Journal of Administrative
Sciences and Digital Technology
https://hicmis.journals.ekb.eg/
Frequency: Quarterly
Print ISSN: 3009-6243 Online ISSN:3009-6936

HICMIS-EGYPT
JASDT
Science Develop | Creates Future

FORMAT PROPOSED APPROACH FOR PREDICTING LIVER DISEASE

Ibrahim M. Attiya[1*]; Rania A. Abouelsoud[2] and A.S. Ismail[1]

1. Information System Dept., Fac. Comp. and Information System, Fayoum Univ., Fayoum, Egypt

1. High Inst. Comp. and Manag. Info. Syst., Minist. High Edu., Egypt.

2. Electrical Eng. Dept., Fac. Eng., Fayoum Univ., Egypt.

ARTICLE INFO

Article history:

Received: 06/11/2023

Revised: 13/12/2023

Accepted: 06/01/2024

Key words:

Prediction Disease, Classification, Prediction, Machine learning algorithms, Liver disease, ILPD, HCV

ABSTRACT

With the development of technology and the use of artificial intelligence in everything, especially in health care systems, where health care systems represents great importance in contributing to preserving human lives. Machine learning algorithms play an important role in classifying and predicting diseases.The efficiency of the machine learning algorithms used is measured by accuracy, recall, precision and F1score. The accuracy of machine learning algorithms depends on quality of data set used in training. Therefore, in this paper we present a proposal to predict the severity of liver disease for a data set obtained from the integration of other data sets, namely HCV ILPD, to improve the accuracy of classification and prediction of liver disease.

INTRODUCTION

Due to the incredible advancements in new technologies, especially artificial intelligence, and the huge prevalence of the disease, the number of deaths from the disease is also increasing due to multiple reasons, including inaccuracies in proper diagnosis of patients and lack of early detection. Several chronic diseases that contribute to their death contribute to the increased death toll.

Therefore, health systems play a vital role in human life by helping to protect human life, especially in predicting diseases because the earlier and more accurate the diagnosis, the faster the patient can be treated.

Chronic liver disease is one of the most common causes of death worldwide, affecting a large proportion of the population. The main cause of this disease is a combination of certain chemicals that damage the liver. Liver disease is the most important organ involved in various processes such as the breakdown of red blood cells.

Any abnormality in the liver is called liver disease. Such cases can manifest in different ways, including inflammation due to viral, non-infectious or autoimmune causes (hepatitis B and C), malignancy, scarring of the liver (cirrhosis) and metabolic disorders. Therefore, it is crucial to use machine learning to predict liver diseases.

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

https://doi.org/10.21608/hicmis.2024.353429
2024 HICMIS-Journal of Administrative Sciences and Digital Technology.
Published by High Institute for Computer and Management Information System.

As machine learning (ML) has made significant progress, it is currently one of the best options for solving problems, especially in a wide range of industries and applications, including speech recognition, data analysis, classification, natural language processing, healthcare, *etc*. . data analysis.

Machine learning models are data-dependent, nonlinear, and nonparametric models that are built based on assumptions about the underlying data generation 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 goal is to learn a mapping from input xi to output yi given a set of labeled input-output pairs M ={(Xi, Yi)}i=1N. M. This is the training set, N is the number of training samples. Supervised learning is divided into classification and regression. In unsupervised or descriptive learning methods, we only receive input, M={(Xi)}i=1N. Sometimes called knowledge discovery. Reinforcement learning is used to learn how to behave or act in response to occasional signals of reward or punishment.

This article only focuses on supervised learning. Two types of problems that supervised machine learning is designed to solve are classification problems and regression problems. Classification methods are used to classify certain data of a statistical population into different groups based on one or more basic attributes of the data. The nature of the data prevents them from choosing the best classification method.

Some researchers have been interested in predicting liver disease, but its application still needs improvement due to low accuracy of results.

With the proposed approach, the use of a set of tools supports the adaptation of machine learning algorithms to the data set (whatever the data set is) by determining the optimal parameters of the algorithm used, which helps to improve the accuracy of predicting the outcome of increased disease.

