mathematics

MDPI

Article

Early Identification of Risk Factors in Non-Alcoholic Fatty Liver Disease (NAFLD) Using Machine Learning

Luis Rolando Guarneros-Nolasco [1], Giner Alor-Hernández [2], Guillermo Prieto-Avalos [3] and José Luis Sánchez-Cervantes [4,*]

[1] Tecnologías de la Información y Comunicación Área Desarrollo de Software, Universidad Tecnológica del Centro de Veracruz, Av. Universidad No. 350, Carretera Federal Cuitláhuac-La Tinaja, Loc. Dos Caminos, Cuitláhuac C.P. 94910, Veracruz, Mexico; luis.guarneros@utcv.edu.mx
[2] Tecnológico Nacional de México Campus Orizaba, Av. Oriente 9 No. 852 Col. Emiliano Zapata, Orizaba C.P. 94320, Veracruz, Mexico; giner.ah@orizaba.tecnm.mx
[3] Departamento de Ingeniería Eléctrica—Electrónica, Tecnológico Nacional de México Campus Mexicali, Av. Instituto Tecnológico S/N, Col. Plutarco Elías Calles, Mexicali C.P. 21376, Baja California, Mexico; guillermoprieto@itmexicali.edu.mx
[4] CONACYT-Instituto Tecnológico de Orizaba, Av. Oriente 9 No. 852 Col. Emiliano Zapata, Orizaba C.P. 94320, Veracruz, Mexico
* Correspondence: jlsanchez@conahcyt.mx; Tel.: +52-229-781-3796

Abstract: Liver diseases are a widespread and severe health concern, affecting millions worldwide. Non-alcoholic fatty liver disease (NAFLD) alone affects one-third of the global population, with some Latin American countries seeing rates exceeding 50%. This alarming trend has prompted researchers to explore new methods for identifying those at risk. One promising approach is using Machine Learning Algorithms (MLAs), which can help predict critical factors contributing to liver disease development. Our study examined nine different MLAs across four datasets to determine their effectiveness in predicting this condition. We analyzed each algorithm's performance using five important metrics: accuracy, precision, recall, f1-score, and roc_auc. Our results showed that these algorithms were highly effective when used individually and as part of an ensemble modeling technique such as bagging or boosting. We identified alanine aminotransferase (ALT), aspartate aminotransferase (AST), alkaline phosphatase (ALP), and albumin as the top four attributes most strongly associated with non-alcoholic fatty liver disease risk across all datasets. Gamma-glutamyl transpeptidase (GGT), hemoglobin, age, and prothrombin time also played significant roles. In conclusion, this research provides valuable insights into how we can better detect and prevent non-alcoholic fatty liver diseases by leveraging advanced machine learning techniques. As such, it represents an exciting opportunity for healthcare professionals seeking more accurate diagnostic tools while improving patient outcomes globally.

Keywords: ensembles; health prevention; machine learning; medical data

MSC: 68T01

Citation: Guarneros-Nolasco, L.R.; Alor-Hernández, G.; Prieto-Avalos, G.; Sánchez-Cervantes, J.L. Early Identification of Risk Factors in Non-Alcoholic Fatty Liver Disease (NAFLD) Using Machine Learning. *Mathematics* **2023**, *11*, 3026. https://doi.org/10.3390/math11133026

Academic Editors: Ravil Muhamedyev and Evgeny Nikulchev

Received: 28 May 2023
Revised: 4 July 2023
Accepted: 4 July 2023
Published: 7 July 2023

1. Introduction

According to the National Institute of Statistics and Geography of Mexico (INEGI, Spanish acronym), liver disease is the sixth leading cause of death in Mexico [1], mainly affecting people over 25. At least in 2020, Mexico recorded 41,492 deaths related to liver disease, so the prevalence of NAFLD exceeds 50% and is considered a national health problem [2].

The liver has many important functions, including food digestion and nutrient processing and distribution. A possible sign of liver disease is when the skin turns yellow, which is known as jaundice. Among the many different liver diseases, some, such as hepatitis, are

caused by viruses, whereas others may be the effect of excessive alcohol consumption or a long-term injury to the liver (i.e., cirrhosis). NAFLD is one of the most important liver diseases [3]. It can be described as a prolonged and progressive disorder characterized by hepatic steatosis in people who consume less than 20g of alcohol per day. Mexico has been attributed to the escalating prevalence of NAFLD risk factors, with more than 50% of its population showing at least one of these factors. Therefore, the medium-term for NAFLD sufferers is very pessimistic if immediate actions are not taken to reduce what is already considered a federal health problem [4].

Several authors have identified biomarkers to be used as a roadmap for diagnosing NAFLD. Among the most important are alanine aminotransferase, gamma-glutamyl transpeptidase [5], aspartate aminotransferase [6], alkaline phosphatase, age, and albumin [7]. Therefore, this analysis aims to identify the main risk factors for NAFLD in four datasets of clinical data from patients with liver disease by applying machine learning (ML) techniques, which is a promising technology for medical data analysis and prediction for disease diagnosis and early detection, to contribute to healthcare professionals improving better preventive diagnosis of NAFLD and more accurate and timely treatment. The algorithms are tested using boosting, bagging (ensemble methods), and non-ensemble methods.

The remainder of this article is organized as follows: Section 2 discusses current research on MLA in clinical datasets, MLA performance metrics, ensembles, and clinical and data sets are available in the data science community repository. Section 3 presents our MLA evaluation methodology. In Section 4, we present and discuss our results. Finally, in Section 5, we conclude and provide suggestions for future work.

2. Related Work

In this section, we analyze a series of research works that use ensemble learning for disease diagnosis and prevention. We classify the examined research into five primary ensemble learning trends: bagging, boosting, stacking, bagging-boosting, and bagging-boosting-stacking. These ensembles are widely implemented for the prediction and diagnosis of diseases such as schizophrenia, breast cancer, cardiovascular diseases, and liver diseases, among others.

2.1. Bagging Ensembles

In their work, Lin et al. [8] established a bagging ensemble and compared it with other algorithms to determine their efficiency in terms of schizophrenia detection. Next, the researchers found that a bagging ensemble with attribute selection could be an applicable method to support software tools for predicting the efficient results of schizophrenia. Ponnaganti and Anitha [9] suggested an Ensemble Bagging Weighted Voting Classification (EBWvc) method for the categorization of breast cancer. The researchers comparatively evaluated five metrics, and their results showed an increased performance of the EBWvc method if compared to similar existing classification methods. Chicco and Jurman [10] analyzed a set of EHRs and relied on MLA to predict the diagnosis of a series of liver diseases. The results confirmed the utility of ensemble learning for predicting the diagnosis of hepatitis C and cirrhosis. In the study of Anisha and Saranya [11], the researchers proposed a system for the early diagnosis of stroke disorder that relies on a homogeneous logistic regression ensemble classifier. The results of the experiment revealed a higher accuracy of the system than simple logistic regression initiatives. In turn, Devi et al. [12] proposed a method to analyze the clinical features influencing liver activity. The researchers applied Gradient Boost Regressor, Extra Tree Regressor, and Random Forest Regressor to the analyzed clinical data to find the highest feature importance. Then, the data were fitted to a set of classifiers to analyze performance metrics before and after feature scaling. As the main findings, Bagging Classifier was found to retain the accuracy of 72% before and after feature scaling with the top features from Random Forest Regressor.

In their work, Lin et al. [13] compared a bagging ensemble with other algorithms and found that a bagging ensemble framework may provide a suitable approach to building

tools for predicting cognitive function in schizophrenia. From a similar perspective, the research of Ejiofor and Ochei [14] relied on the strong capability of the ensemble bagging machine learning technique for predicting breast cancer. The bagging ensemble was implemented on two MLAs as base learners. The validation exhibited an accuracy value of 74% and 51% for Linear Regression (LR) and Decision Tree (DT), respectively. Rahman and Mahmood [15] proposed a model for identifying key biomarkers for heart disease. To this end, the researchers implemented three classification algorithms: Random Forest (RF), K-Nearest Neighbors (KNN), and Naïve Bayes, and three approaches: Lasso, Mutual Information, and Recursive Feature Elimination. Additionally, the researchers employed the bagging ensemble approach with RF to improve the results of their model, which exhibited 85.18% of accuracy through Recursive Feature Elimination with Bagging (Random Forest). In their work, Thomgkam et al. [16] explored the performance and effectiveness of machine learning methods. Their findings revealed that RF is superior to bagging in all the analyzed criteria. In turn, S. Yadav and Singh [17] implemented a model with classifiers to categorize patients affected by Parkinson's disease. As their main findings, Support Vector Classifier exhibited the highest accuracy (93.83%) on the basic classifiers, whereas bagging exhibited 73.28% accuracy from the ensemble technique.

2.2. Boosting Ensembles

In their work, Buyrukoglu [18] suggested ensemble learning methods to improve classification performance compared to single MLAs. The best classification performance of ensemble methods was achieved by boosting the ensemble (AdaBoost) (92.7%). The study found that the classification rate with ensemble learning methods increased between 3.2% and 7.2% compared to the AdaBoost ensemble method with single-based ML approaches. Researchers A. Singh et al. [19] constructed an intelligent hybrid approach for the identification of hepatitis. To avoid clinical experience, the approach relies on ensemble learning to reduce estimation time using many learners to solve a liver disease problem. After comparing different metrics, the findings revealed that both the k-means clustering and improved ensemble learning methodology achieved improved prediction results than other existing individual and integrated models. In Sarvestany et al. [20], the researchers trained six MLAs to identify advanced fibrosis. The MLA exhibiting the highest accuracy was an ensemble algorithm of classical algorithms. In turn, Dutta et al. [21] discussed liver disease diagnosis through different data mining algorithms. The best algorithm for liver disease detection was found to be DT, achieving an accuracy of 99.96%. The study of Verma and Mentha [22] suggested a novel ensemble learning algorithm for the classification of five datasets of the University of California Irvine (UCI). The researchers concluded that the suggested ensemble learning method is remarkably appropriate for handling the classification problem in the bioinformatics domain.

2.3. Stacking Ensembles

In their work, Meng [23] built a stacking ensemble model to predict the disorder evolution and medical results of Alpha-1-antitrypsin deficiency-associated liver disease (AATD-LD). The model uses meta-learning by mixing several supervised MLAs and can be implemented to predict the clinical outcome of other similar diseases. Al Telaq and Hewahi [24] proposed a method to predict liver disease by applying multiple MLAs on a public liver disease dataset. The results revealed that ensemble learning of different classifiers exhibited the highest accuracy (88%). In Gupta and Gupta [25], the researchers recommended a stacking architecture for efficiently predicting the diagnosis of cervical cancer. In Ponnaganti and Anitha [9], the researchers built a decision support system using the ensemble model with different metrics to evaluate the performance of the model. The researchers showed that the proposed method recorded a notable accuracy of 97% in classifying breast cancer data. Pouriyeh et al. [26] research aimed to evaluate the accuracy of different data mining classifiers for heart disease prediction using ensemble ML. The study used the Cleveland heart disease dataset, different classifiers, and applied classifier

ensemble prediction, bagging, boosting, and stacking to the dataset. The results of the experiment revealed that the Support Vector Machines (SVM) method with boosting outperformed the other methods. Kabir and Ludwing [27] attempted to improve the classification performance of multiple MLAs using super learning or stacked ensembles. The experiment results showed that super learning had a better classification performance than the individual base learners. Doganer et al. [28] compared the performance of different MLAs and found that stacking ensembles outperformed boosting, voting, bagging ensembles, and ML methods.

2.4. Bagging and Boosting Ensembles

Hakim et al. [29] evaluated the implementation of two ensemble-based MLAs of bagging and boosting with five different base classifiers for predicting myocardial infarction. The results showed that bagging with RF achieved higher accuracy. The research of Yadav and Pal [30] applied rule-based classification algorithms on a prepared dataset with three selected algorithms by bagging and boosting ensemble methods and calculated four metrics on a diabetes dataset. The research revealed that bagging exhibited the highest accuracy, namely, 98%. In Gao et al. [31], the researchers used different ensemble learning methods to improve heart disease prediction and compared these methods with MLAs. The results revealed that the bagging ensemble learning method with DT achieved the best performance. In the study by Taser [32], bagging and boosting methods were implemented using six different decision tree-based (DTB) classifiers on experimental data to predict diabetes. The experimental results showed that bagging and boosting outperformed the individual DTB classifiers. In Niranjana et al. [33], researchers evaluated the performance of different classification ensembles in the early detection of coronary heart disease based on risk factors. The Bagged Trees Ensemble Classifier was found to have the highest classification accuracy using four performance metrics. The research of Fraiwan et al. [34] studies various DTB ensemble methods to tackle the challenge of multi-class classification, and with their proposed approach, these methods achieved the highest classification accuracy (99.17%). The work of Dhilsath et al. [35] relied on two boosting classifiers and one bagging classifier to build a model for heart disease prediction. The research aims to evaluate the efficiency of grid search algorithms and random search algorithms by tuning the gradient boost parameters (GB), Adaboost, and RF. The findings revealed that the tunning strategy increased the ensemble learner's efficiency.

