Information Sciences Letters
An International Journal

http://dx.doi.org/10.18576/isl/120644

A Proposed Approach for Predicting Liver Disease

Ibrahim Mohamed Attiya[1,], Rania A. Abouelsoud[2], and Ahmed Salama Ismail[1]*

[1]Information System Department, Faculty of Computers and Information System, Fayoum University, Fayoum, Egypt
[2]Electrical Engineering Department, Faculty of Engineering, Fayoum University, Fayoum, Egypt

Received: 2 Apr. 2023, Revised: 13 May 2023, Accepted: 22 May 2023.
Published online: 1 Jun. 2023

Abstract: One of the main challenges is to exploit recent technologies in a way that is able to preserve human life. Liver disease is one of the most influencing and largest organs of the human body, which has a great impact on human life, according to the massive number of deaths of this disease. So, it is important to classify liver disease with the maximum possible accuracy, as the current problem is the weak accuracy of classifying liver disease and not predicting the severity of the liver disease. Thus, through this paper, the aim behind our proposed work is to enhance the performance of classifying liver disease, predicting the severity of liver disease, and then Recommending suitable behaviors to patients using machine learning algorithms and tools like a GridsearchCV tool. Indian liver patient's dataset (ILPD) and the hepatitis C virus (HCV) dataset are our training datasets. Hence, the proposed solution enhanced the classification accuracy of liver disease by 80% and 77 % for extra tree and KNN algorithms when using ILPD datasets. And when using the HCV dataset, the accuracy is achieved by the Gradient boosting algorithm and Logistic Regression by 96% for classifying liver disease, disease severity.

Keywords: Prediction Disease, Machine learning algorithms, Liver disease, ILPD, HCV, Recommendation system, Classification Techniques.

1 Introduction

Due to the incredible improvement of new technologies, especially Artificial Intelligence, and the huge spread of diseases, also the increasing number of deaths through diseases for many reasons, including the inaccuracy of correct diagnosis of the patient and the lack of early detection in some chronic diseases causing that increasing the number of deaths.

So, Health care systems play an essential part in human life in helping to preserve human life, especially in predicting diseases, that is, through early detection of the disease by predicting the disease faster and more accurately and helping doctors to make more suitable decisions for the patients because the early and more accurately the patient is diagnosed, the faster the ability patient is treated.

Chronic Liver Disease, which is among the main causes of death globally and affects a sizable population. A confluence of particular chemicals that harm the liver is the root cause of this disease. Liver disease is the most important organ that is involved in a variety of processes, like the breakdown of red blood cells.

Any abnormality of the liver is referred to as liver disease. This case could manifest in different ways, including inflammation (hepatitis B and C) brought on by viral, non-infectious, or autoimmune causes, malignant tumors, liver scarring (cirrhosis), and metabolic disorders. Therefore, it has become crucial to use machine learning to anticipate liver illness.

Since it has made significant strides, machine learning (ML) is currently one of the best options for handling problems, especially in a wide range of industries and applications, including speech recognition, data analytics, classification, natural language processing, health care, and data analytics.

Machine learning models are data-dependent, nonlinear, and nonparametric models that are constructed based on the assumptions of the underlying data-generating process. Using historical data, machine learning techniques can identify relationships between input and output variables. The three main subcategories of machine learning are supervised learning, unsupervised learning, and reinforcement learning. In supervised or predictive learning, the target is to learn a mapping from inputs xi to output yi, given a labeled set of input-output pairs M ={(Xi, Yi)}i=1N. M . Here is the training set, and N is the number of training examples. Supervised learning is divided into classification and regression. In the unsupervised or descriptive learning approach, we are only given inputs, M={(Xi)}i=1N. Sometimes called knowledge discovery. Reinforcement learning is used for learning how to behave or act, given some reward or

[*]Corresponding author e-mail: im1959@fayoum.edu.eg

punishment signals occasionally.

This paper focuses only on supervised learning. Two kinds of problems that supervised machine learning aim to solve are classification problems and regression problems. The classification methods are used to categorize certain data of statistical communities split to different groups based on one or more of the basic properties of these data. The nature of data restricts it from choosing the best classification method.

Some researchers have been interested in predicting liver disease, but their applications still need improvement due to the poor accuracy of the results.

Through the proposed method, a set of tools that help machine learning algorithms to adapt the data sets are used, whatever they are, by determining the optimal parameters for the algorithm used, which helps to obtain an increase in the accuracy of the results for predicting the disease.

The other advantage is predicting the severity of the disease in the patient, which contributes to the survival of the human element through the speed of disease detection. Also, the preprocessing of data sets seriously affects the accuracy of the prediction results. Therefore, in this paper, we consider this problem in preprocessing the data from stray values and null values before the prediction process. The researchers used a set of machine learning algorithms such as decision tree (DT), Gradient boosting (GB), Support vector machine (SVM), K-Nearest Neighbors (KNN), packing classifier, additive tree, and multilayer sensor (MLP) for the prediction process., which we will discuss briefly.

1) Decision Tree (DT): is a learning algorithm that is used in classification and regression. It is among the supervised algorithms and is widely used in classification processes. This algorithm is represented in the form of a tree. This tree consists of a group of nodes, branches, and leaf nodes, as the internal nodes in this tree represent the features of the data set, and represent the branches are the rules of decisions, and the result is in the form of paper nodes. The decision tree is usually used because it resembles the human way of finding solutions to problems [1].

2) Gradient Boosting (GB): This machine learning method is employed, among other things, for classification and regression tasks. An ensemble of weak predictive models, typically decision trees, functions as a predictive model. The resulting approach is known as a Gradient boosted tree, and it typically outperforms random forests when a decision tree is a weak learner. [2].

3) Support Vector Machine (SVM): SVM is a supervised type of machine learning model that uses classification algorithms to solve binary classification problems. Integrating support vector machines with a set of connected supervised learning methods for regression and classification. SVM is an advanced technique used in mathematical learning theory with a complete classification algorithm. These modeling techniques can be applied to both linear and nonlinear data classification [3].

4) K-Nearest Neighbors (KNN): is a supervised classification algorithm method. It classifies objects based on their nearest neighbors. This is a type of instance-based learning. The distance of a computed property to its neighbors is measured using Euclidean distance. It takes a set of named points and uses them to label another point. Data are grouped according to their similarity, and missing data values can be filled with K-NN. Once missing values are filled, various predictive techniques are applied to the dataset [1].

