Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 19 No. 13 (2023) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v19i13.41871

PAPER

An Integrated Ensemble Learning Framework for Predicting Liver Disease

Soufiane Ardchir[1,2](✉), Youssef Ouassit[1], Soumaya Ounacer[1], Mohammed Yassine El Ghoumari[1,2], Mohamed Azzouazi[1]

[1]Faculty of Sciences Ben M'sik, Hassan II University, Casablanca, Morocco

[2]National School of Business and Management, Hassan II University, Casablanca, Morocco

s.ardchir@encgcasa.ma

ABSTRACT

The liver disease has become a pressing global issue, with a sharp increase in cases reported worldwide. Detecting liver disease can be difficult as it often has few noticeable symptoms, which means that by the time it is detected, it may have already progressed to an advanced stage, resulting in many people dying without even realizing they had it. Early detection is crucial as it enables patients to begin treatment earlier, which can potentially save their lives. This study aimed to assess the efficacy of five ensemble machine learning (ML) models, namely RF, XGBoost, Extra Trees, bagging, and stacking methods, in predicting liver disease. It uses the ILPD dataset. To prevent overfitting and biases in the dataset, several pre-processing statistical techniques were employed to handle missing data, outliers, and data balancing. The study's results underline the importance of using the RFE feature selection method, which allowed the use of only the most relevant features for the model, which may have improved the accuracy and efficiency of the model. The study found that the highest testing accuracy of 93% was achieved by the proposed model, which utilized an improved preprocessing approach and a stacking ensemble classifier with RFE feature selection. The use of ensemble ML has given promising results. Indeed, medical professionals can develop models better equipped to handle the complexity and variability of medical data, resulting in more accurate diagnoses, more effective treatment plans, and better patient outcomes.

KEYWORDS

liver disease, ensemble machine learning (ML), preprocessing, feature selection

1 INTRODUCTION

Liver disease is one of the most dangerous health problems, threatening hundreds of individuals around the world. It results in significant health complications, including liver failure and cancer, and can ultimately result in death if not properly managed. The liver plays a crucial role in the body's overall health and well-being, as it helps to filter out harmful toxins and waste products from the bloodstream.

Ardchir, S., Ouassit, Y., Ounacer, S., El Ghoumari, M.Y., Azzouazi, M. (2023). An Integrated Ensemble Learning Framework for Predicting Liver Disease. *International Journal of Online and Biomedical Engineering (iJOE)*, 19(13), pp. 138–152. https://doi.org/10.3991/ijoe.v19i13.41871

Article submitted 2023-05-31. Revision uploaded 2023-07-04. Final acceptance 2023-07-05.

When the liver is damaged or diseased, it can no longer perform this function efficiently, leading to a range of health problems [1].

There are several different types of liver disease, each with its own unique causes and symptoms. To start with, hepatitis, for example, is a viral infection that can cause liver inflammation, while fatty liver disease is primarily caused by the buildup of fat within liver cells. Cirrhosis is a more severe form of liver disease, characterized by the development of scar tissue within the liver, whereas liver fibrosis is a condition in which the liver becomes stiff and less flexible.

Early detection is crucial when it comes to treating liver disease. In many cases, symptoms may not manifest until the disease has progressed significantly, making it more difficult to treat. Regular check-ups and screenings can help identify liver disease early on, allowing for prompt treatment and better outcomes.

Recent advancements in artificial intelligence have also made it possible to improve liver disease diagnosis and treatment [2]. By analyzing medical data, ML algorithms can detect patterns and trends that could indicate the presence of liver disease. This innovative technology has the potential to assist doctors in making more precise diagnoses and designing highly effective treatment plans. Additionally, digital technologies such as smartphone apps and wearables can help patients monitor their liver health and track symptoms over time, enabling them to take a more proactive role in their care.