2.5. Bagging, Boosting, and Stacking Ensembles

In their work, Khanam et al. [36] applied an ExtraTreeClassifier (ETC) method to find the highly significant features of cervical cancer using different ensemble methods, including bagging, boosting, and stacking. The results of the experiment revealed that stacking combined with RF, SVM, ETC, Extreme Gradient Boosting (XGBoost), and bagging exhibited the highest accuracy (94.4%). In Niranjana et al. [33], the research aimed at predicting coronary heart disease by applying a risk factor method. Using predictive techniques, the researchers evaluated only three metrics, and their results showed that the stacked ensemble was the most effective in terms of accuracy. In Bang et al. [37], the researchers used 18 ML classifiers to determine prediction models of curative resection with different variables on early gastric cancer (U-EGC). As the main finding, XGBoost exhibited the best performance.

According to our literature review, the top eight MLAs used in different ensembles to detect and diagnose diseases such as heart disease, diabetes, breast cancer, liver disease, and schizophrenia include DT, RF, KNN, LR SVM, Artificial Neural Network (ANN), GB, and AdaBoost. In addition, current initiatives to detect and diagnose these chronic diseases are mainly based on the use of bagging and boosting ensembles with KNN, SVM, AdaBoost, RF, DT, NN, and logistic regression algorithms.

3. Materials and Methods

The following section describes how we analyzed the performance of nine MLAs independently and with two ML ensembles (bagging and boosting) on the four liver disease datasets.

3.1. Datasets

We identified four main datasets on clinical open access for liver diseases: the BUPA liver disorders (BLD) dataset, the Hepatocellular Carcinoma (HCC) Survival dataset, the Indian Liver Patient Dataset (ILPD), and the Cirrhosis Prediction Dataset (CPD). Table 1 summarizes the characteristics of each of these datasets.

Table 1. Characteristics of liver disease datasets.

Dataset	Number of Attributes	Number of Classes	Number of Records	Prediction/Diagnosis
BUPA Liver Disorders	6	1	345	Prediction/Diagnosis
HCC Survival	49	1	165	Prediction
ILPD	9	1	313	Diagnosis
CPD	19	1	424	Prediction

The BUPA liver disease database [38] is an open access database which is maintained in the Irvine University of California (UCI) online repository. It is often used to predict liver disorders based on blood tests and alcohol consumption. In addition, the BUPA dataset contains two classes, six numerical attributes, and 345 records (Table 2).

Table 2. Description of the Attributes of the BUPA Liver Disorders Dataset.

Attribute Name	Attribute Description
Mcv	Mean corpuscular volume
Alkphos	Alkaline phosphatase
Sgpt	Alanine aminotransferase
Sgot	Aspartate aminotransferase
Gammagt	Gamma-glutamyl transpeptidase
Drinks	Number of half-pint equivalents of alcoholic beverages drunk per day
Selector	Field created by BUPA researchers to split the data into trains/test sets

The data in the HCC Survival Dataset were collected from a university hospital in Portugal [39]. The dataset is open access, and it is stored in the online repository of UCI [40]. It contains accurate clinical data on 165 HCC patients. In addition, the HCC Survival Dataset contains two classes, 49 numerical attributes, and 165 records (Table 3).

Table 3. Description of the Attributes of the HCC Survival Dataset.

Attribute Name	Attribute Description
Gender	Gender of the patient
Symptoms	Symptoms
Alcohol	Alcohol
HBsAg	Hepatitis B Surface Antigen
HBeAg	Hepatitis B e Antigen
HBcAb	Hepatitis B Core Antibody
HCVAb	Hepatitis C Virus Antibody
Cirrhosis	Cirrhosis
Endemic countries	Endemic countries
Smoking	Smoking
Diabetes	Diabetes

Table 3. *Cont.*

Attribute Name	Attribute Description
Obesity	Obesity
Hemochromatosis	Hemochromatosis
AHT	Arterial Hypertension
CRI	Chronic Renal Insufficiency
HIV	Human Immunodeficiency Virus
NASH	Nonalcoholic Steatohepatitis
Esophageal varices	Esophageal varices
Splenomegaly	Splenomegaly
Portal hypertension	Portal hypertension
Portal vein thrombosis	Portal vein thrombosis
Liver metastasis	Liver metastasis
Radiological hallmark	Radiological hallmark
Age at diagnosis	Age at diagnosis
Grams/day	Grams of Alcohol per day
Packs/year	Packs of cigarettes per day
Performance status	Performance status
Encephalopathy	Encephalopathy
Ascites	Ascites degree
INR	International Normalized Ratio
AFP	Alpha-Fetoprotein (ng/mL)
Hemoglobin	Hemoglobin (g/gL)
MCV	Mean Corpuscular Volume (fl)
Leukocytes	Leukocytes (G/L)
Platelets	Platelets (G/L)
Albumin	Albumin (mg/dL)
Total Bil	Total bilirubin (mg/dL)
ALT	Alanine Transaminase (U/L)
AST	Aspartate Transaminase (U/L)
GGT	Gamma Glutamyl Transferase (U/L)
ALP	Alkaline phosphatase (U/L)
TP	Total proteins (g/dL)
Creatinine	Creatinine (mg/dL)
Number of nodules	Number of nodules
Major dimension	Major dimension of nodule
Dir. Bil	Direct bilirubin (mg/dL)
Iron	Iron (mcg/dL)
Sat	Oxygen saturation (%)
Ferritin	Ferritin (ng/mL)

The ILPD [41] contains 146 liver patient records and 167 non-liver patient records. The dataset was collected from patients in the northeast of Andhra Pradesh, India (Table 4).

Table 4. Description of the Attributes of the ILPD.

Attribute Name	Attribute Description
Age	Age of the patient
Gender	Gender of the patient
TB	Total Bilirubin
DB	Direct Bilirubin
Alkphos	Alkaline Phosphatase
Sgpt	Alanine Aminotransferase
Sgot	Aspartate Aminotransferase
TP	Total Proteins
ALB	Albumin
A/G	Albumin Ratio and Globulin Ratio
Selector	Field used to split the data into two sets (labeled by the experts)

The CPD [42] contains data collected from de Mayo Clinic trial in primary biliary cirrhosis (PBC) with a total of 424 PBC patients and 20 attributes. It is an open access dataset stored in the online repository of Kaggle (Table 5).

Table 5. Description of the Attributes of the CPD.

Attribute Name	Attribute Description
ID	Unique Identifier
N_Days	Number of days between registration and the earlier of death, transplantation, or study analysis time in July 1986
Status	Status of the patient: C (censored), CL (censored due to liver tx), or D (death)
Drug	Type of drug D-penicillamine or placebo
Age	Patient age in days
Sex	M(male) or F (female)
Ascites	Presence of ascites: N (No) or Y (Yes)
Hepatomegaly	Presence of hepatomegaly: N (No) or Y (Yes)
Spiders	Presence of spiders: N (No) or Y (Yes)
Edema	Presence of edema: N (no edema and no diuretic therapy for edema), S (edema present without diuretics, or edema resolved by diuretics), or Y (edema despite diuretic therapy)
Bilirubin	Serum bilirubin in [mg/dL]
Cholesterol	Serum cholesterol in [mg/dL]
Albumin	Albumin in [gm/dL]
Copper	Urine copper in [ug/day]
Alk_Phos	Alkaline phosphatase in [U/L]
SGOT	SGOT in [U/mL]
Triglycerides	Triglycerides in [mg/dL]
Platelets	Platelets per cubic [mL/1000]
Prothrombin	Prothrombin time in seconds [s]
Stage	Histologic stage of disease (1, 2, 3, or 4)

3.2. Machine Learning Classifiers

We have employed nine distinct classifying procedures from various domains of machine learning. The classifiers adopt a linear statistical approach, including LR [43], three tree-based techniques: RF [44], ExtraTrees (ETT) [45], and DT [46]; one SVM model [47]; an instance-based learning algorithm [48]. Additionally, we utilized three ensemble boosting methods consisting of GB [49], LightGBM (LGBM) [50], and AdaBoost [51]. We evaluated the performance of each method independently as well as with two ML ensembles- bagging technique [52] and boosting methodology [53]. Finally, all results were accurately documented for further analysis purposes.

3.3. Methodology

In order to evaluate the main risk factors of liver disease on clinical datasets, we employed a six-staged methodology as outlined in Guarneros-Nolasco et al. [54]. To enhance our analysis, we incorporated ensemble techniques such as bagging and boosting (as depicted in Figure 1) into this process. The Python programming language [55] and the Scikit–Learn [56] library algorithms were used. The stages involved were: (1) loading data dataset; (2) pre-processing data; (3) selecting attributes; (4) running ML models with bagging and boosting ensembles; (5) applying evaluation metrics; and finally; (6) processing MLA/classifier/ensemble performance results.

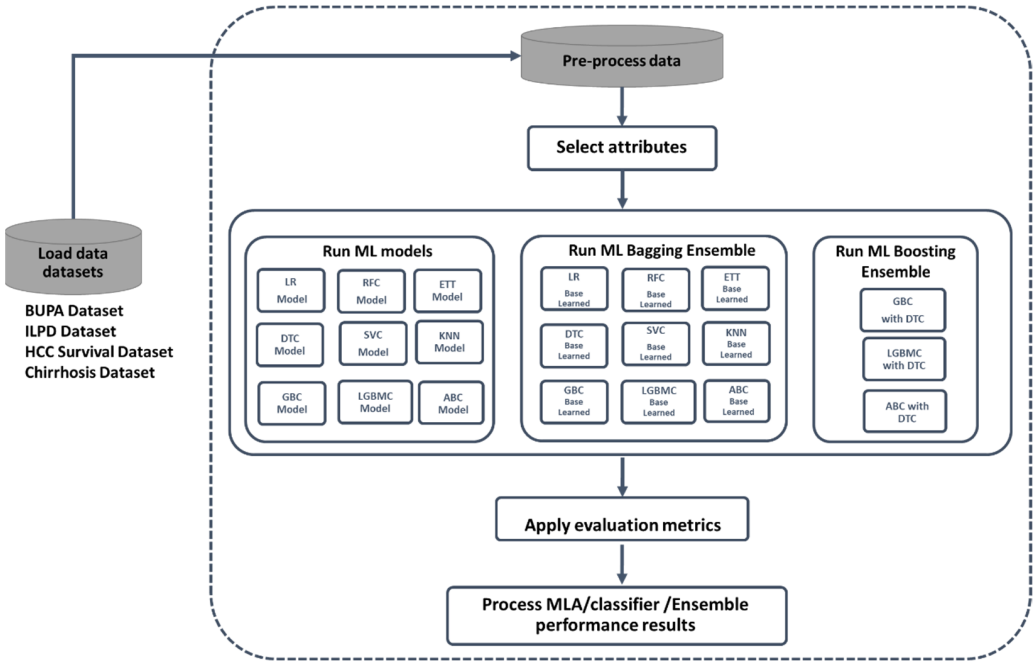


Figure 1. Methodology for evaluating liver disease datasets.

Each strategy arrangement can be depicted as follows:

1. **Load data dataset.** Select and stack information from the dataset containing clinical records of patients with liver diseases;
2. **Pre-process dataset.** Audit stacked information to get their content. At that point, select the classification variable to obtain the best results.
3. **Select attributes or main risk factors.** Use RF to select the top two and top four attributes from each dataset. Split data for training and testing (i.e., 70% for training and 30% for testing) and k = 10 cross-validation. Similary, calculate the best parameters for RandomizedSearchCV for n_estimators, max_attributes, and max_depth. Most of the algorithms have these parameters in common, except for KNN. The parameter ramdom_state was set to 42 in all the assessments;
4. **Run ML classifiers, bagging ensemble, and boosting ensemble:** Apply the nine ML classifiers to observe members with liver illnesses from healthy people. Tune bagging and boosting parameters such as n_estimators and max_samples on the train and test split and cross-validation techniques;
5. **Apply evaluation metrics.** Analyze MLA classification performance with respect to five criteria: accuracy, precision, recall, f1-score, and area under the curve (ROC-AUC);
6. **Process performance results.** Assemble and compare execution values from the nine MLAs with the bagging and boosting ensembles and record such outcomes for further analysis. At that point, select the best-performing MLA or ensemble.

4. Results and Discussion

In this section, we discuss the results of the ML performance analyses for five performance evaluation metrics or criteria (see step 5 of the methodology). We performed the performance evaluations of the classifiers first using the train-test split method (70–30%) and then using k-fold cross-validation (k = 10). We recorded five performance measures during the evaluations: accuracy, precision, recall, f1 score, and roc_auc.

4.1. Attribute Selection in Datasets

We analyzed the performance of RF to identify the top four attributes in the datasets.

