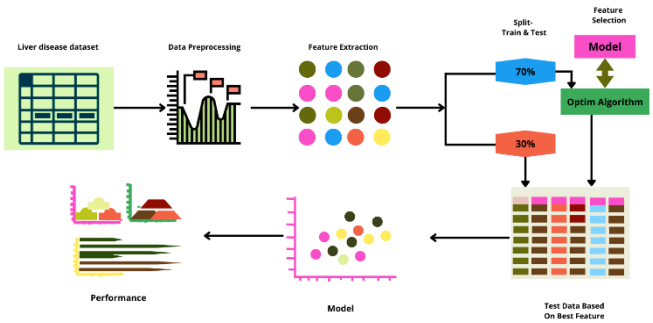


Fig. 3. Working procedure.

*1) Dataset description*: This dataset was collected from the northeastern Andhra Pradesh of India. In addition, this dataset is publicly available in the UCL machine learning repository [13]. There are 583 patients in the ILPD dataset which 441 are males and 142 are females. Anyone over the age of 89 is reported as having an age of 90. There is also a selector field to determine whether the patient having liver disease or not. Non-LD patients (0) total 167, whereas LD patients (1) total 416. Attribute properties of the dataset include multivariate, integer, and real values. Table I represented the dataset contains a total of 11 specific parameters, out of which we selected 10 parameters for our analysis and 1 was used as the target class. These data are used to train and test the models, and the models' performance is assessed based on their own output. In addition, we have divided the dataset into two parts: 70% for training and 30% for testing. Thus, we have 408 samples in our training set and 175 samples in our validation set. In Fig. 4 the dataset's distribution is displayed.

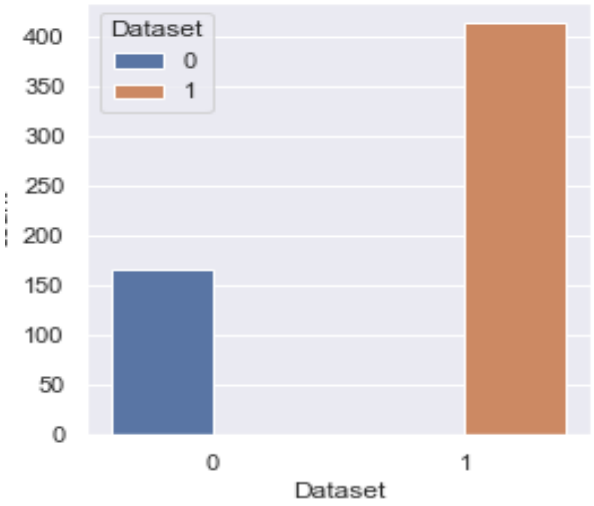


Fig. 4. Dataset description.

TABLE I. DESCRIPTION OF VARIABLES

| Features No | Dataset Information | |
|---|---|---|
| | ***Features Name*** | ***Description*** |
| 1 | Age | Age of the patient |
| 2 | Gender | Gender of the patient |
| 3 | TB | Total Bilirubin |
| 4 | DB | Direct Bilirubin |
| 5 | Alkphos | Alkaline Phosphotase |
| 6 | Sgpt | Alamine Aminotransferase |
| 7 | Sgot | Asparatate Aminotransferase |
| 8 | TP | Total Proteins |
| 9 | ALB | Albumin |
| 10 | AG Ratio | Albumin and Globulin Ratio |

*2) Data Preprocessing*: As stated in the preceding paragraph, the dataset referred to has flaws and scattered data. Pre-processing has been done so that we can get the most out of this dataset. We manually corrected any incorrect data by going through the dataset and looking for any anomalies. When there are no values to fill in, the median of a feature is used. However, Information is extracted from sources and collected in the form of data or discrete analytical data. Each attribute acts as a variable and each instance has specific attributes. Liver disease is predicted using a dataset, which is created through data collection and pre-processing methods. The dataset helps us diagnose the disease based on its various parameters. 10 features are considered to get accurate results in the dataset of the proposed work. Classification is a process of data mining consisting of problem identification. Best performance-based prediction is provided by observing liver disease characteristics in patients and using machine learning algorithms.