To minimize the risk of developing liver disease, individuals can take several preventive measures, including limiting alcohol intake, avoiding exposure to harmful chemicals and toxins, receiving appropriate vaccinations against hepatitis, and maintaining a healthy weight. Regular exercise and a healthy diet are also essential to maintain liver health and minimize the risk of disease development [3].

The remaining part of the paper falls into multiple sections. Section 2 provides a comprehensive literature review of liver disease prediction, outlining previous research in the field and identifying gaps in current knowledge. In Section 3, an exhaustive overview of the proposed framework is provided, with a particular emphasis on the specific methods and techniques employed.

Sections 4 and 5 describe the experimental setup and evaluation results. This section outlines the data used for the experiments and provides an overview of the evaluation metrics that will be used to evaluate the performance of the proposed framework. The results of the experiments are presented, and the strengths and weaknesses of the proposed approach are discussed.

Finally, in Section 6, the research is concluded by emphasizing its primary contributions. Additionally, this section addresses the limitations of the proposed approach and suggests promising avenues for future research.

2 LITERATURE REVIEW

Compared to the last few decades, there has been a noticeable increase in the occurrence of human diseases, and liver diseases, in particular, have increased with a growing number of affected individuals [4]. Nevertheless, in the initial stages of most liver diseases, symptoms may not be noticeable or may be very mild. Thanks to the large amounts of data generated and stored today, researchers have access to a wealth of information that can be used to solve problems in areas such as medical imaging, finance, genomics, and intrusion detection. Acquiring data and obtaining valuable information about liver diseases is indeed crucial for the diagnosis, treatment, and management of various liver conditions. However, it is not a simple task as the liver is a complex organ, and liver diseases can have diverse causes, symptoms, and outcomes [5].

The utilization of ML algorithms for disease prediction has become feasible because of the enhanced accessibility of concealed attributes within medical datasets.

Researchers employ various strategies to extract meaningful information from data sets. Some of these strategies involve the use of ML classifiers to select or extract features, while others do not. However, the presence of large volumes of unnecessary data can harm ML algorithms. To predict diseases or objects, several methods exist to select and extract the most correlated feature space.

In this section, we review existing methods for predicting liver disease and the different machine-learning models that are commonly used.

Bendi et al. [6] evaluated popular classification algorithms to evaluate their classification performance on two hepatic patient datasets (AP Liver and UCLA Liver). The study confirmed the good results of KNN classifiers, backpropagation, and SVM for the AP Liver dataset compared to UCLA datasets.

Amin et al. [7] have proposed a study aimed at classifying patients with liver disease based on the extraction of integrated features. The method begins with a pre-processing step to eliminate missing values and replace outliers, followed by the extraction of the features most significant for classification. In this respect, several categories of classification methods were considered. The proposed system achieved an average accuracy of 91.40% in the ensemble classification algorithm.

Kumar et al. [8] have created a rule-based liver disease prediction model using data mining techniques. They have introduced a novel approach called RBCM, which is used to forecast potential liver diseases. The study used a variety of statistical and machine-learning methods for the classification of liver diseases.

Srilatha et al. [9] suggested employing machine-learning methods to evaluate the overall liver wellness of patients in a comprehensive manner. The incidence rate of liver disease is considered a determinant of information. The percentages of the prediction results are presented in the confusion matrix, which showed high accuracy in predicting the test results.

Another study was proposed by E. Dritsas and M. Trigka [10], where multiple ML models and ensemble methods were analyzed and compared to predict the occurrence of liver disease based on their accuracy, precision, recall, F-measure, and area under the curve (AUC). After performing SMOTE with 10-fold cross-validation, the results of the experiment showed that the voting classifier outperformed the other models, achieving an accuracy of 80.1% as well as a precision of 80.4% and an AUC of 88.4%.

Shaker et al. [11] applied a logistic regression model to the ILPD dataset to anticipate the likelihood of liver disease emergence. The model demonstrated promising performance and, consequently, can serve as a valuable tool for tracking the advancement of liver disease.