(a) **The BUPA Liver Disorders Dataset**. We applied RF with the six numerical attributes of the dataset to identify and select the six most important ones. Figure 2a depicts the ranking of these attributes from the most important to the least important;
(b) **HCC Survival Dataset.** The 49 attributes were ranked using RF on the HCC Survival dataset. Figure 2b depicts a graph of said ranking. As in the previous case, the top six attributes were used in the classifier performance;
(c) **ILPD**. The 10 attributes were ranked using RF on the ILPD dataset. Figure 2c depicts a graph ranking the first six attributes, of which the top six were used in the analysis;
(d) **CPD**. The 19 attributes were ranked using RF in this dataset. Figure 2d graphically shows the ranking of said attributes, of which the top six were used in the analysis.

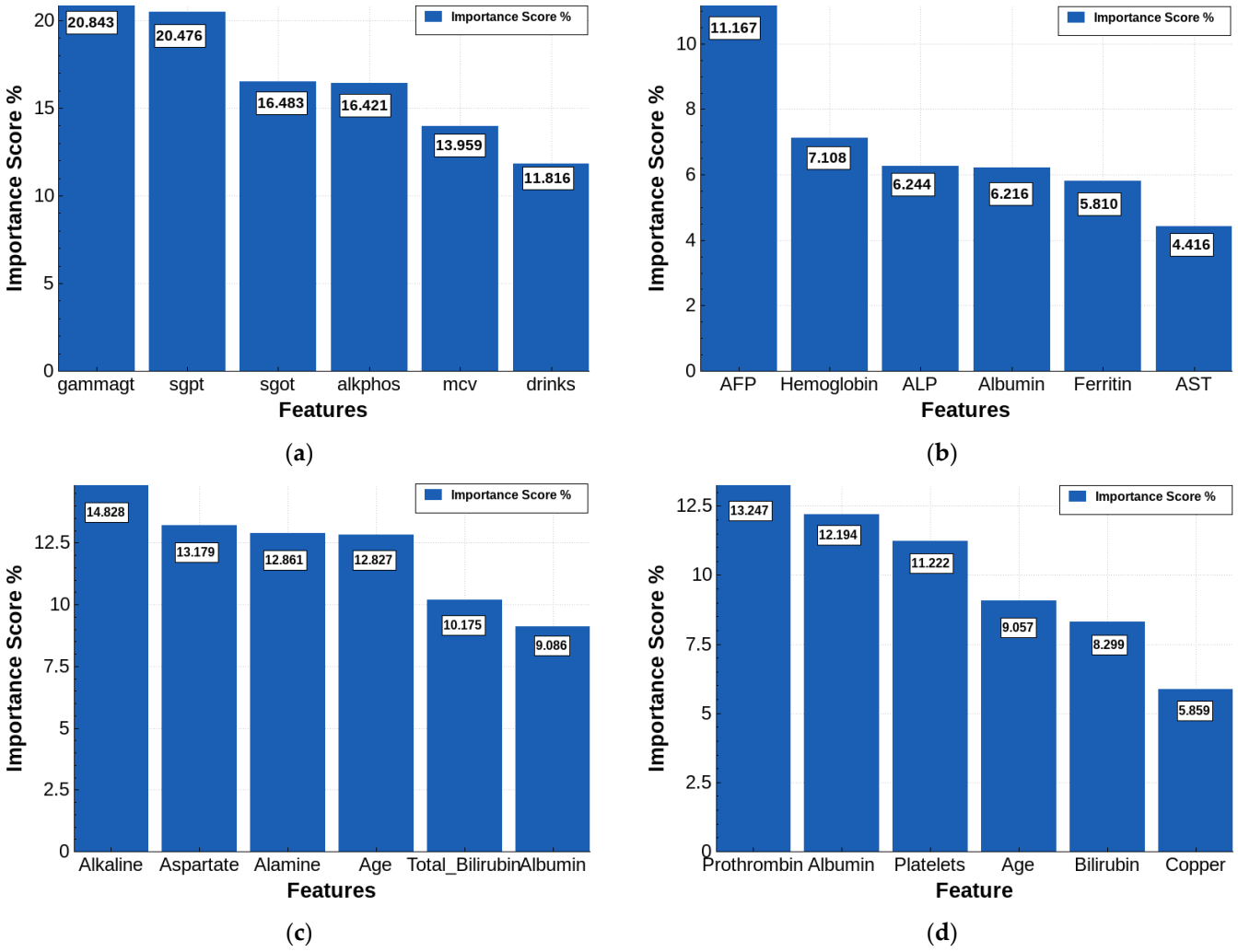


Figure 2. Ranking of attributes selected by RF for predictions on (**a**) 6/6 attributes on the BUPA dataset, (**b**) 6/49 attributes on the HCC Survival Dataset, (**c**) 6/10 attributes on ILPD, and (**d**) 6/19 attributes on CPD.

4.2. Results

We analyzed the performance of the nine ML classifiers (i.e., AdaBoost, DT, ETT, GB, KNN, LGBM, LR, RG, and SVC) on the top four attributes of each dataset using the train-test data split technique (70–30%), cross-validation and no ensembles, bagging ensemble, and boosting ensemble. We tuned bagging and boosting parameters such as n_estimators and max_samples with the train and test split and cross-validation techniques. The analysis results are discussed below.

4.2.1. Classifier Performance on the BUPA Dataset

We tested the performance of the nine ML classifiers (AdaBoost, DT, ET, GB, KNN, LGBM, LR, RF, and SVC) on the top four attributes of the BUPA dataset. These attributes comprised gammagt (score = 20.84), sgpt (score = 20.48), sgot (score = 16.48), and alkphos (score = 16.42). Tables 6–8 summarize the results of our analysis.

Table 6. Non-ensemble learning performance analysis of classifiers on top four attributes—BUPA dataset.

Ensemble	Technique	Predictive Model	Performance Evaluation Metrics				
			% Accuracy	% Precision	% Recall	% f1-Score	% roc_auc
Non-ensemble learning	Train and test split	AdaBoost	66.35	69.23	75.00	72.00	64.77
		DT	65.38	70.00	70.00	70.00	64.55
		ETT	72.12	73.85	80.00	76.80	70.68
		GB	68.27	70.15	78.33	74.02	66.44
		KNN	66.35	**74.51**	63.33	68.47	66.89
		LGBM	65.38	70.00	70.00	70.00	64.55
		LR	**73.08**	74.24	81.67	**77.78**	**71.52**
		RF	70.19	72.31	78.33	75.20	68.71
		SVC	70.19	69.86	**85.00**	76.69	67.50
	Cross-validation	AdaBoost	66.95	69.78	75.97	72.24	70.40
		DT	58.87	64.53	65.88	64.09	57.44
		ETT	68.09	70.25	77.70	73.48	71.96
		GB	66.69	69.32	76.20	71.93	73.02
		KNN	65.19	70.45	67.50	68.44	67.42
		LGBM	70.45	**73.53**	77.94	75.11	73.00
		LR	68.66	70.18	81.42	74.59	71.59
		RF	69.61	71.50	79.66	74.92	71.41
		SVC	**71.60**	71.21	**86.49**	**77.62**	**74.58**

The best-performing results for the critical metrics calculated by each algorithm are shown in bold.

Table 7. Bagging classifier performance analysis using nine base learners on the BUPA dataset (top four attributes).

Ensemble	Technique	Base Estimator	Performance Evaluation Metrics				
			% Accuracy	% Precision	% Recall	% f1-Score	% roc_auc
Bagging ensemble	Train and test split	AdaBoost	68.27	69.57	80.00	74.42	66.14
		DT	71.15	72.73	80.00	76.19	69.55
		ETT	**74.04**	**74.63**	83.33	**78.74**	**72.35**
		GB	67.31	69.70	76.67	73.02	65.61
		KNN	70.19	70.42	83.33	76.34	67.80
		LGBM	65.38	68.18	75.00	71.43	63.64
		LR	73.08	74.24	81.67	77.78	71.52
		RF	68.27	69.01	81.67	74.81	65.83
		SVC	71.15	70.27	**86.67**	77.61	68.33
	Cross-validation	AdaBoost	68.39	70.33	76.46	70.95	69.62
		DT	69.01	**73.99**	76.67	74.12	72.84
		ETT	71.29	70.67	83.09	75.91	73.88
		GB	**71.61**	71.69	80.93	74.26	**75.84**
		KNN	67.50	67.95	81.69	74.11	70.22
		LGBM	69.86	73.17	79.83	74.74	75.28
		LR	69.81	70.29	81.07	74.52	71.73
		RF	69.28	71.83	81.71	76.03	74.51
		SVC	69.55	69.13	**84.87**	**76.38**	75.18

The best-performing results for the critical metrics calculated by each algorithm are shown in bold.

Table 8. Boosting ensemble learning performance analysis of classifiers on the BUPA dataset (top four attributes).

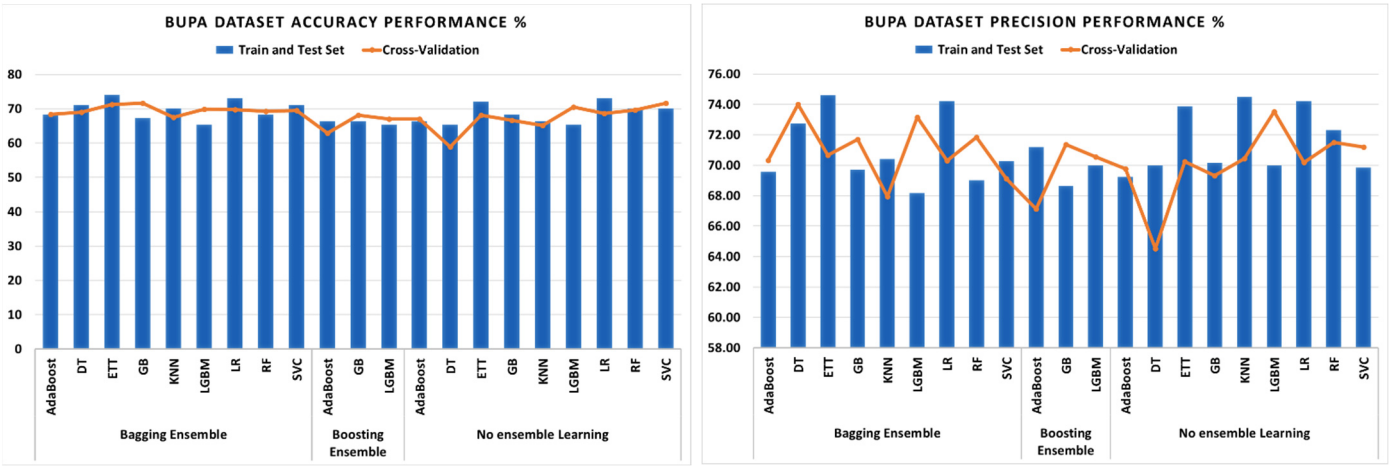
Ensemble	Technique	Predictive Model	Performance Evaluation Metrics				
			% Accuracy	% Precision	% Recall	% f1-Score	% roc_auc
Boosting ensemble	Train and test split	AdaBoost	**66.35**	**71.19**	70.00	70.59	**65.68**
		GB	**66.35**	68.66	**76.67**	**72.44**	64.47
		LGBM	65.38	70.00	70.00	70.00	64.55
	Cross-validation	AdaBoost	62.87	67.14	69.85	67.99	68.76
		GB	**68.14**	**71.37**	**75.35**	**72.60**	69.05
		LGBM	66.97	70.54	74.99	72.15	**71.33**

The best-performing results for the critical metrics calculated by each algorithm are shown in bold.

As can be observed from Table 6, with non-ensemble learning, we calculated algorithm performance using default values. With train and test split, LR exhibited the best results in terms of accuracy (73.08%). When using cross-validation, SVC exhibited the highest accuracy (71.60%). Conversely, the lowest-performing classifiers with respect to accuracy included DT and LGBM (65.38%) in train and test split and DT (58.87%) in cross-validation. Regarding precision, KNN proved to be the best-performing classifier in train and test split, while LGBM in cross-validation, with precision scores of 74.51% and 73.53%, respectively. The lowest-performing classifiers in terms of precision were AdaBoost (69.23%) with train and test split and DT (64.53%) with cross-validation. In conclusion, on the BUPA dataset with train and test split, LR exhibited good performance in terms of accuracy, f1-score, and roc_auc, whereas KNN outperformed in precision, and SVC outperformed in terms of recall. As for cross-validation, SVC exhibited the best results in accuracy, recall, f1-score, and roc_auc, whereas LGBM exhibited the best performance in precision.