*3) Classification and Performance Metrics*: Dorogush et al. [14] developed CatBoost in 2018 based on improvements to XGBoost. Yandex released CatBoost, an open-source machine learning algorithm, in 2017, which is still relatively new [15]. The model is built using a training dataset, which consists of a set of objects with known features and labels. The training dataset is also referred to as the "data set". The validation dataset is only used to evaluate the effectiveness of training and contains similarly organized data, but is not used for training. The basis of CatBoost is gradient-boosted decision trees, where a series of decision trees is generated sequentially during training. Each subsequent tree is built with less damage than the previous tree. The initialization parameter determines how many trees will grow. To prevent overfitting, overfitting detectors are used which stop tree growth when activated. We used CatBoost algorithm to classify liver diseases.

*4) Features selection*: Feature selection is an important process in machine learning, where the most important features (variables or predictors) are selected from the dataset, which play a role in predicting the target variable. This is helpful in reducing dataset dimensions, improving model performance,

and reducing the likelihood of overfitting. In this study, we used two metaheuristic optimization algorithms such as, WOA and HHO. It is well-known swarm-based metaheuristic method for feature selection.

*5) Whale optimization algorithm (WOA)*: WOA is a metaheuristic optimization algorithm that was proposed by Mirjalili and Lewis [16]. The bubble-net hunting method utilized by humpback whales is modelled after and imitated by the algorithm. The entire process can be broken down into three stages: the first stage involves encircling the prey, the second stage involves bubble-net foraging (the exploitation phase), and the third stage involves searching for prey (exploration phase). To better understand these three stages of the WOA strategy we present in Fig. 5. An initial solution candidate is chosen at the beginning of the algorithm, and their fitness is determined with the help of a function. During each cycle of the iterative process, the solution set is either updated through the shrinking encircling mechanism or the spiral updating mechanism (exploitation phase), with the choice of the mechanism being determined by a probability p. In addition, the shrinking encircling mechanism can either update the new solution set so that it is closer to the global best solution or it can use a random search agent. This is determined by a coefficient called A. The equation for the coefficient is presented in the following example:

$$\boldsymbol{A = 2a.r - a} \quad (1)$$

where, a is a random vector in the range [0, 1] and r is a vector that decreases linearly from 2 to 0 over the course of the generation. Because the update that leads to a random agent conducts a worldwide search, the subsequent phase is known as the exploration phase. In Algorithm 1, the pseudo-code for feature selection using WOA is presented.

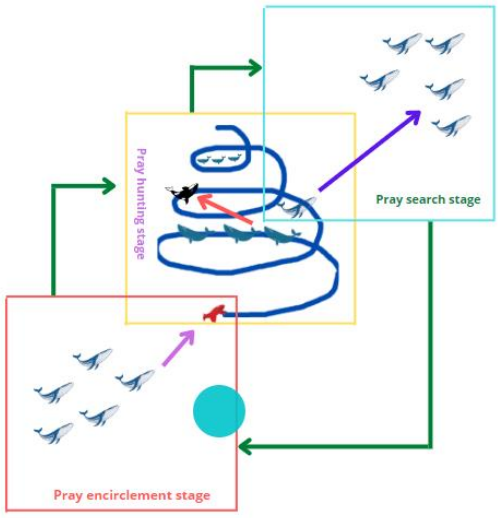


Fig. 5. Whale Optimisation Algorithm.

**Algorithm 1: Feature Selection Using WOA**

Create the first group of n whales xi (1,2,3……., n)
Set the iteration counter $t_{counter}$ = 0
figure out how fit each whale is.
figure out which whale is the fittest, i.e., $Y_{best}$
**for** each whale **do**
  Decode whale position
  Find out the fitness value (F1 score) using
  Feature set using CatBoost classifier
**end**
**while** ($t_{counter}$< Max_Iter number **do**)
**for** each whale **do**
a new parameter has been updated
**if** (p< 0.5) **then**
**if** (|A|< 1) **then**
update the current position of the whale
if (|A > 1) then
**else**
**if** (|A| ≥ 1) **then**
select the random position $X_{rand}$
using the mechanism, adjust the whale's
**end**
**end**
**else**
**if** (p≥0.5) **then**
update the whale's position towards the global
**end**
**end**
**end**
**for** each while **do**
Decode feature set from whale position
Calculate the fitness value using CatBoost classifier
**end**
update X* if a set of the best solutions exists
t = t + 1
**end**
Save the best feature set