Another advantage is that the severity of a patient's disease can be predicted, and the speed with which disease can be detected can help human survival. Furthermore, dataset preprocessing significantly affects the accuracy of prediction results. Therefore, in this paper, we consider this issue when preprocessing data from scattered values and null values before the prediction process. The researchers used a range of machine learning algorithms such as decision trees (DT), gradient boosting (GB), support vector machines (SVM), K-nearest neighbors (KNN), wrapper classifiers, additive trees and multi-layer sensors (MLP) . For the prediction process, we will briefly discuss it.

1) Decision Tree (DT): is a learning algorithm used for classification and regression. It is one of the supervised algorithms that is often used in classification processes. The algorithm is represented in the form of a tree. This tree consists of a set of nodes, branches, and leaf nodes, as the internal nodes of this tree represent features of the dataset, the branches represent decision rules, and the results exist in the form of paper nodes. Decision trees are widely used because they resemble the way humans find solutions to problems (**Shah *et al.*, 2020)**.
2) Gradient Boosting (GB): This machine learning method is used for classification and regression tasks, *etc*. An ensemble of weak predictive models (usually decision trees) acts as a predictive model. The resulting method is called gradient boosted trees and often outperforms random forests when decision trees are weak learners **(Choudhary and Gopalakrishnan, 2021)**.
3) Support Vector Machine (SVM): SVM is a supervised machine learning model that uses a classification algorithm to solve binary classification problems. Integrate

support vector machines with a set of network supervised learning methods for regression and classification. SVM is an advanced technology used in mathematical learning theory and has a complete classification algorithm. These modeling techniques can be applied to both linear and nonlinear data classification **(Sravani *et al.*, 2021)**.

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

5) Bagging Classifier: This is a machine learning technique based on an ensemble of models developed using multiple training data sets derived from the original training data set. It computes multiple models and averages them to create the final ensemble model. Traditional bagging methods create multiple copies of a training set by randomly selecting molecules with replacements from the training set (**Jain *et al.*, 2018**).

6) Extra Trees (ET): Extra trees or extremely random trees are another type of machine learning ensemble classifier like RF. However, there are two fundamental differences between ET and RF. On one hand, these are ET samples without replacement, on the other hand, instead of selecting the best features, random features are selected to split the tree nodes. After creating multiple unpruned trees, ET makes predictions by computing the average of all tree results in the case of regression or by computing the majority decision in the case of classification (**Rabbi *et al.*, 2020**).

7) Random forest (RF): is an example for supervised learning algorithm that isused 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 Sometimes its 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 classification and prediction of liver disease

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.

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 (**Azam *et al.*, 2020**). 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 (**Singh *et al.*, 2020**). 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 Regression 74.36%, Naive Bayes 55.9%, SMO 71%, IBK 67.41, J48 70.67% Random Foreast 71.87% 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. 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 (**Ram *et al.*, 2021**). 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 (**Nandipati *et al.*, 2020**). 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 (**Sravani *et al.,* 2021**). 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 (**Babu *et al.,* 2016**). 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 (**Hassim *et al.,* 2022**). 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 prediction of liver disease and the severity of liver disease as shon in Table 1.

Proposed Method

Our proposed method aims to improve the accuracy of liver disease prediction as shown in Fig. 1.

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

Data collection

In this paper, we used the Indian liver patient dataset (ILPD) and hepatitis c virus HCV dataset, which can be obtained from the University of California (UCI) and Kaggle.

Dataset Description

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

ILPD Dataset

The researchers conducted experiments using the Indian Liver Patient Dataset (ILPD). This dataset was collected by the University of California, Irvine (UCI). The dataset contains 416 liver medical records and 167 non-liver medical records. The dataset was collected from test samples from northeastern Andhra Pradesh, India. "Dataset" is the class name used for classification (liver disease patients or non-liver disease patients). 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 Fig. 2.