5) Bagging Classifier: this is a machine-learning technique based on an ensemble of models developed using multiple training datasets drawn from the original training dataset. It computes multiple models and averages them to produce the final ensemble model. Traditional bagging methods create multiple copies of the training set by randomly selecting molecules with substitutions from the training set [4].

6) Extra Tree (ET): Extra trees or extremely randomized trees are another ensemble machine learning classifier like RF. But there are two fundamental differences between ET and RF. One is ET samples without replacement, and the other is instead of picking the best feature, random features are chosen for splitting the tree nodes. After creating multiple unpruned trees, ET predicts by averaging all the tree outcomes in case of regression or by calculating majority votes in case of classification [5].

7) Random forest (RF): is an example for supervised learning algorithm that is used in classification and regression. It consists of a group of trees, where these trees represent decisions. This is at the time of training to classify the data set. The resulting category is chosen based on the decisions of most trees. Random forests are characterized by their accuracy in classifying the given data, but in some works accuracy decreases depending on the characteristics of the data.

8) Naive Bayes (NB): In statistics, Naive Bayesian classifiers are a class of simple "probabilistic classifiers" based on the application of Bayes' theorem and strong (naive) assumptions of independence between features, but combined with kernel density estimation, they can achieve high accuracy. Naive Bayes classifiers are highly scalable and require a set

of parameters that is linear in the number of variables (features/predictors) in the learning problem.

9) AdaBoost: short for Adaptive Boosting, is a statistical classification meta-algorithm formulated. Any learning algorithm tends to be better suited to certain problem types than others and often requires tuning many different parameters and configurations to achieve optimal performance on a dataset. AdaBoost (using decision trees as weak learners) is often considered the best classifier out of the box.

Hence, the goal of this investigation is to enhance the performance of the classifying of liver disease, classify the danger of liver disease, and recommending suitable behaviors.

In this paper, we used the Indian liver disease dataset (ILPD) to improve the classification of liver disease and compared our results with previous work using the same dataset.

We used an (HCV) data set to classifiy the severity of liver disease and also improve the classification of liver disease. The Datasets were collected by the University of California, Irvine (UCI). The dataset was organized by the University of California, Irvine (UCI).

We found that we have improved the classification liver disease accuracy of DT up to 74%, KNN 77%, MLP 77%, GB 74% for ILPD Dataset and Gradient Boosting 96%, Logistic Regression 96%, MLP 95% and SVC 94% for HCV dataset.

The structure of this research is organized as the following, where the scientific research which is related to our proposed contribution is presented in section 2. Also, the proposed methodology is illustrated in section 3, while our results are discussed in section 4. Finally, the Conclusion and future works are summarized in section 5.

2 Literature Review

In this section, a set of different machine learning algorithms are discussed with other datasets and present their classifiers' accuracy.

Liver Diseases Prediction is developed by a set of Machine Learning Approaches [11]. In this paper, the researcher used a set of algorithms to predict liver disease by classifying a data set (ILPD) such as Decision Tree, Perceptron, Random Forest, K-Nearest Neighbor, Support vector machine, with feature selection and without feature selection, and it indicates that the best performance was the algorithm KNN with an accuracy of 74% with selecting the features, and the accuracy of the performance of the algorithms without selecting the features was as follows: Decision Tree 60%, Perceptron 39%, Random Forest 64%, K-Nearest Neighbor 66%, Support vector machine 71%, And with feature selection was as follows Decision Tree 72%, Perceptron 66%, Random Forest 73%, K-Nearest Neighbor 74%, Support vector machine 72% After studying this research, it is clear from the results it reached that it still needs improvement.

Software-based prediction of liver disease with feature selection and classification techniques [12]. In this paper, the researcher classified a set ILPD patients to predict liver disease through a set of algorithms with feature selection techniques as follows: Logistic Regression74.36%, Naive Bayes 55.9%, SMO 71%, IBK 67.41, J48 70.67% Random Foreast71.87% The performance of algorithms without feature selection technique was accurate Logistic Regression 72.50%, Naive Bayes 55.74%, SMO 71.35%, IBK 67.15%, J48 68.78% Random Forest 71.53. The best classification accuracy was the Logistic Regression algorithm, with an accuracy of 74.36% with the Feature selection technique. The weaknesses of this research are the poor accuracy of the results.

A Comparative Analysis of Classification Algorithms in Liver Disease Detection [13]. In this paper, the researcher used a set of algorithms such as Logistic Regression, Random Forest (RF), KNN, and Decision tree. To identify liver disease in a data set of (ILPD) patients, the performance of machine learning algorithms was as follows: Random forest 65.00%, Logistic Regression 70.15, Decision Tree 63.46, K-Nearest Neighbor 72.04, It was found that the KNN algorithm achieves the best accuracy with 72.04%. In this research, he performed his experiment on a few machine learning algorithms and did not predict the severity of liver disease.

A Fact-Based Liver Disease Prediction by Enforcing Machine Learning Algorithms [14]. In this study, the researcher used twelve classification algorithms represented by: Multilayer perceptron, KNN, Logistic regression, Decision tree, Random forest tree, Gradient boosting, Support vector machine, Naive Bayes, AdaBoost, XGBoost, Bayesian, Bagging, and the accuracy of the performance of each algorithm was as follows: Multilayer perceptron 72.50%, KNN 74.20 %, Logistic regression 74.90 %, Decision tree 65.70 %, Random forest tree 74.60%, Gradient boosting 69.40%, Support vector machine 85.70%, Naive Bayes 62.00%, AdaBoost 68.70%, XGBoost 70.30%, Bayesian 71.40%, Bagging 75.30%. In this research, the results it has been reached still need improvement and did not predict the severity of liver disease.

Hepatitis C virus (HCV) prediction by machine learning techniques. [15] In this paper, the researcher used an Egyptian patient's dataset to predict Hepatitis C Virus disease, and the best accuracy was 51.06% KNN. This researcher used another data set (HCV) to predict liver disease but also reached fragile, unsatisfactory results.