*6) Harris hawk optimization algorithm (HHO)*: Haidari and his colleagues (2019) [17] proposed the use of a new metaheuristic algorithm known as the Harris Hawk Optimization (HHO). HHO imitates the notions of Harris hawks in order to investigate the diverse prey, surprise pounces, and attack techniques used by Harris hawks in the natural world. In HHO, the candidate solutions are symbolized by hawks, and the best solution, which is also referred to as the nearly optimum solution, is referred to as prey. The Harris hawks make use of their keen vision to locate their prey and

then execute a surprise attack in order to successfully capture the target they have located [18].

In most cases, HHO is modeled into two distinct phases: the exploitation phase, and the exploration phase. The HHO algorithm may be used for either exploration or exploitation, and after it has been used for either purpose, the exploration behavior can be altered dependent on the amount of energy that the prey is able to escape with. It is possible to mathematically determine the escape energy of prey using the Eq. (2) to (3):

$$E = 2E_0(t - \frac{t}{T}) \quad (2)$$

$$E_0 = 2r - 1 \quad (3)$$

where t represents the current iteration, T represents the maximum number of iterations, E0 represents the initial energy that is created at random in the range [1,1], and r represents a random value that falls in the range [0, 1]. In Algorithm 2, the pseudo-code for feature selection using HHO is presented.

**Algorithm 02: Feature Selection Using HHO algorithm**

**Inputs:** N is the population size, while T is the maximum

**Outputs:** Rabbit's position and fitness value initialize

**While** (stopping conditions is not met) **do**

**for** (each hawk ($X_i$)) **do**

Update the initial energy E0 and jump strength j

**If** (|E|≥) **then**

Update the location vector

**If** (|E|<1) **then**

**if** (r≥0.5 and |E|≥0.5) **then**

Update the location vector

**else if** (r≥ 0.5 and |E|<0.5) **then**

Update the location vector

**else if** (r <0.5 and |E|≥0.5) **then**

Update the location vector

**else if** (r <0.5 and |E|<0.5) **then**

Update the location vector

**Return** $X_{rabbit}$

Our feature subset was optimized using the WOA and HHO to minimize the number of features while also increasing prediction accuracy. The feature subsets are selected from the WOA and HHO solution sets. The solution set's value indicates whether or not to choose a feature. The CatBoost algorithm was then used to classify liver disease based on the feature subsets. F1 score true and the predicted class is used as the value of an agent's fitness. The advantage of F1 score is that it helps to provide harmonic mean, accuracy and recall. Because of this, it is harsher on values at the extremes. Overall, an agent's fitness value is referred to as the F1 score (feature subset). The WOA and HHO take the best fitness value as a baseline and update the position in accordance with the methodology used by each. This iteration is repeated until a predetermined end point is achieved.

## IV. Results and Analysis

The results of classifier algorithm are detailed in Section A on the experimental evaluation of our proposed CatBoost model. In addition, the relevance of the feature selection with two metaheuristic algorithms mentioned in Section B.

### A. Performance Analysis

In the previous section, we discussed the various contents of the dataset. We used a technique and method to classify the class samples in this dataset. In this section, we will present the research findings. Following the described procedure, we set up a classification model where the CatBoost model was used for training the model and the rest of the samples were used for testing. We have split our dataset in the ratio of 70:30. In this paper we have proposed classification algorithms like CatBoost algorithm. However, after finishing data preprocessing steps without applying feature selection techniques, algorithms are used for classification. Our model provided and accuracy of 74%. Besides accuracy, various evaluation criteria such as precision, recall, and f1-score values are compared in Table II.