HCV Dataset

The assortment comprises of segment data including age, research facility results from hepatitis C patients and blood givers. The UCI Vault gave the information. The only two qualities that lack numbers are category and gender. ALB, ALP, ALT, ALT, AST, BIL, CHE, CHOL, CREA, GGT, and PROT are the characteristics that relate to patient information from 1 to 4 and research center information from 5 to 14 respectively.

HCV dataset comprises of a bunch of components which are portrayed as displayed in Table 3.

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

Table 1. Summary of Literature Review

Method	Dataset	Accuracy	Reference
DT	ILPD	60 %	**. Azam *et al.* (2020)**
RF		64 %	
SVM		71 %	
KNN		66 %	
LR	ILPD	72.50 %	**Singh *et al.* (2020)**
Naive Bayes		55.74 %	
SVM		71.35 %	
KNN		64.15 %	
DT		68.78 %	
RF		71.53 %	
MLP	ILPD	72.50 %	**Ram *et al.* (2021)**
KNN		74.20 %	
LR		74.90 %	
DT		65.70 %	
RF		74.60 %	
GB		69.40 %	
SVM		85.70 %	
Naive Bayes		62.00 %	
AdaBoost		68.70 %	
XGBoost		70.30 %	
Bayesian		71.40 %	
Bagging		75.30 %	
KNN	HCV	47.35 %	**Nandipati *et al.* (2020)**
SVM		52.64 %	
RF		49.15 %	
Bagging		46.63 %	
Adaboost		50 %	
SVM	ILPD	78%	**Sravani *et al.* (2021)**
KNN	ILPD	64 %	**Babu *et al.* (2016)**
DT		69 %	
MLP-BP	ILPD	70.61 %	**Hassim *et al.* (2022)**
FLNN-BP		69.63 %	
FLNN-FA		70.73 %	
KNN	HCV	0.2548 %	**Ahammed *et al.* (2020)**
RF		0.2512 %	
Naïve Bayes		0.2476 %	
DT		0.2433 %	
LR		0.2433 %	

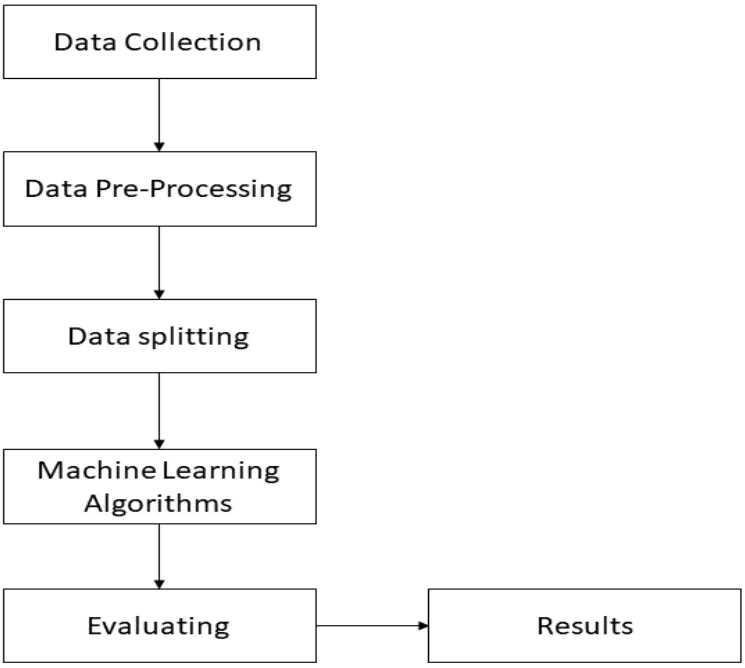

Fig. 1. Our Proposed Architecture

Table 2. Attribute Information ILPD Dataset

No.	Attribute	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.

Table 3. Attribute Information HCV Dataset

No.	Attribute	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 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.

Data Pre-processing

Data pre-processing is a crucial step in the data analysis pipeline, where raw data is transformed into a format suitable for analysis and modeling. It involves various techniques to clean, transform, and prepare the data for further analysis.Where it depends The quality of the output depends on the cleanliness of the data used in the experiment. Therefore, we briefly present the initial data processing steps, which greatly help in improving and increasing the efficiency of the algorithms used in the methodology followed Here's the data pre-processing used steps as shown in Fig 2.