Prediction of Liver Malady Using Advanced Classification Algorithms [3]. In this paper, the author used SVM to predict liver disease using the ILPD dataset and obtained an accuracy of 78%. But in this research, the researcher did not perform his experiment on a set of algorithms, but only one algorithm, and also did not predict the severity of liver disease.

Implementation of partitional clustering on ILPD dataset to predict liver disorders [16]. In this paper, the researcher used a set of algorithms to predict liver disease through the data set (ILPD), (NDS), where he used k-NN, C 4.5 to classify (ILPD) Patients, and the accuracy was obtained k-NN 0.64 %, C 4.5 0.69% but the results it reached still need improvement.

Firefly Algorithm for Functional Link Neural Network Learning [17]. In this paper, the researcher used more than datasets to predict diseases, he used a dataset (ILPD), and the classification results were as follows MLP-BP 70.61%, FLNN-BP 69.63%, FLNN-FA, and 70.73% did not predict the severity of liver disease.

After studying the previous work on predicting liver diseases and other outcomes, it was found that some researchers still need to improve the accuracy of their results. Also, the research did not anticipate the severity of liver disease, and this is what we will do in this paper to improve the classification of liver disease and the severity of liver disease as shown in Table 1.

Table 1: Summary of Literature Review

Method	Dataset	Accuracy	Reference
DT RF SVM KNN	ILPD	60 % 64 % 71 % 66 %	[11]
LR Naive Bayes SVM KNN DT RF	ILPD	72.50 % 55.74 % 71.35 % 64.15 % 68.78 % 71.53 %	[12]
MLP KNN LR DT RF GB SVM Naive Bayes AdaBoost XGBoost Bayesian Bagging	ILPD	72.50 % 74.20 % 74.90 % 65.70 % 74.60 % 69.40 % 85.70 % 62.00 % 68.70 % 70.30 % 71.40 % 75.30 %	[14]
KNN SVM RF Bagging Adaboost	HCV	47.35 % 52.64 % 49.15 % 46.63 % 50 %	[15]
SVM	ILPD	78%	[3]
KNN DT	ILPD	64 % 69 %	[16]
MLP-BP FLNN-BP FLNN-FA	ILPD	70.61 % 69.63 % 70.73 %	[17]
KNN RF Naïve Bayes DT LR	HCV	0.2548 % 0.2512 % 0.2476 % 0.2433 % 0.2433 %	[18]

3 Proposed Method

The proposed model is developed through the Jupyter environment, where the Python language is used to conduct our experiment.

JupyterLab is a web-based interactive development tool for code and data. Through it, users can configure and arrange different aspects of data science, such as scientific computing, computational journalism, and machine learning.

Our proposed methodology aims to enhance the accuracy of the classification of liver disease, classify the severity of the disease, and Recommending suitable behaviours to patients, as shown in Figure 1...

Firstly, the data used in the methodology is collected. Secondly, the dataset is preprocessed, where StandardScaler processes stray values, and empty values are processed by using the Fillna method, which process this problem by calculating the average of data features.

Thirdly, the data sets used are divided into a training part and a test part with a ratio of 80:20, respectively. To avoid choosing similar values during the learning and testing phases of the model, a 10-fold cross-validation operator was used. That allows the data to be clustered into k equal subsets and using each subgroup to be part of both training and testing activities.

The work of the cross-validation operator is considered efficient because it repeats the learning phase k times, each time with a different selection of test data than the previous one. It repeats the experiment k times and uses the averaged result. Cross-validation is an operator widely used for learning and testing purposes.

Fourthly, liver disease and severity of liver disease are classified by using a set of machine learning algorithms. To reach the best parameters for these algorithms that help improve and enhance disease classification, the GridSearchCV tool is used.

Finally, our proposed solution to Recommending suitable behaviours to patients.

The aim behind the proposed method is to improve the classification of liver disease lies in determining the optimal parameters for the machine learning algorithms, and this is done by using a GridSearchCV tool that has a very great benefit which is access to the best parameters of the algorithm used, which improves the performance of the algorithm, as well as pre-processing of data sets by processing stray values Which affects the accuracy of the results and also the treatment of empty values in the data sets, which also causes poor results.

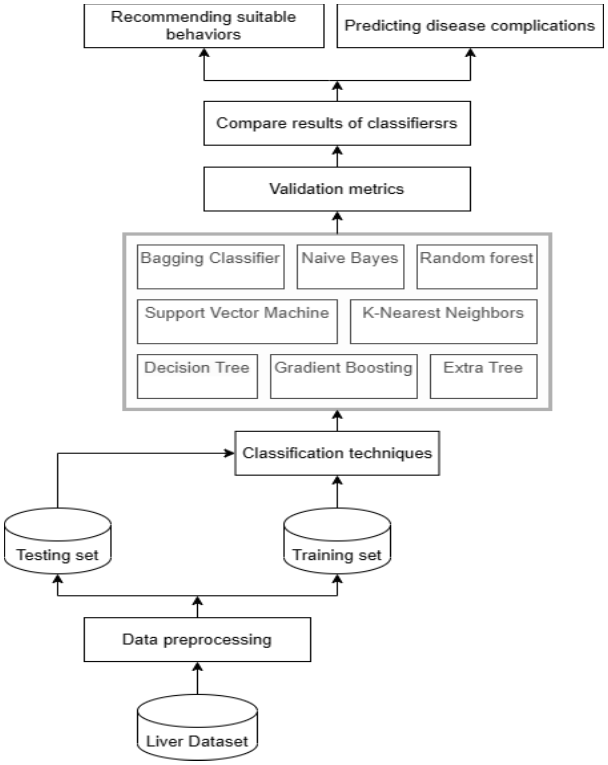

Fig. 1: Our Proposed Architecture

The proposed methodology is covered in detail in the following steps:

Dataset Description

In this research, we used two data sets related to liver disease.

3.1 ILPD Dataset

The researchers conducted their experiments using the Indian Liver Patient Dataset (ILPD). The dataset was collected by the University of California, Irvine (UCI). The dataset contains 416 hepatic medical records and 167 non-liver medical records. This dataset was collected from a test sample in northeastern Andhra Pradesh, India. "Dataset" is the class name used for classification (Patients with Liver Disease or Non-Patients with Liver Disease). The dataset contains 441 male patient records and 142 female patient records. ILPD dataset consists of a set of elements which are described as shown in Table 2. And the correlation of various attributes in the HCV dataset is shown in Figure 2.