TABLE II. Classification Report of our Model

| Class | Precision | Recall | F1-Score |
|---|---|---|---|
| Non-LD | 0.50 | 0.37 | 0.42 |
| LD | 0.80 | 0.87 | 0.84 |

In Fig. 6 shows a confusion matrix, which is used as a powerful tool for evaluating the performance of classification models in machine learning. This matrix clearly shows how the model classified the data into actual and predicted categories. It divides the results into four different categories, providing important insights into the model's strengths and weaknesses: true positives, true negatives, false positives, and false negatives. In this confusion matrix, the result of a binary classification task, where there are two possible outcomes - "0" and "1". The actual value, or true label, is displayed along the vertical axis and the model predicted value along the horizontal axis. The number in each cell shows how many examples fall into that particular category. In this model, it correctly predicted class "1" in 76 instances (true positives) and correctly assigned class "0" in 11 instances (true negatives). However, the model incorrectly classified class "0" as "1" (false positive) in 19 instances and class "1" as "0" (false negative) in 11 instances. These errors show where the model is having trouble, especially distinguishing between two classes. The confusion matrix gives a clear picture of the prediction performance of the model, which helps us better understand the accuracy, precision, and other evaluation metrics of the model.

Fig. 7 shows a receiver operating characteristic (ROC) curve, which is commonly used to evaluate the performance of binary classification models. This curve depicts the relationship between the true positive rate (TPR) and the false positive rate (FPR) at different threshold settings, giving an understanding of how well the model is able to distinguish between the two classes. The dashed diagonal line represents the performance of a random classifier and serves as a baseline with an AUC of 0.5. The orange curve shows the actual performance of the model, which lies above the diagonal, indicating that the model is giving

better results than the random guess. An AUC value of 0.82 suggests that the model is able to distinguish between positive and negative classes and has an 82% chance of correctly identifying a positive instance.

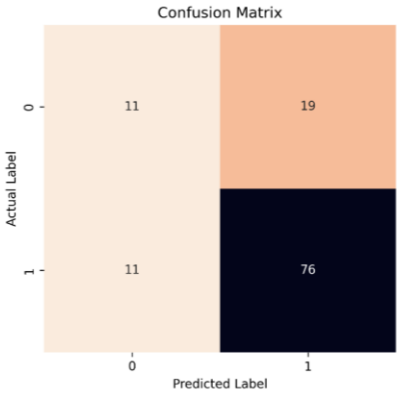


Fig. 6. Confusion matrix.

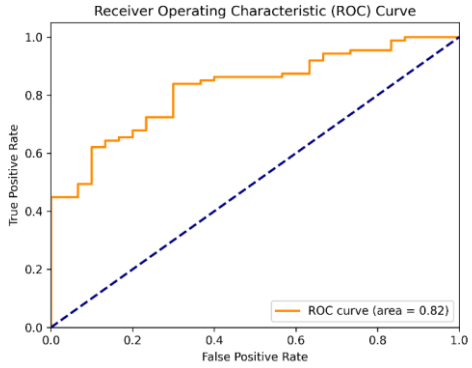


Fig. 7. ROC curves of various classes.

## B. Feature Selection Outcome

We applied the FS algorithm to increase the accuracy of the CatBoost classifier and reduce the dimensionality of the features. The FS process was carried out using two metaheuristic algorithms named WOA and HHO. A fitness function is constructed based on the performance evaluation of the CatBoost classifier. Other performance metrics, such as F1 score, precision and recall, are also taken into account. We checked the 'p_r' parameter of WOA and HHO between 0.21, indicating the learning rate potential. A value of 0.25 was identified as optimal for 'p_r', while values were 50 for 'pop_size' and 'epoch' parameters. Various 'p_r' values are shown in Table III, which highlights the set of features returned from the WOA and HHO processes.

TABLE III. BEST FEATURE SOLUTION

| Optimization Algorithm | Feature Name | Optimization Algorithm | Feature Name |
|---|---|---|---|
| HHO | Total Bilirubin | WOA | Age |
| | Direct Bilirubin | | Total_Bilirubin |
| | Alamine Aminotransferase | | Alamine Aminotransferase |

Table III outlines the best features obtained from two optimization algorithms, HHO and WOA, which have been used to identify the most important features for liver disease detection. Selected parameters for HHO include Total Bilirubin, which is important in evaluating liver function because high bilirubin levels usually indicate liver problems. It also selects Direct Bilirubin and Alanine Aminotransferase, where it is an enzyme that increases in the blood when liver cells are damaged. On the other hand, the WOA algorithm identified Age, Total Bilirubin, and Alanine Aminotransferase as important characteristics that are influential in liver disease. Both algorithms selected Total Bilirubin and Alanine Aminotransferase, indicating their high importance in the diagnosis of liver disease, and proved to be important features for accurate detection.