With the bagging ensemble classifier, we used number of trees = 50 and max_samples = 0.5. As can be observed from Table 7, ETT achieved the highest accuracy (74.04%) with train and test split, whereas GB exhibited the highest accuracy with cross-validation (71.61%). Conversely, the lowest-performing classifiers with respect to accuracy included LGBM (65.38%) in train and test split and KNN (67.50%) in cross-validation. As for precision, ET proved to be the best-performing classifier (74.63%) when using train and test split and DT when using cross-validation (73.99%). The lowest-performing classifiers in terms of precision were LGBM (65.18%) in train and test split and KNN (67.95%) in cross-validation. In conclusion, on the BUPA dataset, when using the train and test split, ETT exhibited good performance in terms of accuracy, precision, f1-score, and roc_auc, whereas SVC exhibited the best results in recall. Conversely, when using cross-validation, GB achieved the best results in accuracy, and roc_auc and DT outperformed the other classifiers in terms of precision, and SVC exhibited the best performance in recall and f1-score.

With the boosting ensemble classifier, we used the best arguments for the BUPA dataset, including n_estimators = 200, max_depth = None, and max_features = sqrt. Our results revealed that when using train and test split, AdaBoost and GB exhibited the best accuracy results (66.35%), whereas GB outperformed when using cross-validation (68.14%). Conversely, the lowest-performing classifiers with respect to accuracy included LGBM (65.38%) with train and test split and AdaBoost (62.87%) with cross-validation. As regards precision, AdaBoost proved to be the best-performing classifier with train and test split and GB with cross-validation (scores of 71.19% and 71.37%, respectively). On the other hand, the lowest-performing classifiers in terms of precision were GB (68.66%) when using train and test split and AdaBoost (67.14%) when using cross-validation (see Figure 3). In conclusion, on the BUPA dataset using a boosting ensemble AdaBoost exhibited good performance with train and test split in terms of accuracy, precision, and roc_auc, whereas GB proved to be the best-performing classifier in recall and f1-score. On the other hand, with cross-validation, GB outperformed in accuracy, precision, recall, and f1-score, whereas LGBM exhibited the best performance in terms of roc_auc.



(**a**) (**b**)

Figure 3. ML classifier performance in accuracy (**a**) and precision (**b**): train and test split vs. K-Fold cross-validation—BUPA dataset.

4.2.2. Classifier Performance on the CPD

We tested the performance of the nine ML classifiers on the top four attributes of the CPD. The selected attributes comprised prothrombin (score = 13.25), albumin (score = 12.19), platelets (score = 11.22), and age (score = 9.05). Tables 9–11 summarize the results of the analysis.

Table 9. Non-ensemble learning performance analysis of classifiers on the CPD (top four attributes).

Ensemble	Technique	Predictive Model	Performance Evaluation Metrics				
			% Accuracy	% Precision	% Recall	% f1-Score	% roc_auc
Non-ensemble learning	Train and test split	AdaBoost	**69.05**	**57.14**	45.45	50.63	63.58
		DT	57.14	41.07	**52.27**	46.00	56.01
		ETT	65.08	50.00	40.91	45.00	59.48
		GB	66.67	52.78	43.18	47.50	61.23
		KNN	62.70	44.00	25.00	31.88	53.96
		LGBM	**69.05**	56.76	47.73	**51.85**	**64.11**
		LR	65.87	51.43	40.91	45.57	60.09
		RF	66.67	52.78	43.18	47.50	61.23
		SVC	64.29	47.37	20.45	28.57	54.13
	Cross-validation	AdaBoost	71.79	60.04	48.54	53.06	68.53
		DT	61.98	45.64	**50.57**	47.17	59.29
		ETT	68.20	55.01	44.48	48.57	72.01
		GB	**72.74**	62.84	49.90	**54.83**	71.64
		KNN	67.69	54.42	28.71	37.10	62.58
		LGBM	69.86	58.37	49.44	52.36	68.32
		LR	72.25	66.00	41.75	50.04	**75.84**
		RF	70.10	59.51	46.65	51.60	72.01
		SVC	69.36	**70.17**	22.43	32.04	65.46

The best-performing results for the critical metrics calculated by each algorithm are shown in bold.

As can be observed from Table 9, with non-ensemble learning, LGBM achieved the best results in terms of accuracy (69.05%) with the train and test split technique, whereas GB exhibited the highest accuracy (72.74%) with cross-validation. Conversely, DT was the lowest-performing classifier in accuracy with both train and test split and cross-validation, with scores of 57.14% and 61.98%, respectively. Regarding precision, AdaBoost proved to be the best-performing classifier in train and test split (57.14%), while SVC showed the best results in cross-validation (70.17%). The lowest-performing classifier in terms of precision

was DT with both train and test split and cross-validation, with scores of 41.07% and 45.64%, respectively. In conclusion, on the CPD with train and test split, LGBM exhibited good performance in accuracy, f1-score, and roc_auc, whereas AdaBoost showed the best results in precision. Also, DT exhibited the best results in terms of recall. On the other hand, with cross-validation, GB outperformed the other classifiers in accuracy and f1-score. SVC exhibited the best performance in precision, DT showed the highest score in recall, and LR outperformed the other classifiers in terms of roc_auc.

Table 10. Bagging classifier performance using nine base learners on the CPD (top four attributes).

Ensemble	Technique	Base Estimator	Performance Evaluation Metrics				
			% Accuracy	% Precision	% Recall	% f1-Score	% roc_auc
Bagging ensemble	Train and test split	AdaBoost	66.67	52.78	43.18	47.50	61.23
		DT	66.67	52.94	40.91	46.15	60.70
		ETT	65.87	51.61	36.36	42.67	59.04
		GB	65.87	51.52	38.64	44.16	59.56
		KNN	61.90	42.86	27.27	33.33	53.88
		LGBM	**67.46**	**54.05**	**45.45**	**49.38**	**62.36**
		LR	65.87	51.43	40.91	45.57	60.09
		RF	66.67	52.94	40.91	46.15	60.70
		SVC	65.08	50.00	20.45	29.03	54.74
	Cross-validation	AdaBoost	**73.21**	**68.28**	**50.39**	**55.73**	72.68
		DT	69.15	58.06	44.34	51.55	72.80
		ETT	71.05	65.23	43.33	53.12	74.97
		GB	71.54	66.56	49.38	55.51	74.62
		KNN	69.37	59.01	31.13	41.82	65.58
		LGBM	70.81	65.67	49.62	54.10	72.20
		LR	72.01	67.23	41.75	49.06	**75.99**
		RF	72.02	66.37	43.09	54.18	75.15
		SVC	68.39	60.33	16.24	29.15	65.09

The best-performing results for the critical metrics calculated by each algorithm are shown in bold.

Table 11. Boosting ensemble learning performance analysis of classifiers on the CPD (top four attributes).

Ensemble	Technique	Predictive Model	Performance Evaluation Metrics				
			% Accuracy	% Precision	% Recall	% f1-Score	% roc_auc
Boosting ensemble	Train and test split	AdaBoost	**66.67**	**52.38**	**50.00**	**51.16**	**62.80**
		GB	63.49	47.62	45.45	46.51	59.31
		LGBM	65.08	50.00	47.73	48.84	61.06
	Cross-validation	AdaBoost	66.51	50.60	44.22	46.69	63.68
		GB	67.50	52.35	45.32	48.10	**68.25**
		LGBM	**67.72**	**53.89**	**48.65**	**50.00**	67.33

The best-performing results for the critical metrics calculated by each algorithm are shown in bold.

As summarized in Table 10, with the bagging ensemble classifier, we used number of trees = 50 and max_samples = 0.5. With the train and test split strategy, LGBM achieved the highest accuracy (67.46%), whereas AdaBoost exhibited the highest accuracy (73.21%) with cross-validation. Conversely, the lowest-performing classifiers with regards to accuracy were KNN (61.90%) in train and test split and SVC (68.39%) in cross-validation. Regarding precision, LGBM proved to be the best-performing classifier (54.05%) in train and test split, and AdaBoost exhibited the highest score during cross-validation (68.28%). The lowest-performing classifiers in precision included KNN (42.86%) in train and test split and DT (58.06%) in cross-validation. In conclusion, on the CPD with train and test split, LGBM exhibited good performance in accuracy, precision, recall, f1-score, and roc_auc.

Conversely, with cross-validation, AdaBoost showed the best results in accuracy, precision, recall, and f1-score, while LR outperformed the other classifiers in roc_auc.

As observed in Table 11, we analyzed the performance of the boosting ensemble using the best arguments, i.e., n_estimators = 500, max_depth = 3, and max_features = auto. With the train and test split technique, AdaBoost achieved the highest accuracy (66.67%), whereas LGBM displayed the highest accuracy score (67.72%) with the cross-validation technique. On the other hand, the lowest-performing classifiers with regard to accuracy included GB (63.49%) with train and test split and AdaBoost (66.51%) with cross-validation. As for precision, AdaBoost proved to be the best-performing classifier in train and test split (52.38%), while LGBM achieved the highest score with cross-validation (53.89%). On the other hand, GB (47.62%) was the lowest-performing classifier in terms of precision with train and test split, and AdaBoost showed the lowest score (50.60%) with cross-validation (see Figure 4). In conclusion, in the boosting ensemble applied on the CPD, AdaBoost exhibited good performance with train and test split across the five metrics. However, with cross-validation, LGBM outperformed the other classifiers in accuracy, precision, recall, and f1-score, whereas GB exhibited the best performance in roc_auc.

(**a**) (**b**)

Figure 4. ML classifier performance in accuracy (**a**) and precision (**b**): train test split vs. K-fold cross-validation—CPD.

4.2.3. Classifier Performance on the HCC Survival Dataset

We tested the performance of the nine ML classifiers on the top four attributes of the HCC Survival dataset. The selected attributes comprised AFP (score = 11.17), hemoglobin (score = 7.11), ALP (score = 6.24), and albumin (score = 6.22). Tables 12–14 summarize the results of the analysis.

As summarized in Table 12, with non-ensemble learning, DT and GB achieved the highest accuracy (75.81%) with train and test split, while GB exhibited the highest accuracy (74.60%) with cross-validation. On the other hand, the lowest-performing classifier in terms of accuracy was SVC in both train and test split and cross-validation, with scores of 45.16% and 53.95%, respectively. As for precision, GB and DT proved to be the best-performing classifiers (75.81%) with the train and test split strategy, whereas KNN exhibited the best result in cross-validation (76.00%). SVC in train test split and LR in cross-validation were the lowest-performing classifiers in precision, with values of 43.33% and 41.01%, respectively. In conclusion, on the HCC Survival dataset with train and test split, GB and DT exhibited good performance in accuracy, precision, f1-score, and roc_auc, whereas SVC outperformed the other classifiers in recall with a value of 100%. When using cross-validation, GB outperformed the other classifiers in accuracy, f1-score, and roc_auc. Additionally, KNN exhibited the best performance in precision, and SVC displayed the best score in recall.

Table 12. Non-ensemble learning performance analysis of classifiers on the HCC Survival dataset (top four attributes).

Ensemble	Technique	Predictive Model	Performance Evaluation Metrics				
			% Accuracy	% Precision	% Recall	% f1-Score	% roc_auc
Non-ensemble learning	Train and test split	AdaBoost	67.74	59.38	73.08	65.52	68.48
		DT	**75.81**	68.97	76.92	**72.73**	**75.96**
		ETT	69.35	62.07	69.23	65.45	69.34
		GB	**75.81**	**68.97**	76.92	**72.73**	**75.96**
		KNN	67.74	60.71	65.38	62.96	67.41
		LGBM	70.97	64.29	69.23	66.67	70.73
		LR	62.90	54.29	73.08	62.30	64.32
		RF	70.97	65.38	65.38	65.38	70.19
		SVC	45.16	43.33	**100.00**	60.47	52.78
	Cross-validation	AdaBoost	72.55	73.41	70.56	71.06	78.57
		DT	66.71	66.81	63.89	64.47	66.81
		ETT	74.55	72.74	74.87	73.09	82.39
		GB	**74.60**	74.20	73.14	**73.17**	**83.65**
		KNN	73.12	**76.00**	67.81	70.12	75.54
		LGBM	69.64	70.41	69.57	68.76	79.87
		LR	57.38	41.01	52.90	45.51	74.11
		RF	73.98	74.27	71.91	72.49	82.53
		SVC	53.95	52.51	**99.23**	67.73	80.02

The best-performing results for the critical metrics calculated by each algorithm are shown in bold.

Table 13. Bagging classifier performance analysis using nine base learners on top four attributes—HCC Survival dataset.