Table 2: Attribute Information ILPD Dataset

n	Attributes	description
1	Age	Age of the patient
2	Gender	Gender (Male or Female) of the patients.
3	TB	Total Bilirubin: This blood test calculates how much bilirubin is present. It serves as a measure of the liver's effectiveness.
4	DB	Conjugated or direct bilirubin moves freely through your blood to your liver. Most of the bilirubin ends up in the small intestine. A very small amount of bilirubin
5	Alkphos	The Alkaline Phosphatase test quantifies the blood's level of ALP. It is frequently used to identify bone disease or liver damage.
6	Sgpt	ALT stands for alanine aminotransferase, which refers to the enzyme found in the liver.
7	Sgot	Aspartate aminotransferase is an enzyme that is mostly located in the liver but is also present in muscles and other body organs.
8	TP	Total protein testing is usually done as part of a regular checkup. It measures the levels of two proteins in your body, albumin, and globulin.
9	ALB	Albumin: is a blood plasma protein that is synthesized in the liver.
10	A/G	ratio of albumin to globulin The total amount of protein is measured by total protein and albumin/globulin (A/G) ratio assays. Albumin and globulin are the two primary proteins found in blood.
11	Dataset	Dataset target class; data is split into two sets: 1. Patient with liver disease. 2. Patient with no disease.

3.2 HCV Dataset

The collection consists of demographic information including age, laboratory results from hepatitis C patients and blood donors. The UCI Repository provided the data. The only two qualities that lack numbers are category and gender. ALB, ALP, ALT, AST, BIL, CHE, CHOL, CREA, GGT, and PROT are the qualities that correspond to patient data from 1 to 4 and laboratory data from 5 to 14 respectively.

HCV dataset consists of a set of elements which are described as shown in Table 3.

Table 3: Attribute Information HCV Dataset

n	Attributes	Description
1	X	(Patient ID/No)
2	Category (diagnosis)	(values: '0= Patient with no disease, '1=Hepatitis', '2=Fibrosis', '3=Cirrhosis').
3	Age	Age of the patient (in years).
4	Sex	Gender (Male or Female) of the patients.
5	ALB	Albumin: a blood plasma protein synthesized in the liver.
6	ALP	The alkaline phosphatase (ALP) test quantifies the blood's level of ALP. It is frequently used to identify bone or liver illness.
7	ALT	ALT stands for alanine aminotransferase: It is an enzyme found in the liver.
8	AST	The enzyme known as AST (aspartate aminotransferase) is mostly present in the

n	Attributes	Description
		liver.
9	BIL	Total Bilirubin: This blood test calculates how much bilirubin is present. How well your liver functions will be determined by this test.
10	CHE	A blood test called serum cholinesterase measures the amounts of two compounds that support the healthy operation of the nervous system. Acetylcholinesterase and pseudocholinesterase are their names.
11	CHOL	“Fat-like substance found in all of body’s cells" is cholesterol.
12	CREA	A creatinine test measures how well your kidneys filter waste products from your blood. Creatinine is a compound left over from the process of producing energy in muscles.
13	GGT	A gamma-glutamyltransferase: test measures the amount of A gamma-glutamyltransferase in the blood.
14	PROT	Muscle, bone, skin, hair, and practically every other biological part or tissue can be found to have protein. Haemoglobin, which transports oxygen in the blood, and enzymes that power numerous chemical reactions are both produced by it.

The target attribute for classification is Category: Patient with no disease vs. Hepatitis C patients (including its progress ('just' Hepatitis C, Fibrosis, and Cirrhosis)