In Fig. 8, the confusion matrix of the HHO algorithm shows that the model correctly classified 82 cases as true negative (TN) and 96 cases as true positive (TP). However, it misclassified 17 cases as false positive (FP) and 3 cases as false negative (FN). The HHO model achieved 85% accuracy, indicating strong performance in liver disease detection. A particularly low number of false negatives makes the model useful in medical diagnosis, as it indicates that very few cases of true disease are missed.

In Fig. 9, the confusion matrix of the WOA shows that the model correctly classified 13 cases as True Negative (TN) and 84 cases as True Positive (TP). At the same time it misclassified 17 cases as false positive (FP) and 3 cases as false negative (FN). The WOA model achieved 82% accuracy. Although the number of false negatives is low, the model lags slightly behind HHO in detecting true negatives, showing slight weakness in detecting cases without disease.

On the other hand, the previously used CatBoost algorithm achieved only 74% accuracy, which is significantly lower than HHO and WOA. This proves these two optimization algorithms more effective in liver disease detection. Overall, the HHO model shows the best performance in liver disease detection, as it is able to maintain a good balance between high accuracy and true positive and true negative detection.

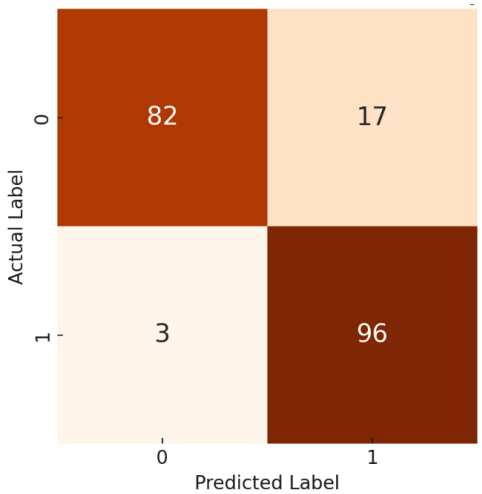


Fig. 8. Confusion matrix for HHO.

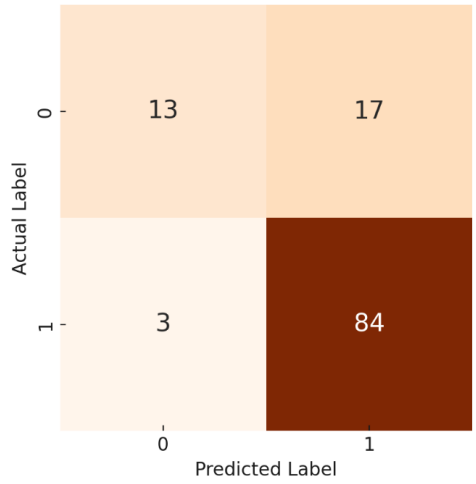


Fig. 9. Confusion matrix for WOA.

## V. Conclusion

This study successfully identified key features for liver disease detection using two optimization algorithms: Harris Hawk Optimization (HHO) and Whale Optimization Algorithm (WOA). Both algorithms highlighted Total Bilirubin and Alanine Aminotransferase as critical indicators for diagnosing liver disease. Additionally, Direct Bilirubin and Age were also recognized as significant factors in assessing liver function. Our findings align with existing research while offering new insights that can enhance diagnostic accuracy using clinical data. These results underscore the efficacy and potential of machine learning models combined with optimization algorithms in advancing liver disease diagnosis. This work contributes to the growing evidence that such computational approaches can significantly improve early detection and intervention strategies in healthcare. Future research could explore integrating additional datasets and machine learning techniques to further refine these predictive models and expand their applicability across diverse populations.