M. Alghobiri et al., in a study referred to by [12], used various ML models to select important features and predict liver disease patients. Classification models are selected with great care, taking into account their global performance, and evaluated using a comprehensive training and test set. To enhance the evaluation of the models' practical performance, 10-fold cross-validation is utilized as a supporting test. The results indicate that the logistic regression and decision tree models are the best classifiers, achieving 72% and 71% accuracy, respectively, in cross-validation. Furthermore, although the Naive Bayes model does not perform well during training, it achieves 92% accuracy during the cross-validation phase.

Hassannataj Joloudari et al. [13] compared various predictive models of liver disease using the ELTA approach to select significant features. The study elaborated on the five most commonly used ML classification models, namely the Bayesian network, MLP-Neural Network, Random Forest, SVM, and also the PSO-SVM method. After the evaluation and estimation of the average accuracy, the PSO-SVM model achieved better accuracy compared to the other models.

3 MATERIALS AND METHODS

This section begins by introducing the dataset employed in this study. Following this introduction, we outline the architecture of our framework, the fundamental principles and stages of feature engineering, and examine the various classification models implemented in this analysis. Lastly, we provide a brief explanation of the feature selection techniques utilized in machine learning.

3.1 Data description

The data used in this study on liver disease was acquired from the UCI ML Repository [14], which is well-known in the scientific research and ML communities as a centralized location for accessing a wide range of resources. The ILPD dataset consists of 583 records, each with 11 features such as gender, age, DB, TB, Alkphos, Sgot, Sgpt, ALB, A/G ratio, TP, and the target variable. A comprehensive explanation of the features and their types can be found in Table 1. This dataset was partitioned into two separate groups. Group 1, consisting of 416 records, represented liver patients, while Group 2, consisting of 167 records, represented non-liver patients. Additionally, Figure 1 provides a visual representation of the number of patients in each category of the dataset. The histogram displays the frequency of patients classified as having liver disease or not.

Table 1. ILPD dataset features description

NO	Attribute Information	Feature Type	Missing Values	Domain	Measurement		
					Mean	Std. Error	Std. Deviation
1	Age (The patient's age)	Integer	No	(4–90)	44.75	0.67	16.19
2	Gender (The patient's gender)	Categorical	No	(Male–Female)			
3	TB (Total Bilirubin)	floats	No	(0.4–75)	3.3	0.26	74.6
4	DB (Direct Bilirubin)	floats	No	(0.1–19.7)	1.49	0.12	19.6
5	Alkphos (Alkaline Phosphatase)	Integer	No	(63–2110)	290.58	10.06	2047
6	SGPT (Alamine Aminotransferase)	Integer	No	(10–2000)	80.71	7.56	1990
7	SGOT (Aspartate Aminotransferase)	Integer	No	(10–4929)	109.91	11.97	4919
8	TP (Total Proteins)	floats	No	(2.7–9.6)	6.48	0.05	6.9
9	ALB (Albumin)	floats	No	(0.9–5.5)	3.14	0.03	4.6
10	A/G Ratio (Albumin and Globulin Ratio)	floats	4	(0.3–2.8)	0.95	0.01	2.5
11	Target (Disease/non-Disease)	Integer	No	(1,2)	1.29	0.02	1

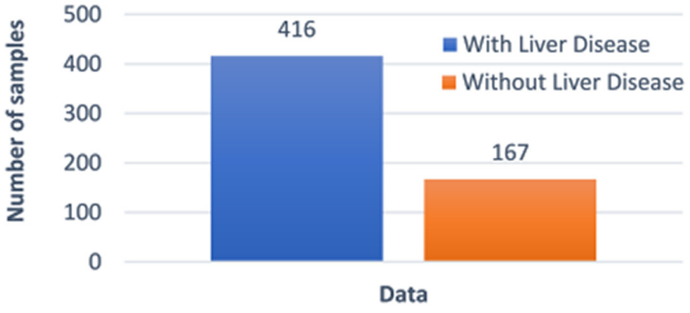

Fig. 1. Histogram of the frequency of liver patients

3.2 The proposed methodology

To better classify patients with liver disease, this study used a multi-step methodology. The first step was data preprocessing, which involved cleaning and transforming the original data to a suitable level for analysis. This process included the removal of missing values, the treatment of outliers, and the scaling of the data to ensure that all characteristics were on a comparable scale. In addition, the RFE feature selection method was employed to address the most important features that have a significant impact on the accuracy of the models employed.