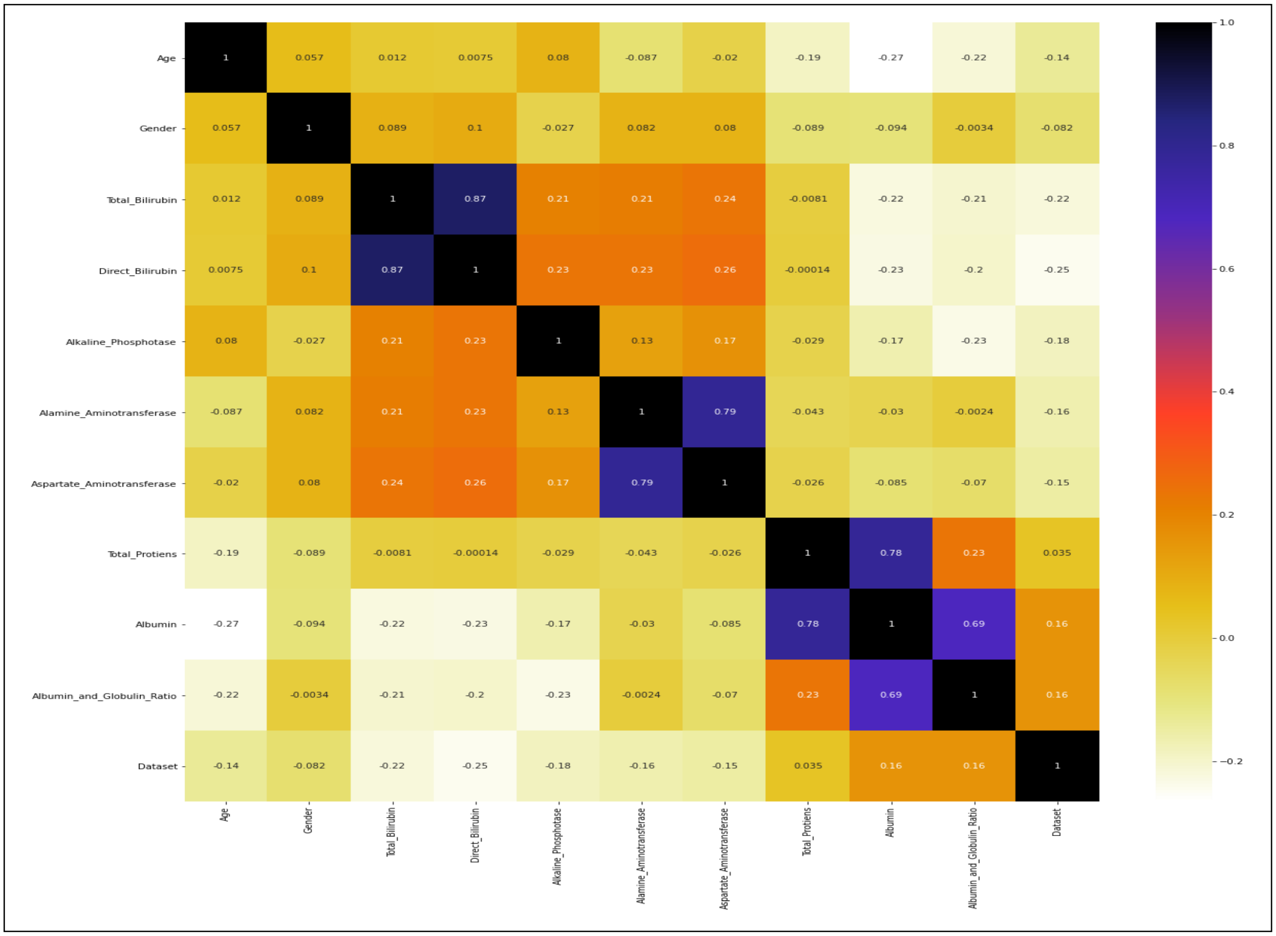

Fig. 2: Correlation of Various Attributes in the ILPD Dataset

3.3 Data Preprocessing

The excellent quality of the data and how clean the data is from outliers, missing values, and non-relevant features, as the superb quality of classification and prediction of disease. That is because the more these outliers increase, the lower the overall prediction accuracy.

Our research objective is to enhance the prediction accuracy for prediction liver disease, predict the dangers of liver disease and recommend to patients the best action or instructions. So, in our research, we focused on increasing the prediction accuracy for liver disease.

We used the label encoding in this dataset, where attribute gender contains categorical values, which are Male and Female. Label encoder is used to convert the absolute values into numerical labels, such as 1 for Males and 0 for Females, as we processed the missing values.

Also, we processed outlier values to improve the quality of the data. We also used a (GridSearchCV) tool to determine the best parameters for algorithms to increase accuracy.

After preprocessing the data sets, they data sets were divided into a training part and a testing part, where the ratio was 80:20, respectively.

3.4 Classification process

This step shows the machine learning algorithms that we used in the research, as we used supervised machine learning algorithms such as Decision Tree (DT), Gradient Boosting (GB), Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Bagging Classifier, Extra Tree (ET), and MLP.

One of the problems facing the researchers is the poor results they reached in classifying liver disease due to the inability to achieve the optimal parameters of the machine learning algorithms used in their experiments and research.

In order to process this problem, we used the GridSearchCV tool to determine the optimal parameters for the used machine learning algorithms in order to reach the best performance for each algorithm, which contributes to reaching the goals of the methodology. Which achieves and enhances the accuracy of liver disease classification.

After training the machine learning algorithm on the dataset, model save is used. It is the process of saving the model after it has been trained as if it had memorized its weights, so it can be used and predicted later without wasting time again in training. It is used via the **externals. joblib** module from the Scikit-learn library.

Scikit-learn (Sklearn) is the most valuable and powerful machine-learning library in Python. It provides a range of efficient machine learning and statistical modeling tools, including classification, regression, clustering, and dimensionality reduction, through a consistent interface in Python. This library is primarily written in Python and built based on NumPy, SciPy, and Matplotlib.

Support Vector Machines

Through this algorithm, each data item is designed as a point in multi-dimensional space (where n is the number of features) by considering that the value of each feature refers to the value of a particular coordinate. Then, the classification is done. The pseudo-code of the SVM classifier is represented in the following Fig 3:

```
    • Input: S, λ, T
      where S= data set samples,
      λ= the regularization parameter of SVM for a linear kernel,
      T= number of iteration.
    • Initialize: Set W1 = 0          where W = weight
    • For t= 1, 2………...., T
    • Choose $i_t$ € {1... $|S|$} uniformly at random.    Where I = index
    • Set $\eta_t = \frac{1}{\lambda t}$
    ▪ If $y_{i_t}$ ( $w_t$ , $w_{i_t}$ ) < 1, then :
          • Set $w_{t+1}$= (1-ηtλ)$w_t$+ ηt $y_{i_t}$ $x_{i_t}$
    ▪ Else ( if $y_{i_t}$ ( $w_t$ , $w_{i_t}$ ) >= 1 ):
          • Set $w_{t+1}$= (1-ηtλ)$w_t$
 [Optional: $w_{t+1}$= min {1,$\frac{1/\sqrt{\lambda}}{||w_{t+1}||}$} $w_{t+1}$ ]
          • Output: $w_{T+1}$
```

Fig. 3: Pseudo Code of SVM Classifier

Random forest

It provides as an example of how regression and classification problems may both be solved using supervised machine learning [34,35]. The trees grow parallel to one another in the random forests. There is no interaction between the trees as they are being created. It works by creating a sizable number of levels from the decision tree that are inherited using the training data, and then extracting the category that is the mode of the types (classification) or means prediction (regression), which combines the results of multiple predictions, that aggregates several decision trees, with some helpful modifications: the amount of features that will be split at each node is Forbidden to some share of the entire (which is though).This makes sure the ensemble model uses all the most likely prophetic options while without placing an excessive amount of weight on any one person's attributes. In order to add more randomization and avoid overfitting, each tree generates a random sample from the starting information set.

$$g(x) = f_0(x) + f_1(x) + f_2(x) + \cdots \quad (1)$$

g is the sum of sample base models f_i

- Randomly select “k” features from total “m” features.

  Where k < m
- Among the “k” features, calculate the node “d” using the best split point.
- Split the node into daughter nodes using the best split.
- Repeat 1 to 3 steps until “1” number of nodes has been reached.
- Build forest by repeating steps 1 to 4 for “n” number times to create “n” number of trees.