Ensemble	Technique	Base Estimator	Performance Evaluation Metrics				
			% Accuracy	% Precision	% Recall	% f1-Score	% roc_auc
Bagging ensemble	Train and test split	AdaBoost	**74.19**	66.67	76.92	**71.43**	**74.57**
		DT	69.35	65.22	57.69	61.22	67.74
		ETT	69.35	62.96	65.38	64.15	68.80
		GB	**74.19**	**69.23**	69.23	69.23	73.50
		KNN	70.97	65.38	65.38	65.38	70.19
		LGBM	69.35	62.07	69.23	65.45	69.34
		LR	62.90	54.84	65.38	59.65	63.25
		RF	**74.19**	**69.23**	69.23	69.23	73.50
		SVC	45.16	43.33	**100.00**	60.47	52.78
	Cross-validation	AdaBoost	75.98	77.71	**79.60**	**78.59**	82.99
		DT	73.50	74.86	72.99	76.24	82.59
		ETT	76.50	75.91	75.55	74.41	83.74
		GB	76.98	**78.45**	77.67	76.48	**84.98**
		KNN	73.57	75.28	72.17	71.95	75.82
		LGBM	74.55	75.53	73.42	71.08	82.81
		LR	73.14	71.86	74.91	72.02	80.06
		RF	**77.45**	76.27	76.75	76.42	84.56
		SVC	49.60	45.26	69.23	51.99	79.92

The best-performing results for the critical metrics calculated by each algorithm are shown in bold.

With the bagging ensemble classifier, we used number of trees = 50 and max_samples = 0.5. When using the train and test split strategy, AdaBoost, GB, and RF achieved the highest accuracy score (74.19%), whereas RF exhibited the highest accuracy score (77.45%) with cross-validation. On the other hand, SVC proved to be the lowest-performing classifier in accuracy, with values of 45.16% and 49.60% in train and test split and cross-validation, respectively. As regards precision, GB and RF exhibited the best score with train and test split (69.23%), while GB outperformed the other classifiers when using the cross-validation

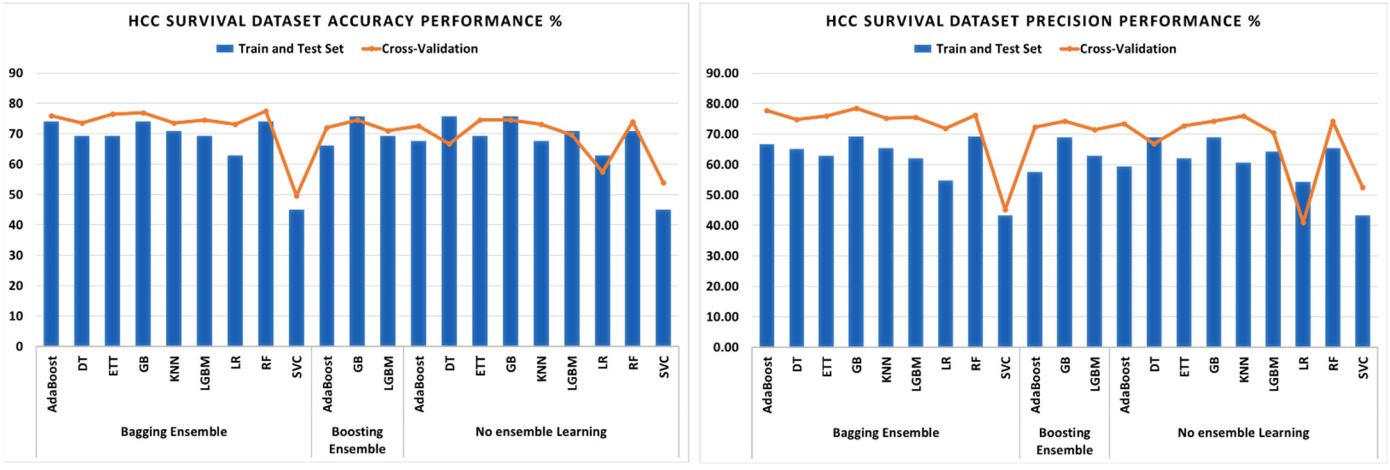
technique (78.45%). The lowest-performing classifier in precision was SVC, with scores of 43.33% in train and test split and 45.26% in cross-validation. In conclusion, on the HCC Survival dataset with train and test split, AdaBoost exhibited good performance in accuracy, f1-score, and roc_auc. GB performed best in accuracy and precision and SVC in recall. Conversely, with cross-validation, AdaBoost displayed the highest recall and f1-scores, RF outperformed in accuracy, and GB exhibited the highest scores in precision and roc_auc.

Table 14. Boosting ensemble learning performance analysis of classifiers on the HCC Survival dataset (top four attributes).

Ensemble	Technique	Predictive Model	Performance Evaluation Metrics				
			% Accuracy	% Precision	% Recall	% f1-Score	% roc_auc
Boosting ensemble	Train and test split	AdaBoost	66.13	57.58	73.08	64.41	67.09
		GB	**75.81**	**68.97**	**76.92**	**72.73**	**75.96**
		LGBM	69.35	62.96	65.38	64.15	68.80
	Cross-validation	AdaBoost	72.07	72.24	71.54	71.07	78.95
		GB	**74.60**	**74.20**	**73.14**	**73.17**	**83.65**
		LGBM	71.07	71.52	71.74	70.64	80.95

The best-performing results for the critical metrics calculated by each algorithm are shown in bold.

We analyzed the performance of the ML classifiers (see Table 14) with the boosting ensemble using best arguments n_estimators = 100, max_depth = 3, and max_features = auto. With the train and test split strategy, GB displayed the best accuracy score (75.81%), whereas GB exhibited the highest accuracy (74.60%) with cross-validation. Conversely, the lowest-performing classifiers in accuracy included AdaBoost Classifier (66.13%) with train and test split and LGBM (71.07%) with cross-validation. As regards precision, GB proved to be the best-performing classifier in both train and test split and cross-validation with scores of 68.97% and 74.20%, respectively. The lowest-performing classifiers in precision were AdaBoost (57.58%) with train and test split and LGBM (71.52%) with cross-validation (see Figure 5). In conclusion, on the HCC Survival dataset using a boosting ensemble, GB exhibited the highest performance scores across the five metrics and with the two validation strategies: train and test split and cross-validation.



(**a**) (**b**)

Figure 5. ML classifier performance in accuracy (**a**) and precision (**b**): train test split vs. K-fold cross-validation—HCC Survival dataset.

4.2.4. Classifier Performance on ILPD Dataset

We tested the performance of the nine ML classifiers (AdaBoost, DT, ETT, GB, KNN, LGBM, LR, RG, and SVC) on the top four attributes of the ILPD dataset. The selected

attributes comprised Alkphos (score = 14.83), Sgot (score = 13.18), Sgpt (score = 12.86), and age (score = 12.83). Tables 15–17 summarize the results of the analysis.

Table 15. Non-ensemble learning performance analysis of classifiers on top four attributes—ILPD.

Ensemble	Technique	Predictive Model	Performance Evaluation Metrics				
			% Accuracy	% Precision	% Recall	% f1-Score	% roc_auc
Non-ensemble learning	Train and test split	AdaBoost	73.14	79.41	85.04	82.13	63.35
		DT	66.29	78.81	73.23	75.92	60.57
		ETT	69.71	76.43	84.25	80.15	57.75
		GB	68.57	76.09	82.68	79.25	56.96
		KNN	67.43	78.69	75.59	77.11	60.71
		LGBM	73.14	**79.85**	84.25	81.99	**64.00**
		LR	**74.29**	74.40	98.43	**84.75**	54.42
		RF	70.29	77.37	83.46	80.30	59.44
		SVC	72.57	72.57	**100.00**	84.11	50.00
	Cross-validation	AdaBoost	69.46	76.06	83.34	79.35	69.51
		DT	62.95	74.09	73.86	73.75	53.92
		ETT	68.95	76.02	82.61	78.85	71.34
		GB	69.46	76.01	84.11	79.54	71.56
		KNN	63.29	75.04	73.01	73.83	66.49
		LGBM	70.66	**77.69**	82.90	79.89	71.36
		LR	**71.51**	73.88	93.55	82.22	**72.52**
		RF	68.44	75.83	81.81	78.48	71.39
		SVC	71.35	71.35	**100.00**	**83.08**	61.78

The best-performing results for the critical metrics calculated by each algorithm are shown in bold.

Table 16. Bagging classifier performance using nine base learners on the ILPD (top four attributes).

Ensemble	Technique	Base Estimator	Performance Evaluation Metrics				
			% Accuracy	% Precision	% Recall	% f1-Score	% roc_auc
Bagging ensemble	Train and test split	AdaBoost	**76.00**	**78.15**	92.91	84.89	**62.08**
		DT	70.86	76.76	85.83	81.04	58.54
		ETT	71.43	75.16	90.55	82.14	55.69
		GB	71.43	75.84	88.98	81.88	56.99
		KNN	72.00	74.68	92.91	82.81	54.79
		LGBM	71.43	76.92	86.61	81.48	58.93
		LR	74.86	74.85	98.43	85.03	55.46
		RF	70.86	75.33	88.98	81.59	55.95
		SVC	72.57	72.57	**100.00**	84.11	50.00
	Cross-validation	AdaBoost	70.65	**76.32**	89.09	83.01	71.23
		DT	70.32	75.93	84.97	79.85	72.07
		ETT	70.15	74.56	88.62	81.15	73.62
		GB	71.69	75.69	87.93	81.74	73.31
		KNN	69.97	73.82	87.99	80.24	68.17
		LGBM	69.12	74.86	87.00	79.67	72.74
		LR	**72.03**	74.49	93.10	82.32	72.62
		RF	70.65	75.22	88.14	80.97	72.72
		SVC	71.35	71.35	**100.00**	**83.08**	69.71

The best-performing results for the critical metrics calculated by each algorithm are shown in bold.

As summarized in Table 15, with non-ensemble learning, LR displayed the best results in accuracy with both train and test split and cross-validation (74.29% and 71.51%, respectively). However, DT proved to be the lowest-performing classifier in accuracy with both train and test split and cross-validation (66.29% and 62.95%, respectively). As far as precision is concerned, LGBM proved to be the best-performing classifier, with scores of 79.85% in train and test split and 77.69% in cross-validation. The lowest-performing

classifier in precision was SVC with both train and test split and cross-validation (72.579% and 71.35%, respectively). In conclusion, on the ILPD with train and test split, LR exhibited good performance in accuracy and f1-score, LGBM outperformed the other classifiers in precision and roc_auc, and SVC exhibited the highest score in recall. With cross-validation, LR exhibited the highest score in accuracy and roc_auc, LGBM in precision, and SVC in f1-score and recall.

Table 17. Boosting ensemble learning performance analysis of classifiers on the ILPD (top four attributes).

Ensemble	Technique	Predictive Model	Performance Evaluation Metrics				
			% Accuracy	% Precision	% Recall	% f1-Score	% roc_auc
Boosting ensemble	Train and test split	AdaBoost	72.00	79.55	82.68	81.08	63.21
		GB	68.00	77.10	79.53	78.29	58.51
		LGBM	**73.14**	**79.85**	**84.25**	**81.99**	**64.00**
	Cross-validation	AdaBoost	68.59	75.69	82.28	78.64	68.07
		GB	67.57	**75.89**	79.94	77.64	68.97
		LGBM	**68.77**	75.79	**82.53**	**78.75**	**71.52**

The best-performing results for the critical metrics calculated by each algorithm are shown in bold.

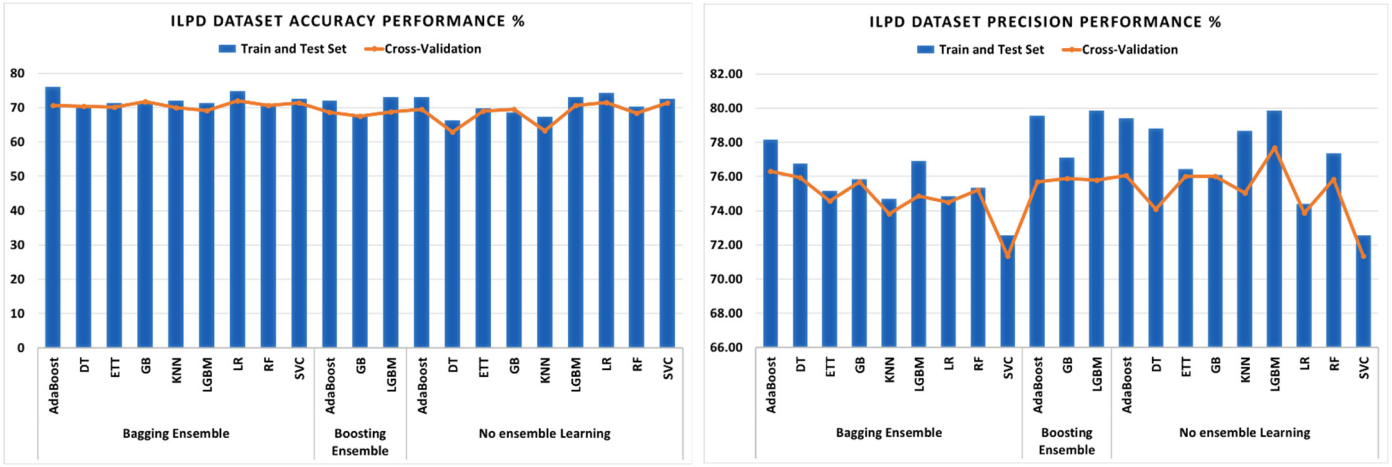
As summarized in Table 16, with the bagging ensemble classifier, we used number of trees = 100 and max_samples = 0.4. With the train and test split strategy, AdaBoost exhibited the best result in accuracy (76.00%), while LR exhibited the highest accuracy (72.03%) with cross-validation. Conversely, the lowest-performing classifiers in accuracy were DT (70.86%) with train and test split and LGBM (69.12%) with cross-validation. Regarding precision, AdaBoost proved to be the best-performing classifier with both strategies, train and test split and cross-validation, with scores of 78.15% and 76.32%, respectively. In precision, the lowest-performing classifiers were SVC (72.57%) with train and test split and 71.35% with cross-validation. In conclusion, on the ILPD, AdaBoost exhibited good performance in accuracy, precision, and roc_auc with train and test split. LR outperformed the other classifiers in f1-score and SVC in terms of recall. On the other hand, when using cross-validation, LR outperformed in accuracy, AdaBoost in precision, SVC in f1-score and recall, and ETT in roc_auc.