As for the second step, the dataset is divided into two distinct subsets: the training set and the test set. The first sub-set is used to train our learning models, and subsequently, the models are tested on a set of data (the test set) that was not previously observed during the learning phase. In the third step, various ML models are trained to classify liver disease using the features available in the dataset. The fourth stage encompasses evaluating the trained model on the test set. This involves utilizing the trained model to predict the existence or lack of liver disease in the test set. Various metrics are employed to evaluate the performance of the model. The diagram depicting the process is presented in Figure 2.

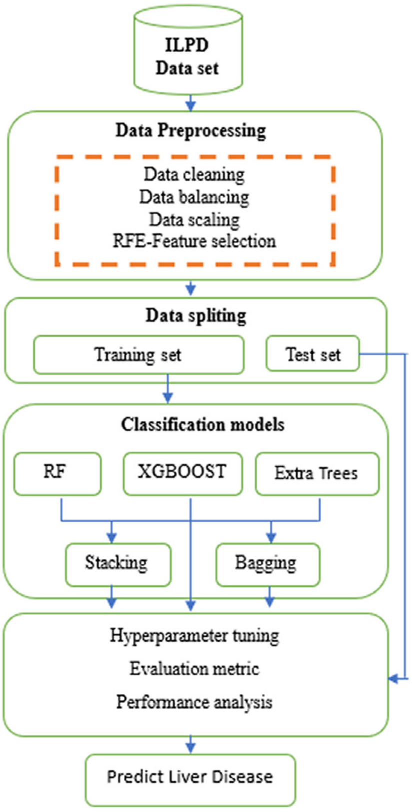

Fig. 2. Flowchart of the proposed model

3.3 Data preprocessing

To achieve accurate results from ML models, data preparation is a crucial step that cannot be overlooked [15]. When datasets are not properly handled, ML performance can suffer. One potential issue is a divergence between the model's performance during the training and testing phases. Factors such as data errors, noise, and omissions can all contribute to this problem. To prevent inconsistencies and ensure accuracy, it is important to preprocess the data by eliminating duplicates, anomalies, and other inconsistencies before comparing them.

Data encoding, a process that transforms categorical data into numerical values, is frequently necessary prior to instructing different methods. The Indian Liver dataset contains solely one categorical characteristic, namely gender. This feature contains two classes: female and male. To enable the use of this feature in various models, the gender column has been encoded in such a way that the female class is represented by 0 and the male class is represented by 1.

It is important to identify missing data in a dataset before applying a ML algorithm. Indeed, many of the latter are based on statistical methods that assume receiving a complete data set as input. Otherwise, the ML algorithm runs the risk of providing a poor predictive model. One way of solving this problem is either to apply various imputation techniques or to eliminate rows containing empty values. Imputation methods can be classified as univariate or multivariate. The univariate method works by utilizing merely the available non-missing values of that feature to estimate the missing values. On the other hand, multivariate imputation methods estimate missing values by utilizing all the features available in the dataset. In this study, the latter approach was used by employing a regression method to predict missing values. This method is considered to be one of the most efficient currently available.

Checking for skewness in data is important because it can considerably affect the accuracy and performance of ML models [15]. Skewed data can violate model assumptions, leading to suboptimal performance or biased predictions. Additionally, skewed data can affect the interpretation of feature importance, which can lead to incorrect conclusions about the significance of certain features in the model. To determine whether the data is skewed or not, it is possible to plot distribution curves. Figure 3a shows that the albumin_and_globulin_ration feature is slightly skewed; however, the other features shown from Figure 3b to Figure 3f are strongly skewed.