Label Encoding

Label encoding is a technique used to convert categorical data into numerical format, which many machine learning algorithms require.

Handling Missing Values

Missing values are a common issue in datasets and can occur due to various reasons such as data collection errors, sensor malfunctions, or simply because the information is not available.

There are several approaches to handling missing values: Imputation: Replace missing values with estimated ones, such as the mean, median, or mode of the feature. Deletion: Remove records or features with missing values.The choice of method depends on the nature of the data and the analysis objectives.

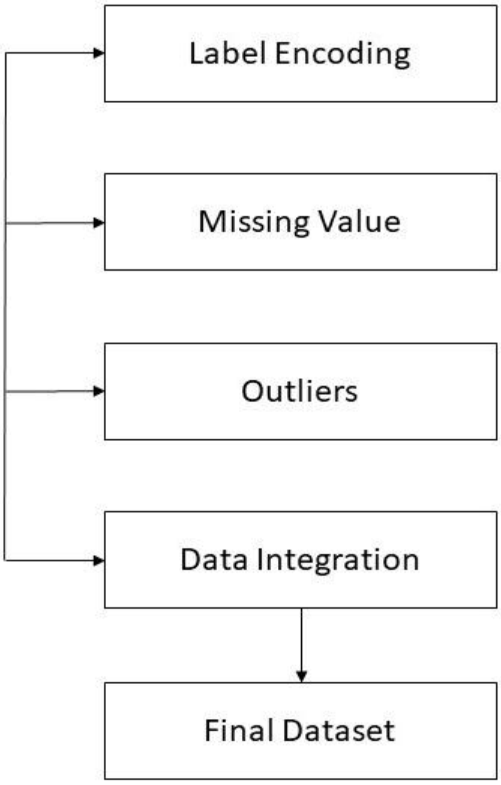

Fig. 2. Pre-Processing Steps Architecture

Outliers

Outliers are data points that significantly differ from the rest of the observations in a dataset. Outliers can arise due to measurement errors, data corruption, or genuine anomalies in the data. Handling outliers is important because they can skew statistical analyses and machine learning models.

Transformation: Applying transformations such as log transformation to make the distribution more symmetrical.Each of these techniques plays a crucial role in ensuring the quality and reliability of data used for analysis and modeling.

In machine learning classification, data integration refers to the process of combining data from multiple sources or datasets to create a unified dataset for training a classification and prediction model. This can involve integrating data from different sources. In this paper we integrate data between a ILPD and HCV for data integration.

Overall, data integration plays a crucial role in machine learning classification by enriching feature representation, addressing class imbalance and data sparsity issues, and enhancing model robustness, ultimately leading to more accurate and reliable classification models.

Data Splitting

Data splitting is a fundamental step in machine learning where the available dataset is divided into multiple subsets for training, testing purposes. This process is crucial for evaluating the performance of machine learning models and assessing their generalization ability.

In this research paper, the data set was divided 80 : 20 for training and testing respectively.

Machine Learning algorithms

In this process, we used a set of machine learning algorithms for the classification and prediction process.

Classification process

In the classification process, seven machine learning algorithms were used to classify liver disease, such as Decision Tree (DT), Gradient Boosting (GB), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Bagging Classifier, Extra Tree (ET), and Logistic Regression.

Prediction 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 predicting 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 prediction.

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, two different classes are clasified well.

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:

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

It can be transformed into: -

$$ln\left(\frac{p(x)}{1-p(x)}\right) = b0 + b1x \qquad (2)$$

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.