As observed in Table 17, we analyzed the performance of the classifiers with the boosting ensemble using best arguments for ILPD, i.e., n_estimators = 100, max_depth = 7, and max_features = auto. With both train and test split and cross-validation, LGBM exhibited the best result accuracy, with values of 73.14 and 68.77%, respectively. Conversely, GB was the lowest-performing classifier with respect to accuracy, with values of 68.00% during train and test split and 67.57% during cross-validation. As for precision, LGBM proved to be the best-performing classifier with the train and test split technique (79.85%) and GB with cross-validation (75.89%). The lowest-performing classifiers in precision included GB (77.10%) with train and test split and AdaBoost (75.69%) with cross-validation (see Figure 6). In conclusion, on the ILPD using a boosting ensemble, LGBM exhibited the best performance across the five metrics with train and test split. On the other hand, with cross-validation, LGBM outperformed in accuracy, recall, f1-score, and roc_auc, and GB exhibited the best results in precision.

4.3. Most Important Dataset Attributes

This research aims to identify the top four attributes or risk factors for the detection and prevention of liver disease by finding the best precision and accuracy results from the nine ML classifiers in different ensembles. We compared the results obtained from all the non-ensemble and ensemble (bagging and boosting) analyses and found out that the nine ML classifiers performed adequately on all the datasets with the two performance analysis strategies: train and test split and cross-validation (see Figure 7). Similarly, we noted that the bagging ensemble improved the accuracy of ET and AdaBoost on the BUPA and ILPD

datasets with train and test split. However, in three of the four datasets (BUPA, CPD, and HCC Survival Dataset), the algorithms performed much better with cross-validation.



(**a**) (**b**)

Figure 6. ML classifier performance in accuracy (**a**) and precision (**b**): train test split vs. K-fold cross-validation—ILPD.

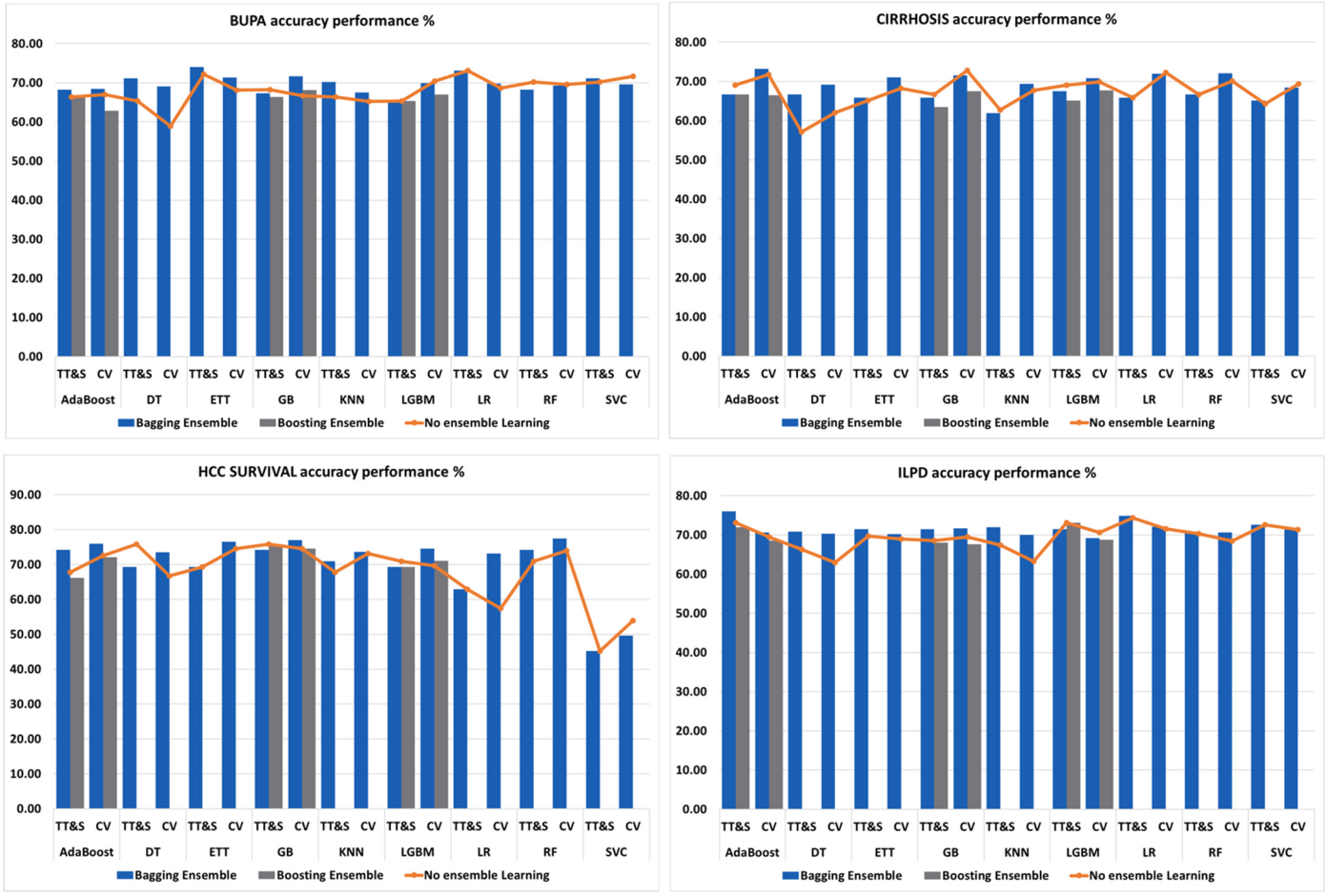


Figure 7. ML classifier accuracy performance top four attributes of liver disease datasets: train and test split vs. K-fold cross-validation.

Concerning the precision metrics, Figure 8 shows that the classifiers adequately performed for all the top four attribute classifications. Additionally, we found that when working with the BUPA and the HCC Survival datasets, the classifiers performed better in precision when using the bagging ensemble, whereas, in the ILPD, the classifiers performed much better with the boosting ensemble. On the other hand, when dealing with the non-ensemble method on the datasets, we achieved better classifier performance results on the ILPD and the HCC Survival dataset. As for the best-evaluated technique, train-and-test split validation worked best on the BUPA dataset, the CPD, and the ILPD. On the other hand, on the HCC Survival dataset, some algorithms performed better when using train-and-test split with the boosting ensemble, whereas some others worked best when using both the bagging ensemble and the non-ensemble models.

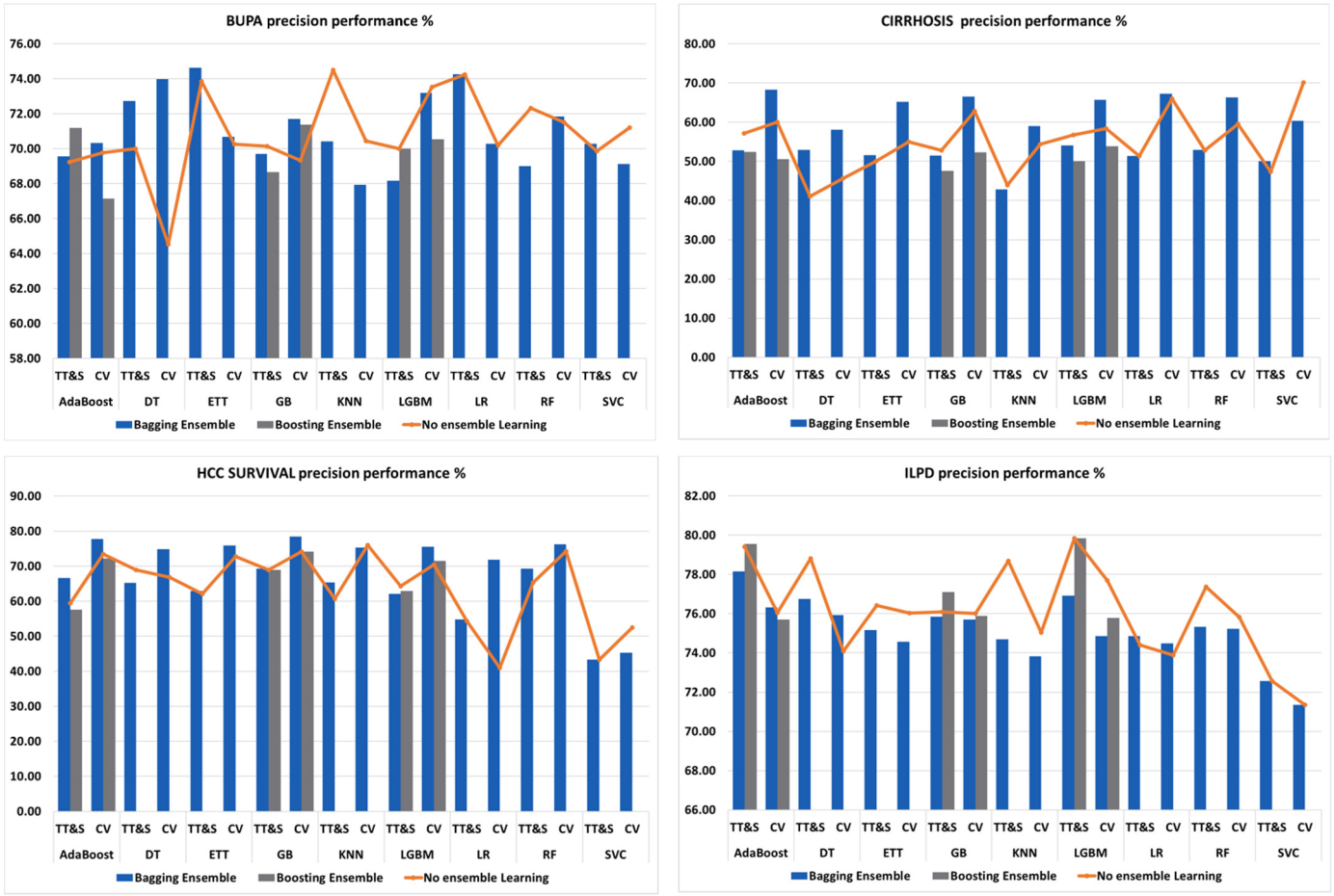


Figure 8. ML classifier precision performance on top four attributes of liver disease datasets: train and test split vs. K-fold cross-validation.

We identified the main attributes for diagnosing liver disease in the four datasets due to the previous analysis (Table 18). On the BUPA dataset, such attributes include gammagt (gamma-glutamyl transpeptidase), sgpt (alanine aminotransferase), sgot (aspartate aminotransferase), and alkphos (alkaline phosphatase). Regarding the best ensemble performance with train and test split, the bagging ensemble with ETT achieved the best results in accuracy (74.04%), precision (74.63%), f1-score (78.74%), and roc_auc with (72.35%). On the other hand, SVC yielded the best recall performance (86.67%). When using k-fold cross-validation, the bagging ensemble with GB exhibited the best performance in accuracy (71.61%) and roc_auc (75.84%), while DT outperformed the other classifiers in precision (73.99%). SVC with no ensemble displayed the best results in recall (84.87%) and f1-score (76.38%).

Table 18. Main risk factors identified in the four datasets for the diagnosis and prediction of liver disease.