Various techniques exist for managing skewed data; in this work, we have used the "log" transformation technique. This transformation is effective in balancing the distribution of the curve; therefore, it is selected as the method of choice. The result of the new distributions of the different skewed features after the logarithmic transformation is displayed in Figure 4.

RobustScaler was applied to the ILPD dataset; it would involve the subtraction of the median from every feature value and subsequently dividing by the (IQR), which stands for the Inter-Quartile Range of that feature. This scaling technique would help to normalize the features and ensure that they have a similar range, even in the presence of outliers or skewed data.

The robust scaling equation is given in (1).

$$X_{new} = \frac{X - X_{mediane}}{IQR} \tag{1}$$

Imbalanced data is a frequently encountered issue in ML. The figures pertaining to positive instances are significantly smaller than the negative instances. The ILPD dataset used in this study comprised 167 records denoting individuals not afflicted with liver disease and 416 records corresponding to individuals diagnosed with the ailment. This indicates that the dataset is imbalanced, with a significantly smaller number of negative class observations. When the minority class is of interest, the imbalanced nature of data has the potential to significantly influence the performance and accuracy of ML models.

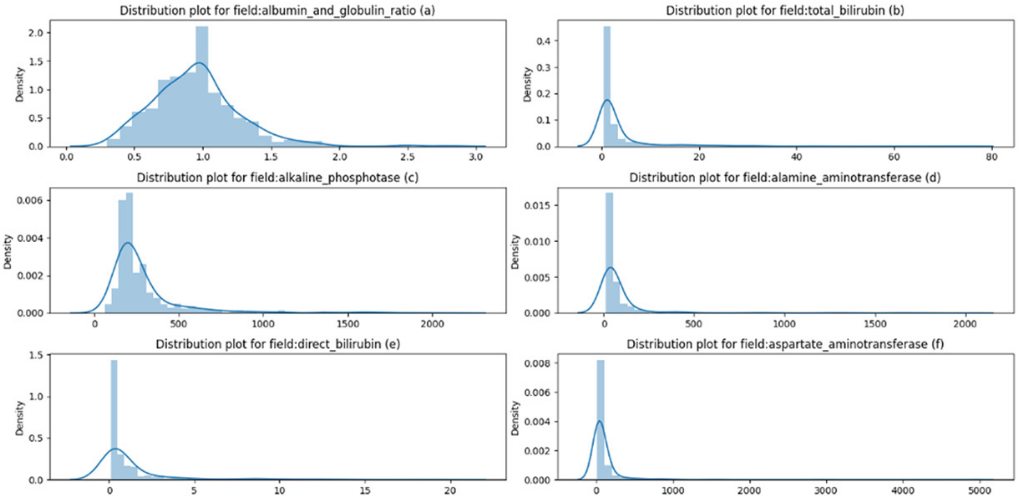

Fig. 3. Distribution plot for skewness features

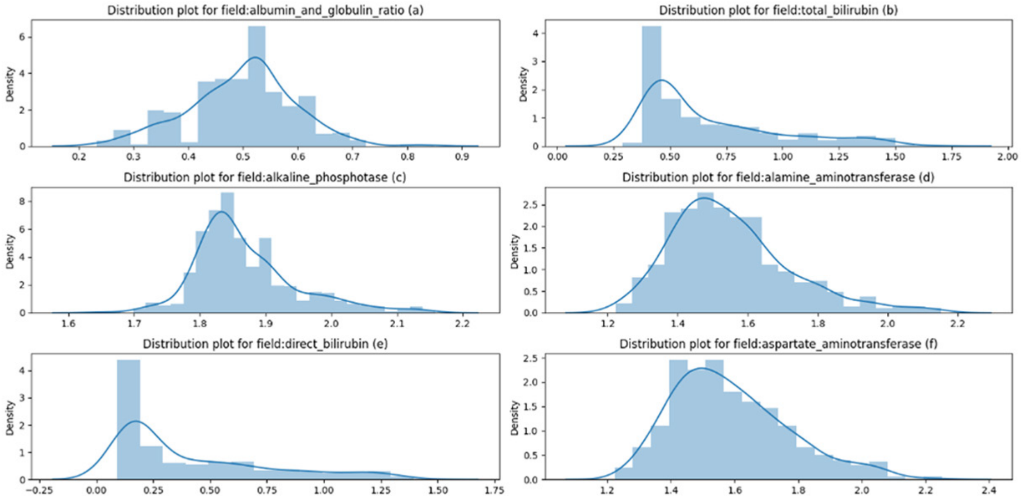

Fig. 4. Log1 p transformation of skewness features

Therefore, appropriate techniques should be applied to address the imbalance in the data and ensure reliable model performance [16].

In the case of the ILPD dataset characterized by an imbalanced class distribution, the application of the SMOTE technique can rectify the issue by oversampling the minority class (individuals without liver disease) and balancing instance counts in both classes. By generating synthetic samples of the minority class and implementing the SMOTE technique, ML models can more effectively capture the hidden patterns in the data and prevent bias towards the majority class.

After applying the SMOTE resampling technique, the total number of records increased to 832, with an equal number of instances in both classes. This balanced class distribution can potentially increase the performance and accuracy of ML algorithms, particularly in cases where the minority class is of interest.

3.4 Feature selection approach

Feature selection consists of identifying and selecting a smaller set of important features or variables from a larger pool of features in a dataset that are deemed to be most pertinent to the specific problem being investigated. In ML, feature selection is a crucial step in the preprocessing phase, as it helps to reduce the dimensionality of the data and can improve model accuracy and efficiency.