Fig. 4: Pseudo Code of Random Forest Classifier

3.5 prediction process

In this step, the complications of the disease are predicted

Logistic Regression

A mathematical analytical technique called logistic regression may be used to forecast an information value supported by prior data set observations. Within the field of machine learning, logistic regression is becoming a crucial tool. The method enables the use of an algorithmic programme in a machine learning application to categorise incoming data based on prior data. The computer programme should get better at guessing classes inside data sets as more pertinent data is added. When using the (ETL) method to stage the data for analysis, Logistic regression can be utilised to enable data sets to be arranged into precisely predefined blocks through the data preparation phase.The model is represented by the following Eq:

$$p(x) = e^{b0+b1x}/(1 + e^{b0+b1x}) \quad (2)$$

It can be transformed into: -

$$ln\left(\frac{p(x)}{1-p(x)}\right) = b0 + b1x \quad (3)$$

Where p(x) represents the predicted value, where the value of b0 is the intercept term, and the importance of b1 is the coefficient for the single input value (x). The aim of using the training data is to get the values of both coefficients b0 and b1 to shrink the error gap between the predicted data and the actual data.

3.6 Recommending suitable behaviors

After classifying liver patients and finding out the machine learning algorithm that achieves the best performance in the classification stage, a recommendation form is built to recommend the appropriate behavior to the patient according to the condition of his disease.

4 Validation and Evaluation of the Proposed Method

The process of validating and assessing a data classification model is one of the most crucial parts of creating a model. Estimating the level of performance that might be anticipated from models produced by the modelling process is the goal of validation. The classification of as many future units as possible is the primary goal of developing the classification rule. A confusion matrix is the simplest and most popular criterion for judging a set of classification rules out of the many that are available [34]. Where N is the total number of target values (classes), the confusion matrix is N×N. The information in the matrix is frequently used to assess the effectiveness of such models. The 2×2 is shown in the following table 4:

Table 4: Confusion Matrix

Actual	**predicated**	
	Positive	**Negative**
Positive	True Positive (TP)	Negative
Negative	False Negative (FN)	True Negative (TN)

As can be seen from Table 4, True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are four different possibilities for a case classification prediction with two Class classes. The results are "1" ("Yes") and "0" ("No"). A false positive result occurs when a result is misclassified as "yes" (or "positive") when in fact, it is "no" (or "negative"). A false negative result is when a result is classified as negative when it is actually positive. True positives and true negatives are obviously the correct classifications. The following formulas are used to calculate accuracy, sensitivity (recall), precision, and F1-score.

On the other hand, The Accuracy factor refers to the proportion of true results for both true the true positives and negatives values in the population as obtained in the following Eq. (4):

$$\text{Accuracy} = \frac{TN+TP}{TN+TP+FN_{+}FP} \tag{4}$$

Precision refers to the ratio of the predicted positive observation values to the total predicted positive observations in the correct form.[35] Where the high accuracy means a low wrong positive rate, which can be obtained by the following Eq. (5):

$$\text{Precision} = \frac{TP}{FP+TP} \tag{5}$$

Sensitivity or recall (the true positive value) is the percentage of positive patient's cases that are indicated through the test. In other words, sensitivity measures how effective a test is for people who are positive. With a sensitivity of 1, the test works well for positive people, and with a sensitivity of 0.5, this is equivalent to a random draw. If it's below 0.5, the test is counterproductive, and it makes sense to invert the rules so that the sensitivity is above 0.5 (assuming this doesn't affect specificity). The mathematical definition is as follows Eq. (6):

$$\text{Recall} = \frac{TP}{FN+TP} \tag{6}$$

The value of the F1 score refers to the weighted average of precision and recall. Therefore, the score the value of false positive and false negative results. Intuitively, F1 is often more useful than classification accuracy, especially when the classes are not evenly distributed. When the value of both false positives and false negatives differs significantly, it is best to consider both precision and recall can be obtained by the following Eq (7).

$$\text{F1-Score} = \frac{2TP}{FN+FP+2TP} \tag{7}$$

5 Results and Discussion

Table 5: Confusion Matrix for Applied classification Model for ILPD Dataset

Classification Algorithms	**Terms**			
	TP	**TN**	**FP**	**FN**
DT	85	1	2	29
Gradient Boosting	79	8	8	22
SVC	87	1	0	29
K-Nearest Neighbors	80	10	7	20
Bagging Classifier	81	8	6	22
MLP	79	11	8	19

Classification Algorithms	**Terms**			
	TP	**TN**	**FP**	**FN**
Extra tree	83	11	4	19
Random Forest	75	13	12	17

Through this section, the results of our experiment on a set of algorithms to classifiy data set of ILPD patients and HCV is presented. It was found that KNN, MLP, Gradient Boosting, and Extra Tree Classifiers achieve the best results, and the results are as follows in Table 6:

Table 6: Consequences of Prediction Algorithms for ILPD Dataset

Algorithms	**Accuracy**	**Precision**	**Recall**	**F1 Score**
DT	74%	0.64	0.74	0.64
Gradient Boosting	74%	0.70	0.74	0.70
SVC	75%	0.81	0.75	0.65
K-Nearest Neighbors	77%	0.75	0.77	0.75
Bagging Classifier	76%	0.73	0.76	0.73
MLP	77%	0.79	0.80	0.78
Extra tree	80%	0.75	0.77	0.75
Random Forest	75%	0.74	0.75	0.74

Table 6 shows the results that we reached by using the proposed method using the same dataset (ILPD) used in previous works for better comparison. We find that the accuracy of KNN has been improved up to 77%.