Random forest

It provides as an example of how regression and classification problems may both be solved using supervised machine learning (**Fawzy *et al.*, 2021; Alrweili and Fawzy, 2022**). 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.

$$g(x) = f_0(x) + f_1(x) + f_2(x) + \cdots \quad \textbf{(3)}$$

g is the sum of sample base models f_i

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 (**Fawzy *et al.*, 2021**). 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):

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) :

$$\text{Accuracy} = \frac{TN+TP}{TN+TP+FN_+FP} \qquad (4)$$

Table 4. Confusion Matrix

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

Precision refers to the ratio of the predicted positive observation values to the total predicted positive observations in the correct form (**Alrweili and Fawzy, 2022)**. Where the high accuracy means a low wrong positive rate, which can be obtained by the following Eq. (5):

$$\text{Precision} = \frac{TP}{FP+TP} \qquad (5)$$

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):

$$\text{Recall} = \frac{TP}{FN+TP} \qquad (6)$$

The value of the F1 score refers to the weighted average of precision and recall. Therefore, the score consider 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).

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

RESULTS AND DISCUSSION

Through this section, the results of our experiment on a set of algorithms to predict 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.

Classification process

Table 6 shows the results that we reached by using the proposed method using the the integrated dataset

Table 7 present our results for Predicting an Integrated data set as shown In table 8 , we show a comparison between the results obtained when using a set of different machine learning algorithms when classifying liver disease using HCV and ILPD dataset.

Table 6. Consequences of Classification Algorithms for Integrated Dataset

Algorithm	Accuracy	Precision	Recall	F1 Score
DT	98%	0.97	0.90	0.93
Gradient Boosting	99.8%	0.99	1.00	99.8
SVC	98.7%	0.99	0.94	0.97
K-Nearest Neighbors	97.7%	0.99	0.90	0.94
Bagging Classifier	98%	0.97	0.90	0.98
Logistic Regression	98%	0.98	0.92	0.95
Extra tree	99.2%	1.00	0.97	0.98

Table 7. Consequences of Prediction Algorithms for Integrated Dataset

Algorithm	Train Score	Accuracy
Decision Tree Regressor	1.0%	96.3%
Support Vector Regressor	98.6%	92%
Gradient Boosting Regressor	98.9%	96.1%