There are several methods for selecting features in ML [17], including:

- Filter methods: These methods employ statistical tests or other measures to rank the features based on their relevance to the target variable. Examples include correlation analysis and mutual information.
- Wrapper methods: These methods involve assessing the model's performance with different subsets of features, followed by selecting the subset that delivers the optimal performance. Examples include recursive feature elimination (RFE) and forward selection.
- Embedded methods: These methods select the features as part of the model training process. Examples include Lasso and Ridge regression.

The selection of a feature selection method is contingent on the particular problem and dataset at hand.

It is worth mentioning that feature selection is not always necessary, as some ML algorithms can handle high-dimensional data. However, when dealing with large datasets or limited computational resources, feature selection can be a useful tool for enhancing the model's performance and decreasing the computational cost.

In this study, the RFE method was chosen for feature selection on the Indian Liver Patient Dataset (ILPD). RFE is a feature selection method that operates in a backward fashion, wherein it recursively removes features and creates a model from the remaining ones [18]. It then ranks the features by their importance and eliminates the least important features. The process is repeated until the specified number of features is obtained.

By applying RFE to the ILPD dataset, the study aimed to select the most relevant features for the problem at hand and improve the performance of the ML method. The specific reasons for choosing RFE as the feature selection method may vary with the objectives of the study, the characteristics of the data, and the learning algorithm used. However, in general, RFE is a reliable method for feature selection as it considers the importance of each feature concerning the chosen model, which can help reduce the risk of overfitting.

4 ENSEMBLE MACHINE LEARNING

In the context of the ILPD dataset, ensemble-based ML algorithms may be used to improve model performance and handle the challenges posed by imbalanced and skewed data. Ensemble models typically involve using multiple base estimators or base learners to generate predictions, which are then combined to obtain a final prediction. Therefore, ensemble-based machine-learning algorithms can be a powerful tool for addressing the challenges posed by the ILPD dataset and enhancing the accuracy and performance of machine-learning models.