## Fund

This research Supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF- RS-2023-00237287, NRF-2021S1A5A8062526) and local government-university cooperation-based regional innovation projects (2021RIS-003).

## References

[1] S. Afrin et al., "Supervised machine learning based liver disease prediction approach with LASSO feature selection," Bull. Electr. Eng. Inform., vol. 10, no. 6, pp. 3369–3376, 2021.

[2] M. A. Kuzhippallil, C. Joseph, and A. Kannan, "Comparative analysis of machine learning techniques for indian liver disease patients," in 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), IEEE, 2020, pp. 778–782.

[3] M. F. Rabbi, S. M. Hasan, A. I. Champa, M. AsifZaman, and M. K. Hasan, "Prediction of liver disorders using machine learning algorithms: a comparative study," in 2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT), IEEE, 2020, pp. 111–116.

[4] C. Geetha and A. R. Arunachalam, "Evaluation based Approaches for Liver Disease Prediction using Machine Learning Algorithms," in 2021 International Conference on Computer Communication and Informatics (ICCCI), IEEE, 2021, pp. 1–4.

[5] S. Ambesange, A. Vijayalaxmi, R. Uppin, S. Patil, and V. Patil, "Optimizing Liver disease prediction with Random Forest by various Data balancing Techniques," in 2020 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM), IEEE, 2020, pp. 98–102.

[6] S. Rajesh, N. A. Choudhury, and S. Moulik, "Hepatocellular carcinoma (HCC) liver cancer prediction using machine learning algorithms," in 2020 IEEE 17th India Council International Conference (INDICON), IEEE, 2020, pp. 1–5.

[7] A. S. Rahman, F. J. M. Shamrat, Z. Tasnim, J. Roy, and S. A. Hossain, "A comparative study on liver disease prediction using supervised machine learning algorithms," Int. J. Sci. Technol. Res., vol. 8, no. 11, pp. 419–422, 2019.

[8] A. S. Singh, M. Irfan, and A. Chowdhury, "Prediction of liver disease using classification algorithms," in 2018 4th international conference on computing communication and automation (ICCCA), IEEE, 2018, pp. 1–3.

[9] S. Dhamodharan, "Liver disease prediction using bayesian classification," 2016.

[10] A. H. Roslina and A. Noraziah, "Prediction of hepatitis prognosis using support vector machines and wrapper method," in 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, IEEE, 2010, pp. 2209–2211.

[11] O. S. Soliman and E. A. Elhamd, "Classification of hepatitis C virus using modified particle swarm optimization and least squares support vector machine," Int. J. Sci. Eng. Res., vol. 5, no. 3, p. 122, 2014.

[12] S. Vijayarani and S. Dhayanand, "Liver disease prediction using SVM and Naïve Bayes algorithms," Int. J. Sci. Eng. Technol. Res. IJSETR, vol. 4, no. 4, pp. 816–820, 2015.

[13] "Indian Liver Patient Records." Accessed: Jun. 23, 2022. [Online]. Available: https://www.kaggle.com/uciml/indian-liver-patient-records

[14] A. V. Dorogush, V. Ershov, and A. Gulin, "CatBoost: gradient boosting with categorical features support," Oct. 24, 2018, arXiv: arXiv:1810.11363. doi: 10.48550/arXiv.1810.11363.

[15] S. Thiesen, "CatBoost regression in 6 minutes," Medium. Accessed: Jun. 27, 2022. [Online]. Available: https://towardsdatascience.com/catboost-regression-in-6-minutes-3487f3e5b329

[16] S. Mirjalili and A. Lewis, "The whale optimization algorithm," Adv. Eng. Softw., vol. 95, pp. 51–67, 2016.

[17] A. A. Heidari, S. Mirjalili, H. Faris, I. Aljarah, M. Mafarja, and H. Chen, "Harris hawks optimization: Algorithm and applications," Future Gener. Comput. Syst., vol. 97, pp. 849–872, 2019.

[18] J. Too, A. R. Abdullah, and N. Mohd Saad, "A new quadratic binary harris hawk optimization for feature selection," Electronics, vol. 8, no. 10, p. 1130, 2019.