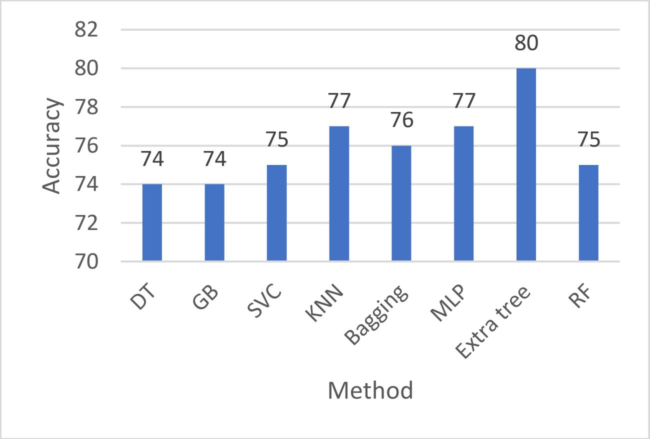

Fig. 5: Comparing Classifiers Results for ILPD Dataset

Table 7 present our results for classifying an HCV data set

Table 7: Consequences of Classification Algorithms for HCV Dataset

Algorithms	**Accuracy**	**Precision**	**Recall**	**F1 Score**
DT	93%	0.91	0.93	0.92
Gradient Boosting	96%	0.96	0.96	0.96
SVC	94%	0.94	0.94	0.94
Random Forest	93%	0.93	0.93	0.92
MLP	95%	0.95	0.95	0.95
Naïve Bayes	89%	0.92	0.91	0.91
K-Nearest Neighbors	93%	0.93	0.93	0.91
Bagging Classifier	93%	0.93	0.93	0.91
AdaBoost	92%	0.92	0.93	0.92
Logistic Regression	96%	0.96	0.96	0.96
ExtraTree	93%	0.91	0.93	0.92

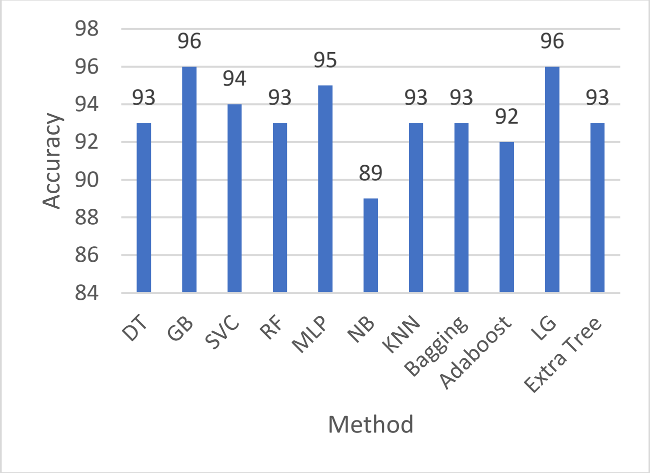

Fig. 6: Comparing Classifiers Results for HCV Dataset

To justify our work, the following Table 8 shows and compares our work and confirms that we have obtained better accuracy than the previous works mentioned:

Table 8: Comparison of Accuracy with Previous Studies for ILPD Dataset

Previous work			**Our implementation**	
Methods	**Accuracy**	**Reference**	**Methods**	**Accuracy**
DT	60 %	[11]	DT	74%
KNN	66 %		KNN	77%
SVM	71 %		SVM	75%
SMO	71.3%	[12]	SVM	75%
J48	68.7%		DT	74%
MLP	72.50%	[14]	MLP	77%
KNN	74.20%		KNN	77%
DT	65.70%		DT	74%
GB	69.40%		GB	74%
SVM	85.70%		SVM	75%
Bagging	75.30%		Bagging	76%

Table 8 presents the results of the algorithms in our experiment in order to compare them with the algorithms used in the previous works in order to improve performance in the algorithms used and to choose the algorithm that achieves better accuracy. We found that we have improved the accuracy of DT 74%, KNN 77%, MLP 77%, and GB 74%.

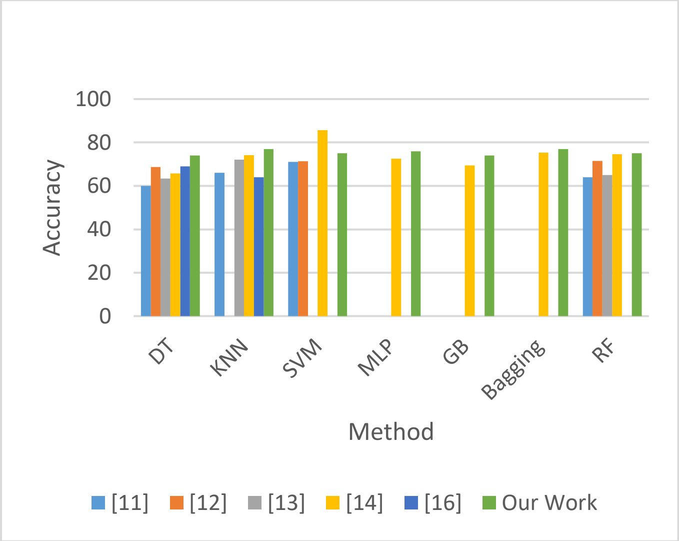

Fig. 7: Comparing Classifiers Results with Previous Work for the ILPD Dataset

6 Conclusion and Future Works

Liver disease is one of the crucial diseases that cause death for many people. Many researchers are interested in predicting liver disease, but their implementations still need improvement. According to the low level of results accuracy, they reached. Through this research, a proposed methodology is represented to improve the classification of liver disease and contribute to the classification of the severity of liver disease. In order to improve the accuracy of disease prediction, we used a GridSearchCV tool that helps to reach the best parameters for the algorithm used, which improves the accuracy of performance. To conduct our experiment, we used the Indian liver patient dataset (ILPD) that was used in previous work. The accuracy of prediction was enhanced compared to previous work, as shown in Table 8. To predict the severity of liver disease, we used another data set (HCV) with the same methodology used to show the results, as shown in Table 7.

We plan for future work of this research to use more than one large data set to achieve the best and highest accuracy, and we also plan to use deep learning techniques to solve disease prediction problems.

Conflict of interest: The authors declare that there is no conflict regarding the publication of this paper.

References

[1] D . Shah, S. Patel, S. K. Bharti, Heart disease prediction using machine learning techniques, SN Computer Science, vol. 1, pp. 1–6, 2020.

[2] R. Choudhary, T. Gopalakrishnan, An Efficient Model for Predicting Liver Disease Using Machine Learning, Data Analytics in Bioinformatics: A Machine Learning Perspective, pp. 443–457, 2021.

[3] K. Sravani, G. Anushna, I. Maithraye, P. Chetan, Prediction of Liver Malady Using Advanced Classification Algorithms, In Machine Learning Technologies and Applications, Springer, Singapore, pp. 39-49, 2021.

[4] S. Jain, E. Kotsampasakou, G. F. Ecker, Comparing the performance of meta-classifiers—a case study on selected imbalanced data sets relevant for prediction of liver toxicity, Journal of computer-aided molecular design, vol. 32, pp. 583–590, 2018.