Data	Best Rated Feature	Description
BUPA	Gammagt	Gamma-glutamyl transpeptidase
	Sgpt	Alanine aminotransferase
	Sgot	Aspartate aminotransferase
	alkphos	Alkaline phosphatase
HCC SURVIVAL DATASET	AFP	Alpha-Fetoprotein (ng/mL)
	Hemoglobin	Hemoglobin (g/dL)
	ALP	Alkaline phosphatase (U/L)
	Albumin	Albumin (mg/dL)
ILPD	Alkphos	Alkaline Phosphatase
	Sgot	Aspartate Aminotransferase
	Sgpt	Alanine Aminotransferase
	Age	Age of the patient
CPD	Prothrombin	prothrombin time in seconds [s]
	Albumin	albumin in [gm/dL]
	Platelets	platelets per cubic [mL/1000]
	Age	Age of the patient

In the HCC Survival dataset, the main attributes identified included AFP (alpha-fetoprotein), hemoglobin (hemoglobin), ALP (alkaline phosphatase), and albumin (albumin). With train and test split, DT without ensemble achieved the best performance in accuracy (75.81%), precision (68.97%), f1-score (72.33%), and roc_auc (75.96%), whereas SVC exhibited the best results in recall (100.00%). Regarding precision, RF and GB with the bagging ensemble yielded the best performance (69.23%). On the other hand, when using k-fold cross-validation with the bagging ensemble, RF exhibited the highest score in accuracy (77.45%), GB in precision (78.45%), roc-auc (84.98%), and AdaBoost in f1-score (78.59%). SVC exhibited the best performance with no ensemble in recall (99.23%).

In the ILPD dataset, the main attributes identified included Alkphos (alkaline phosphatase), Sgot (aspartate aminotransferase), Sgpt (alanine aminotransferase), and age (age of the patient). Regarding the best ensemble performance with train and test split, the bagging ensemble with AdaBoost achieved the best accuracy (76.00%), while LR displayed the best score in the f1-score (85.03%). With no ensemble, LGBM showed the best performance in precision (79.85%), SVC in recall (100.00%), and LGBM in roc_auc (64.00%). On the other hand, when using k-fold cross-validation with the bagging ensemble, LR showed the best results in accuracy (72.03%) and ETT in roc_auc (73.62%). With no ensemble, LGBM exhibited the best performance in precision (77.69%), while SVC outperformed the other classifiers in recall (100.00%) and f1-score (83.08%)

In the CPD, the main attributes identified included prothrombin (prothrombin time in seconds [s]), albumin (albumin in [gm/dl]), platelets (platelets per cubic [ml/1000]), and age (Age of the patient). The non-ensemble method achieved the best accuracy (69.05%) and precision (57.14%) with AdaBoost, whereas, in recall, DT outperformed the other classifiers (52.27%). Similarly, LGBM exhibited the best score in f1-score (51.85%) and roc_auc (64.11%). When using k-fold cross-validation, the bagging ensemble performed best with AdaBoost in accuracy (73.21%) and f1-score (55.73%) and with LR in roc_auc (75.99%). With no ensemble, SVC showed the best performance in precision (70.17%) and DT in recall (50.57%).

From our previous discussion, we concluded that the nine ML classifiers achieved adequate performance in the classification of the top four dataset attributes and can be successfully used for liver disease prediction.

The implemented algorithms validate the main features considered relevant for the diagnosis of liver diseases according to the literature, as shown in Table 18. Of the variables identified in the BUPA, HCC Survival, ILPD, CPD, alanine aminotransferase is strongly related to metabolic syndrome in NAFLD [57], alkaline phosphatase is an important serum

analyte that can be elevated in liver disease [58], and aspartate aminotransferase is a significant marker for alcoholic liver disease [6]. Testing of alanine aminotransferase [59] and aspartate aminotransferase is important in children with symptoms of possible liver disease, such as jaundice, dark urine, nausea, vomiting, or belly pain. Alkaline phosphatase in older female patients with isolated elevated alkaline phosphatase and risk factors for NAFLD should be evaluated for evidence of significant steatohepatitis [60]. Two other important risk factors are prothrombin [61] (identified in the CPD) and albumin [62] (identified in the HCC Survival dataset and the CPD). Prothrombin levels in liver disease are important indicators of liver function and pathology [63], and albumin binding function is a novel biomarker for early liver damage in NAFLD [64]. Both prothrombin and albumin must be primarily monitored when suspecting the presence of NAFLD. Platelets are another critical risk factor for liver disease [62,65]. Platelet counts can be performed to monitor or diagnose liver diseases or to look for the cause of too much bleeding or clotting. As for patient age, it was identified as a top risk factor for liver disease in the CPD and ILPD, which is consistent with the fact that cirrhosis can start at an early age [62]. In the BUPA dataset, gamma-glutamyl transpeptidase surfaced as a major risk factor for liver disease. Serum gamma GT activity is a valuable diagnostic tool for liver disease in children [66]. A GGT test is often used to diagnose liver disease and determine whether liver damage is due to liver disease or bone disease. The test is also used to check for blocked bile ducts and detect or monitor alcohol use disorder. When monitoring results are higher than normal, it may be a sign of liver damage caused by hepatitis, cirrhosis, alcohol use disorder, pancreatitis, diabetes, congestive heart failure, or a side effect of a medication. Finally, Alpha-fetoprotein (AFP) is a biomarker that can be used in the diagnosis and monitoring of liver diseases, particularly hepatocellular carcinoma (HCC) [67]. Accordingly, the algorithms implemented in the different ensembles identified the markers considered the main risk factors for liver disease in the datasets evaluated.

5. Conclusions and Future Directions

In this research, we aimed to compare the performance of nine machine learning algorithms (MLAs) across four datasets in predicting liver disease based on the top four attribute classifications using different ensembles with train-test split strategy and k-fold cross-validation methods. Our results indicated that alanine aminotransferase, aspartate aminotransferase, alkaline phosphatase, and albumin were identified as significant risk factors for liver diseases such as hepatitis, cirrhosis, alcohol use disorder, pancreatitis, diabetes, and mainly the aim of our study NAFLD. Additionally, gamma-glutamyl transpeptidase, hemoglobin, age, prothrombin, alpha-fetoprotein, and platelets contributed significantly toward detection. Our main findings revealed that the analyzed MLAs exhibited the best performance in the BUPA dataset and the CPD across the five performance metrics with non-ensemble learning. However, in the bagging ensemble, only ETT and LR exhibited high accuracy. The studied algorithms were classified appropriately to predict if a person were to have non-alcoholic fatty liver disease and exhibited good accuracy and precision across the four datasets. As for which classifier exhibited the highest accuracy with train and test split, AdaBoost outperformed in the ILPD using a bagging ensemble. As for cross-validation, the HCC Survival dataset obtained the best performance with RF using a bagging ensemble. The main contribution of this research is to validate the top risk factors for NAFLD: alanine aminotransferase, alkaline phosphatase, aspartate aminotransferase, alpha-fetoprotein, and gamma-glutamyl transpeptidase. Having proper medical follow-up on these attributes can contribute to the early diagnosis and treatment of non-alcoholic fatty liver disease. As for future proposals, research into medical databases for other common ailments, such as colon and breast cancer, is suggested. In addition, the risk factors detected in this research can be prioritized in mobile applications aimed at diagnosing and monitoring liver diseases. Finally, creating a database with the main attributes of liver diseases (risk factors) from various sources, such as clinical datasets, portable devices, mobile applications, and medical records, would be interesting. This goal could be achieved using big data methodologies

combined with machine learning, which will play a crucial role in improving our standard of living.

Author Contributions: Conceptualization, G.A.-H.; methodology, L.R.G.-N.; software L.R.G.-N. and J.L.S.-C.; validation, L.R.G.-N.; formal analysis, G.P.-A.; investigation, J.L.S.-C., L.R.G.-N. and G.A.-H.; data curation, L.R.G.-N.; visualization, G.P.-A.; supervision, G.A.-H.; project administration, J.L.S.-C.; funding acquisition, J.L.S.-C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Council for Scientific Research and Technological Development in Veracruz (COVEICYDET) through the project Prevention and Early Detection of Cardiovascular Diseases (Arrhythmias and Tachycardias) using Machine Learning Techniques, Big Data, and the Internet of Things, grant number 12 1806.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. These data can be found here: Liver Disorders Data Set (https://archive.ics.uci.edu/ml/datasets/liver+disorder) (accessed on 22 May 2022); HCC Survival Dataset (https://archive.ics.uci.edu/ml/datasets/HCC+Surviva) (accessed on 22 May 2022); ILPD Dataset (https://archive.ics.uci.edu/ml/datasets/ILPD+%28Indian+Liver+Patient+Dataset%29) (accessed on 22 May 2022); Cirrhosis Prediction Dataset (https://www.kaggle.com/datasets/fedesoriano/cirrhosis-prediction-dataset) (accessed on 22 May 2022).

Acknowledgments: L.G.N. and G.P.A. thank CONACYT for supporting their postdoctoral stays at Tecnológico Nacional de México. J.L.S.C. thanks CONACYT for the research position granted under the Cátedras-CONACYT program.

Conflicts of Interest: The authors declare no potential conflict of interest with respect to the publication of this research.

References

1. INEGI. INEGI Instituto Nacional de Estadística, Geografía e Informática. Características de las Defunciones Registradas en México Durante Enero a Agosto de 2020. *INEGI*. 28 June 2022. Available online: https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2021/EstSociodemo/DefuncionesRegistradas2020_Pnles.pdf (accessed on 27 June 2023).
2. Lee, H.W.; Sung, J.J.Y.; Ahn, S.H. Artificial intelligence in liver disease. *J. Gastroenterol. Hepatol.* **2021**, *36*, 539–542. [CrossRef] [PubMed]
3. Goldman, O.; Ben-Assuli, O.; Rogowski, O.; Zeltser, D.; Shapira, I.; Berliner, S.; Zelber-Sagi, S.; Shenhar-Tsarfaty, S. Non-alcoholic Fatty Liver and Liver Fibrosis Predictive Analytics: Risk Prediction and Machine Learning Techniques for Improved Preventive Medicine. *J. Med. Syst.* **2021**, *45*, 22. [CrossRef] [PubMed]
4. Kwak, M.S.; Kim, D. Non-alcoholic fatty liver disease and lifestyle modifications, focusing on physical activity. *Korean J. Intern. Med.* **2018**, *33*, 64–74. [CrossRef]
5. Ahmed, M.H. Biochemical Markers the Road Map for the Diagnosis of Nonalcoholic Fatty Liver Disease. *Am. J. Clin. Pathol.* **2007**, *127*, 20–22. [CrossRef]
6. Aravind, G.N.; Abhilash, K.; Syed, U.F. A study of alanine aminotransferase—Aspartate aminotransferase as a marker of advanced alcoholic liver disease. *Int. J. Adv. Med.* **2020**, *7*, 551–553. [CrossRef]
7. Pancreas, J.J.; Das, R.N.; Mukherjee, S.; Sharma, I. Alkaline Phosphatase Determinants of Liver Patients. 2018. Available online: http://pancreas.imedpub.com/ (accessed on 27 June 2023).
8. Lin, E.; Lin, C.H.; Lane, H.Y. Applying a bagging ensemble machine learning approach to predict functional outcome of schizophrenia with clinical symptoms and cognitive functions. *Sci. Rep.* **2021**, *11*, 6922. [CrossRef] [PubMed]
9. Ponnaganti, N.D.; Anitha, R. A Novel Ensemble Bagging Classification Method for Breast Cancer Classification Using Machine Learning Techniques. *Trait. Signal* **2022**, *39*, 229–237. [CrossRef]
10. Chicco, D.; Jurman, G. An ensemble learning approach for enhanced classification of patients with hepatitis and cirrhosis. *IEEE Access* **2021**, *9*, 24485–24498. [CrossRef]
11. Anisha, C.D.; Saranya, K.G. Early diagnosis of stroke disorder using homogenous logistic regression ensemble classifier. *Int. J. Nonlinear Anal. Appl.* **2021**, *12*, 1649–1654. [CrossRef]
12. Devi, M.S.; Swathi, P.; Upadhyay, S.S.; Sah, N.K.; Budhia, A.; Srivastava, S.; Rohella, M. Feature Predominance Ensemble Inquisition towards Liver Disease Prediction using Machine Learning. In Proceedings of the International Conference on Innovative Computing & Communication (ICICC), Delhi, India, 20–21 February 2021. [CrossRef]