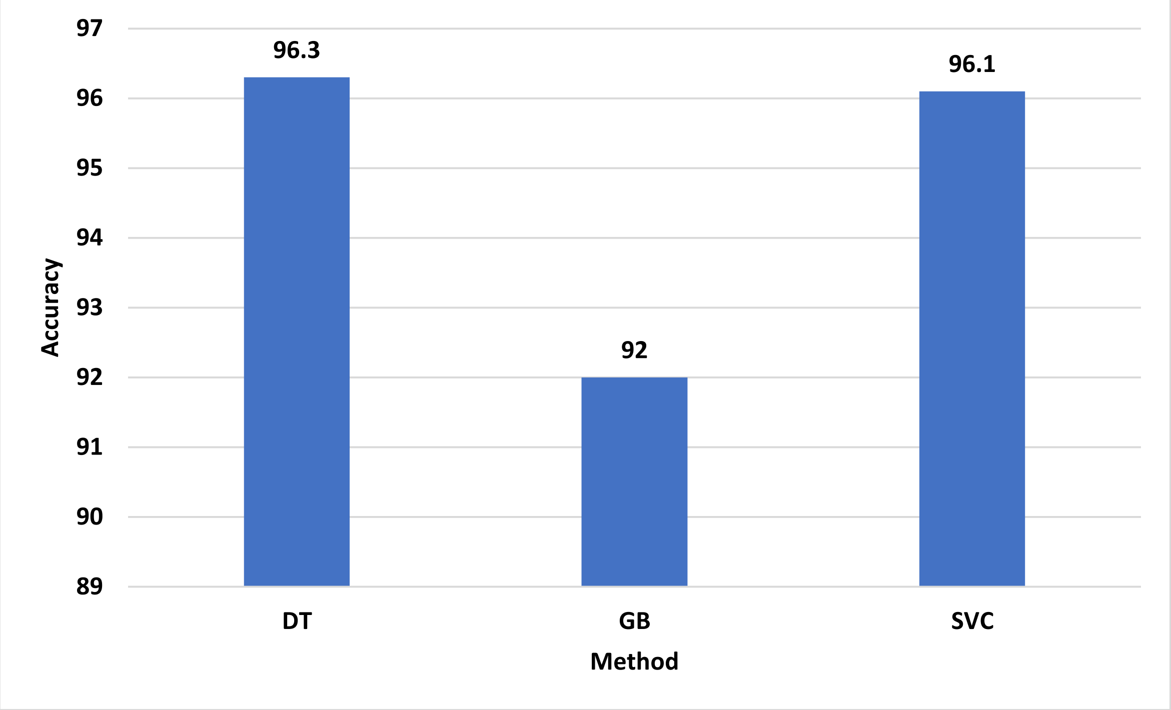

Fig. 3. Comparing Prediction Results for Integrated Dataset

Table 8. Comparison of Accuracy with Previous Studies for ILPD and HCV Dataset

ILPD			HCV		
Methods	**Accuracy**	**Reference**	**Methods**	**Accuracy**	**Reference**
DT	60 %	**. Azam *et al.* (2020)**			
KNN	66 %				
SVM	71 %				
SMO	71.3%	**Singh *et al.* (2020)**	DT	74%	**Attiya *et al.* (2023)**
J48	68.7%		GB	77%	
MLP	72.50%	**Ram *et al.* (2021)**	SVC	75%	
KNN	74.20%		RF	75%	
DT	65.70%		MLP	74%	
GB	69.40%		NB	77%	
SVM	85.70%		K-NN	77%	
Bagging	75.30%		Bagging	74%	
GB	74%	**Attiya *et al.* (2023)**	Adaboost	74%	
SVC	75%		LR	75%	
K-NN	77%		ET	76%	
Bagging	76%				
ET	80%				

REFERENCES

Alrweili, H. and H. Fawzy (2022). Forecasting crude oil prices using an ARIMA-ANN hybrid model. J. Stat Appl. Probab, 11 (3): 845-855.

Azam, M.S., A. Rahman and S.M.H.S. Iqbal (2020). Prediction of liver diseases by using few machine learning based approaches, Aust. J. Eng. Innov. Technol., 2: 85–90.

Babu, M.S.P., M. Ramjee, S. Katta (2016). Implementation of partitional clustering on ILPD dataset to predict liver disorders, In IEEE Int. Conf. on Software Eng. and Service Sci. (ICSESS), IEEE, 1094–1097.

Choudhary, R. and T. Gopalakrishnan (2021). An efficient model for predicting liver disease using machine learning, Data Analytics in Bioinformatics: A Machine Learning Perspect., 443–457.

Fawzy, H., E.H.A. Rady and A.M.A. Fattah (2021). Forecasting time series using a hybrid ARIMA-ANN methodology. J. Appl. Probab. Stat, 16: 95-106.

Hassim, Y.M.M., R. Ghazali, N. Hassan and N. Arbaiy (2022). Firefly algorithm for functional link neural network learning, In Recent Trends in Mechatronics Towards Industry 4.0, Springer, Singapore, 941–948.

Jain, S., E. Kotsampasakou and G.F. Ecker (2018). 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, 32: 583–590.

Nandipati, S.C.R., C. XinYing and K.K. Wah (2020). Hepatitis C virus (HCV) prediction by machine learning techniques, Applic. Mod. and Simul., 4 : 89 – 100.

Rabbi, M.F., S.M.M. Hasan and A.I. Champa (2020). Prediction of liver disorders using machine learning algorithms: a comparative study, In 2020 2nd Int. Conf. Advanced Information and Commun. Technol., (ICAICT), IEEE, 111 – 116.

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

Shah, D., S. Patel and S.K. Bharti (2020). Heart disease prediction using machine learning techniques, SN Computer Sci., 1: 1–6.

Singh, J., S. Bagga and R. Kaur (2020). Software-based prediction of liver disease with feature selection and classification techniques, Procedia Comp. Sci., 167: 1970 –1980.

Sravani, K., G. Anushna, I. Maithraye and P. Chetan (2021). Prediction of liver malady using advanced classification algorithms, In Machine Learning Technologies and Applications, Springer, Singapore, 39-49.