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

[6] F. S. Alotaibi, Implementation of a machine learning model to predict heart failure disease, International Journal of Advanced Computer Science and Applications, Vol. 10, 2019.

[7] H. El Massari, N. Gherabi, S. Mhammedi, An ontological model based on machine learning for predicting breast cancer, International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 13, 2022.

[8] G. Sailasya, G. L. A. Kumari, Analyzing the performance of stroke prediction using ML classification algorithms, International Journal of Advanced Computer Science and Applications, Vol. 12, 2021.

[9] A. Alfaidi, R. Aljuhani, B. Alshehri, Machine Learning: Assisted Cardiovascular Diseases Diagnosis, International Journal of Advanced Computer Science and Applications, Vol. 13, 2022.

[10] P. B. K. Chowdary, R. U. Kumar, An Effective Approach for Detecting Diabetes using Deep Learning Techniques based on Convolutional LSTM Networks, International Journal of Advanced Computer Science and Applications, Vol. 12, 2021.

[11] M. S. Azam, A. Rahman, S. M. H. S. Iqbal, Prediction of liver diseases by using few machine learning based approaches, Aust. J. Eng. Innov. Technol, vol. 2, pp. 85–90, 2020.

[12] J. Singh, S. Bagga, R. Kaur, Software-based prediction of liver disease with feature selection and classification techniques, Procedia Computer Science, vol. 167, pp. 1970–1980, 2020.

[13] A. Soni, A. Rai, A Comparative Analysis of Classification Algorithms in Liver Disease Detection, JNNCE Journal of Engineering & Management, Vol 5, No.1, 2021.

[14] M. K. Ram, C. Sujana, R. Srinivas, G. S. N. Murthy, A fact-based liver disease prediction by enforcing machine learning algorithms, In Computational Vision and Bio-Inspired Computing, Springer, Singapore, pp. 567–586, 2021.

[15] S. C. R. Nandipati, C. XinYing, K. K. Wah, Hepatitis C virus (HCV) prediction by machine learning techniques, Applications of Modelling and Simulation, vol. 4, pp. 89–100, 2020.

[16] M. S. P. Babu, M. Ramjee, S. Katta, Implementation of partitional clustering on ILPD dataset to predict liver disorders, In IEEE International Conference on Software Engineering and Service Science (ICSESS), IEEE, pp. 1094–1097, 2016.

[17] Y. M. M. Hassim, R. Ghazali, N. Hassan, N. Arbaiy, Firefly Algorithm for Functional Link Neural Network Learning, In Recent Trends in Mechatronics Towards Industry 4.0, Springer, Singapore, pp. 941–948, 2022.

[18] K. Ahammed, M. S. Satu, M. I. Khan, M. D. Whaiduzzaman, Predicting the infectious state of hepatitis c virus affected patient's applying machine learning methods, In 2020 IEEE Region 10 Symposium (TENSYMP), pp. 1371-1374, 2020.

[19] M. Pavithra, A. M. Sindhana, T. Subajanaki, Effective Heart Disease Prediction Systems Using Data Mining Techniques, Annals of the Romanian Society for Cell Biology, pp. 6566–6571, 2021.

[20] S. Sharma, M. Parmar, Heart diseases prediction using deep learning neural network model, International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 9, pp. 124–137, 2020.

[21] D. Waigi, D. S. Choudhary, D. P. Fulzele, D. Mishra, Predicting the risk of heart disease using advanced machine learning approach, Eur. J. Mol. Clin. Med, vol. 7, pp. 1638–1645, 2020.

[22] S. Mohan, C. Thirumalai, G. Srivastava, Effective heart disease prediction using hybrid machine learning techniques, IEEE Access, vol. 7, pp. 81542–81554, 2019.

[23] C. B. C. Latha, S. C. Jeeva, Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques, Informatics in Medicine Unlocked, vol. 16, pp. 100203, 2019.

[24] P. Rani, R. Kumar, N. M. O Ahmed, A. Jain, A decision support system for heart disease prediction based upon machine learning, Journal of Reliable Intelligent Environments, vol. 7, pp. 263–275, 2021.

[25] M. A. Khan, An IoT framework for heart disease prediction based on MDCNN classifier, IEEE Access, vol. 8, pp. 34717–34727, 2020.

[26] E. A. Zanaty, Support vector machines (SVMs) versus multilayer perception (MLP) in data classification, Egyptian Informatics Journal, vol. 13, pp. 177-183, 2012.

[27] R. E. Ali, H. El-Kadi, S. S. Labib, Y. I. Saad, Prediction of potential-diabetic obese-patients using machine learning techniques, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, 2019.

[28] N. Khan, I. Ahmed, M. Kiran, A. Adnan, Overview of technical elements of liver segmentation, International Journal of Advanced Computer Science and Applications, Vol. 7, 2016.

[29] M. Sharma, R. Parveen, The Application of Image Processing in Liver Cancer Detection, International Journal of Advanced Computer Science and Applications, Vol. 12, 2021.

[30] M. A. Farahat, K. A. Bahnasy, A. Abdo, S. M. Kamal, S. K. Kassim, A. Sharaf Eldin. Response prediction for chronic HCV genotype four patients to DAAs, International Journal of Advanced Computer Science and Applications, Vol. 7, 2016.

[31] N. Kumar, K. Sikamani Prediction of chronic and infectious diseases using machine learning classifiers-A systematic approach, Int J Intell Eng Syst, Vol, 13,2020.

[32] Y. Khourdifi, M. Bahaj Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization, International Journal of Intelligent Engineering and Systems, Vol. 12, 2019.

[33] M. Manur, A. K. Pani, P. Kumar A prediction technique for heart disease based on long Short term memory recurrent neural network. International Journal of Intelligent Engineering and Systems, Vol. 13, 2020.

[34] Fawzy, H., Rady, E. H. A., & Fattah, A. M. A. Forecasting time series using a hybrid ARIMA-ANN methodology. J. Appl. Probab. Stat, Vol 16, 95-106, 2021.

[35] Alrweili, H., & Fawzy, H. (2022). Forecasting crude oil prices using an ARIMA-ANN hybrid model. J Stat Appl Probab, Vol 11(3), 845-855, 2022.