13. Lin, E.; Lin, C.H.; Lane, H.Y. A bagging ensemble machine learning framework to predict overall cognitive function of schizophrenia patients with cognitive domains and tests. *Asian J. Psychiatr.* **2022**, *69*, 103008. [CrossRef]
14. Ejiofor, C.I.; Ochei, L.C. Application of Heterogenous Bagging Ensemble Model for predicting Breast Cancer. *J. Comput. Sci. Its Appl.* **2021**, *28*. [CrossRef]
15. Rahman, F.; Mahmood, M.A. A Dynamic Approach to Identify the Most Significant Biomarkers for Heart Disease Risk Prediction utilizing Machine Learning Techniques. Available online: https://www.researchgate.net/publication/357458668 (accessed on 28 April 2023).
16. Thomgkam, J.; Sukmak, V.; Klangnok, P. Application of Machine Learning Techniques to Predict Breast Cancer Survival. In *Lecture Notes in Computer Science, Proceedings of the 14th Multi-disciplinary International Conference on Artificial Intelligence (MIWAI 2021), Online, 2–3 July 2021*; Springer: Cham, Switzerland, 2021; Volume 12832, pp. 141–151. [CrossRef]
17. Yadav, S.; Singh, M.K. Hybrid Machine Learning Classifier and Ensemble Techniques to Detect Parkinson's Disease Patients. *SN Comput. Sci.* **2021**, *2*, 189. [CrossRef]
18. Buyrukoglu, S. Improvement of Machine Learning Models Performances based on Ensemble Learning for the detection of Alzheimer Disease. In Proceedings of the 6th International Conference on Computer Science and Engineering (UBMK), Ankara, Turkey, 15–17 September 2021; pp. 102–106. [CrossRef]
19. Singh, A.; Mehta, J.C.; Anand, D.; Nath, P.; Pandey, B.; Khamparia, A. An intelligent hybrid approach for hepatitis disease diagnosis: Combining enhanced k-means clustering and improved ensemble learning. *Expert Syst* **2021**, *38*, e12526. [CrossRef]
20. Sarvestany, S.S.; Kwong, J.C.; Azhie, A.; Dong, V.; Cerocchi, O.; Ali, A.F.; Karnam, R.S.; Kuriry, H.; Shengir, M.; Candido, E.; et al. Development and validation of an ensemble machine learning framework for detection of all-cause advanced hepatic fibrosis: A retrospective cohort study. *Lancet Digit Health* **2022**, *4*, e188–e199. [CrossRef] [PubMed]
21. Dutta, K.; Chandra, S.; Gourisaria, M.K. Early-Stage Detection of Liver Disease Through Machine Learning Algorithms. *Lect. Notes Netw. Syst.* **2022**, *318*, 155–166. [CrossRef]
22. Verma, A.; Mehta, S. A comparative study of ensemble learning methods for classification in bioinformatics. In Proceedings of the 7th International Conference on Cloud Computing, Data Science & Engineering—Confluence, Noida, India, 12–13 January 2017; pp. 155–158. [CrossRef]
23. Meng, L.; Treem, W.; Heap, G.; Chen, J. Predicting Clinical Outcomes of Alpha-1 Antitrypsin Deciency-Associated Liver Disease Using a Stacking Ensemble Machine Learning Model Based on UK Biobank Data. 2022; *preprint*. [CrossRef]
24. Al Telaq, B.H.; Hewahi, N. Prediction of Liver Disease using Machine Learning Models with PCA. In Proceedings of the 2021 International Conference on Data Analytics for Business and Industry (ICDABI), Sakheer, Bahrain, 25–26 October 2021; pp. 250–254. [CrossRef]
25. Gupta, S.; Gupta, M.K. Computational Prediction of Cervical Cancer Diagnosis Using Ensemble-Based Classification Algorithm. *Comput. J.* **2021**, *65*, 1527–1539. [CrossRef]
26. Pouriyeh, S.; Vahid, S.; Sannino, G.; De Pietro, G.; Arabnia, H.; Gutierrez, J. A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease. In Proceedings of the 2017 IEEE Symposium on Computers and Communications (ISCC), Heraklion, Greece, 3–6 July 2017; pp. 204–207. [CrossRef]
27. Kabir, M.F.; Ludwig, S.A. Enhancing the Performance of Classification Using Super Learning. *Data-Enabled Discov. Appl.* **2019**, *3*, 5. [CrossRef]
28. Doğaner, A.; Çolak, C.; Küçükdurmaz, F.; Ölmez, C. Prediction of Renal Cell Carcinoma Based on Ensemble Learning Methods. *Middle Black Sea J. Health Sci.* **2021**, *7*, 104–114. [CrossRef]
29. Hakim, M.A.; Jahan, N.; Zerin, Z.A.; Farha, A.B. Performance Evaluation and Comparison of Ensemble Based Bagging and Boosting Machine Learning Methods for Automated Early Prediction of Myocardial Infarction. In Proceedings of the 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 6–8 July 2021. [CrossRef]
30. Yadav, D.C.; Pal, S. An Experimental Study of Diversity of Diabetes Disease Features by Bagging and Boosting Ensemble Method with Rule Based Machine Learning Classifier Algorithms. *SN Comput. Sci.* **2021**, *2*, 50. [CrossRef]
31. Gao, X.Y.; Ali, A.A.; Hassan, H.S.; Anwar, E.M. Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method. *Complexity* **2021**, *2021*, 6663455. [CrossRef]
32. Taser, P.Y. Application of Bagging and Boosting Approaches Using Decision Tree-Based Algorithms in Diabetes Risk Prediction. *Proceedings* **2021**, *74*, 6. [CrossRef]
33. Murthy, H.S.N.; Manjunatha, M.N. Early Prognosis of Coronary Heart Disease using Ensemble Classifiers: A Comparative Analysis. *Volatiles Essent. Oils* **2021**, *8*, 2136–2142.
34. Fraiwan, L.; Hassanin, O. Computer-aided identification of degenerative neuromuscular diseases based on gait dynamics and ensemble decision tree classifiers. *PLoS ONE* **2021**, *16*, e0252380. [CrossRef]
35. Dhilsath, F.M.; Samuel, S.J. Hyperparameter Tuning of Ensemble Classifiers Using Grid Search and Random Search for Prediction of Heart Disease. *Comput. Intell. Healthc. Inform.* **2021**, 139–158. [CrossRef]
36. Khanam, F.; Mondal, M.R.H. Ensemble Machine Learning Algorithms for the Diagnosis of Cervical Cancer. In Proceedings of the 2021 International Conference on Science and Contemporary Technologies, ICSCT, Dhaka, Bangladesh, 5–7 August 2021. [CrossRef]

37. Bang, C.S.; Ahn, J.Y.; Kim, J.H.; Kim, Y.I.; Choi, I.J.; Shin, W.G. Establishing Machine Learning Models to Predict Curative Resection in Early Gastric Cancer with Undifferentiated Histology: Development and Usability Study. *J. Med. Internet Res.* **2021**, *23*, e25053. [CrossRef]
38. UCI Machine Learning Repository: Liver Disorders Data Set. Available online: https://archive.ics.uci.edu/ml/datasets/liver+disorders (accessed on 22 May 2023).
39. Santos, M.S.; Abreu, P.H.; García-Laencina, P.J.; Simão, A.; Carvalho, A. A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. *J. Biomed. Inf.* **2015**, *58*, 49–59. [CrossRef]
40. UCI Machine Learning Repository: HCC Survival Data Set. Available online: https://archive.ics.uci.edu/ml/datasets/HCC+Survival# (accessed on 22 May 2023).
41. UCI Machine Learning Repository: ILPD (Indian Liver Patient Dataset) Data Set. Available online: https://archive.ics.uci.edu/ml/datasets/ILPD+%28Indian+Liver+Patient+Dataset%29 (accessed on 22 May 2023).
42. Cirrhosis Prediction Dataset. Kaggle. Available online: https://www.kaggle.com/datasets/fedesoriano/cirrhosis-prediction-dataset (accessed on 22 May 2023).
43. Iyer, R.; Hosmer, D.W.; Lemeshow, S. Applied Logistic Regression. *J. R. Stat. Soc. Ser. D* **1991**, *40*, 458. [CrossRef]
44. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
45. Sharma, J.; Giri, C.; Granmo, O.-C.; Goodwin, M. Multi-layer intrusion detection system with ExtraTrees feature selection, extreme learning machine ensemble, and softmax aggregation. *EURASIP J. Inf. Secur.* **2019**, *2019*, 15. [CrossRef]
46. Lecun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
47. Sorich, M.J.; Miners, J.O.; McKinnon, R.A.; Winkler, D.A.; Burden, F.R.; Smith, P.A. Comparison of Linear and Nonlinear Classification Algorithms for the Prediction of Drug and Chemical Metabolism by Human UDP-Glucuronosyltransferase Isoforms. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2019–2024. [CrossRef] [PubMed]
48. Ramana, B.V.; Babu, M.S.P.; Venkateswarlu, N.B. A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis. *Int. J. Database Manag. Syst.* **2011**, *3*, 101–114. [CrossRef]
49. Biau, G.; Cadre, B.; Rouvière, L. Accelerated gradient boosting. *Mach. Learn.* **2019**, *108*, 971–992. [CrossRef]
50. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 2017. Available online: https://github.com/Microsoft/LightGBM (accessed on 31 May 2021).
51. Zhu, J.; Zou, H.; Rosset, S.; Hastie, T. Multi-class AdaBoost. *Stat. Its Interface* **2009**, *2*, 349–360.
52. Dietterich, T.G. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Mach. Learn.* **2000**, *40*, 139–157. [CrossRef]
53. Zhang, W.; Zeng, F.; Wu, X.; Zhang, X.; Jiang, R. A comparative study of ensemble learning approaches in the classification of breast cancer metastasis. In Proceedings of the 2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing, IJCBS, Shanghai, China, 3–5 August 2009; pp. 242–245. [CrossRef]
54. Guarneros-Nolasco, L.R.; Cruz-Ramos, N.A.; Alor-Hernández, G.; Rodríguez-Mazahua, L.; Sánchez-Cervantes, J.L. Identifying the main risk factors for cardiovascular diseases prediction using machine learning algorithms. *Mathematics* **2021**, *9*, 2537. [CrossRef]
55. Van Rossum, G.; Drake, F.L. *Python 3 Reference Manual*; CreateSpace: Scotts Valley, CA, USA, 2009.
56. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
57. Chen, Z.; Chen, L.; Dai, H.; Chen, J.; Fang, L. Relationship between alanine aminotransferase levels and metabolic syndrome in nonalcoholic fatty liver disease. *J. Zhejiang Univ. Sci. B* **2008**, *9*, 616–622. [CrossRef]
58. Grytczuk, A.; Gruszewska, E.; Panasiuk, A.; Cylwik, B.; Chrostek, L. Serum Profile of Lactate Dehydrogenase (LDH) and Alkaline Phosphatase (ALP) in Alcoholic Liver Diseases. 2020, *preprint*. [CrossRef]
59. Arsik, I.; Frediani, J.K.; Frezza, D.; Chen, W.; Ayer, T.; Keskinocak, P.; Jin, R.; Konomi, J.V.; Barlow, S.E.; Xanthakos, S.A.; et al. Alanine Aminotransferase as a Monitoring Biomarker in Children with Nonalcoholic Fatty Liver Disease: A Secondary Analysis Using TONIC Trial Data. *Children* **2018**, *5*, 64. [CrossRef] [PubMed]
60. Pantsari, M.W.; Harrison, S.A. Nonalcoholic fatty liver disease presenting with an isolated elevated alkaline phosphatase. *J. Clin. Gastroenterol.* **2006**, *40*, 633–635. [CrossRef] [PubMed]
61. Tripodi, A.; Caldwell, S.H.; Hoffman, M.; Trotter, J.F.; Sanyal, A.J. Review article: The prothrombin time test as a measure of bleeding risk and prognosis in liver disease. *Aliment Pharmacol. Ther.* **2007**, *26*, 141–148. [CrossRef] [PubMed]
62. Angulo, P.; Hui, J.M.; Marchesini, G.; Bugianesi, E.; George, J.; Farrell, G.C.; Enders, F.; Saksena, S.; Burt, A.D.; Bida, J.P.; et al. The NAFLD fibrosis score: A noninvasive system that identifies liver fibrosis in patients with NAFLD. *Hepatology* **2007**, *45*, 846–854. [CrossRef] [PubMed]
63. Stancu, G.; Iliescu, E.L. The Influence of Liver Transplant on Serum Cholinesterase Levels: A Case Report. *Cureus* **2023**, *15*, e33761. [CrossRef]
64. Sun, L.; Wang, Q.; Liu, M.; Xu, G.; Yin, H.; Wang, D.; Xie, F.; Jin, B.; Jin, Y.; Yang, H.; et al. Albumin binding function is a novel biomarker for early liver damage and disease progression in non-alcoholic fatty liver disease. *Endocrine* **2020**, *69*, 294–302. [CrossRef]
65. Enomoto, H.; Bando, Y.; Nakamura, H.; Nishiguchi, S.; Koga, M. Liver fibrosis markers of nonalcoholic steatohepatitis. *World J. Gastroenterol.* **2015**, *21*, 7427–7435. [CrossRef]

66. Maggiore, G.; Bernard, O.; Hadchouel, M.; Lemonnier, A.; Alagille, D. Diagnostic value of serum gamma-glutamyl transpeptidase activity in liver diseases in children. *J. Pediatr. Gastroenterol. Nutr.* **1991**, *12*, 21–26. [CrossRef]
67. Luo, X.; Cui, H.; Cai, L.; Zhu, W.; Yang, W.-C.; Patrick, M.; Zhu, S.; Huang, J.; Yao, X.; Yao, Y.; et al. Selection of a Clinical Lead TCR Targeting Alpha-Fetoprotein-Positive Liver Cancer Based on a Balance of Risk and Benefit. *Front. Immunol.* **2020**, *11*, 623. [CrossRef]