Some popular ensemble methods are [19]:

- Bagging: Training the same algorithm multiple times on several subsets of the training data and then averaging the predictions. This helps reduce variance. Examples are Random Forests (RF) and Decision Trees (DT).

- Boosting: Building models sequentially, where each model is trained on the instances that the previous models misclassified. This helps reduce bias. Examples are AdaBoost and XGBoost.
- Stacking: Training multiple first-level models and then using a second-level model to learn how to best combine the first-level models.

In this study, three popular supervised algorithms were identified and selected for the ensemble process, namely the RF, XGBoosting algorithm, and Extra-tree algorithm. The DT algorithm, which is commonly applied in supervised learning, is one of the selected algorithms.

The subsequent sections offer an extensive and comprehensive analysis of the algorithms employed in this research study.

4.1 XGBoosting classification algorithm

XGBoost (eXtreme Gradient Boosting) is an ensemble-based learning algorithm for boosting decision trees. It is designed to be efficient, flexible, and highly performant. The basic principle of the XGBoost algorithm is to combine several simple decision tree models to create a more complex and accurate model [20]. Each tree is trained on the examples that were misclassified by the previous trees to correct prediction errors. This boosting approach improves the model's accuracy by ensuring that each subsequent tree in the sequence corrects the errors of the previous ones.

4.2 Random forest classification algorithm

The Random Forest algorithm is an exceptional algorithm of learning that excels at supervised learning tasks. It is an ensemble learning method that involves building various decision trees on randomly selected subsets of the training data and then averaging their predictions [21]. The Random Forest algorithm has several advantages in comparison with other ML algorithms. First, it can handle high-dimensional datasets with many features. Second, it can avoid overfitting by building various decision trees on random training data subsets. Third, it can estimate the importance of each feature in the dataset, which can be useful for feature selection.

4.3 Extra tree classification algorithm

Extra Trees, also known as Extremely Randomized Trees, is a machine-learning algorithm that uses multiple decision trees. It constructs the final prediction by aggregating the predictions of models built on random subsets of the training data. This algorithm is used for various tasks, including classification and regression [21]. By introducing additional levels of randomness in the tree-building process, Extra Trees can be less prone to overfitting and more robust to noisy data.

4.4 Ensemble stacking algorithm

Stacking ensemble methods refer to the use of stacking in combination with other ensemble learning techniques to improve the predictive performance of ML models.

By combining these techniques with stacking, the base models can be trained on different subsets of the data, with the meta-learner using their predictions to generate the final prediction [22]. This approach can lead to even better predictive performance. Indeed, each technique addresses different sources of error and enhances the strengths of the other techniques. As demonstrated in Algorithm 1, in this particular study, ensemble models are used as the base models for stacking.

Algorithm 1: Ensemble Stacking Algorithm

```
Input: Training data
Output: Classification result (liver disease or no liver disease)
1. base models = {random forest, Boost, Extra Tree}
2. Initialize the base classifiers: RF, XGBoost, and Extra Tree
3. Train and use the base classifiers to predict liver disease on the test set.
4. Combine the base classifiers predictions using a meta-classifier (e.g., Logistic Regression) to create a
   new feature set.
5. Train the meta-classifier on the newly created feature set in step 4.
6. Use the trained meta-classifier to predict liver disease in the test set.
7. Evaluate the performance of the stacking ensemble algorithm on the test set.
```

4.5 Ensemble bagging algorithm

The ensemble Maximum of Bagging the (bootstrap aggregating) algorithm is a method of ensemble learning that associates the different ML methods predictions trained on different training subsets [22]. The algorithm works by creating multiple subsets of the training data by randomly sampling with replacement. Each subset is used to train a separate base model, such as a decision tree or random forest. In this study, the Bagging ensemble models are applied to the ILPD data set as shown in Algorithm 2.

Algorithm 2: Ensemble Bagging Algorithm

```
Input: Training data
Output: Classification result (liver disease or no liver disease)
1. base models = {random forest, Boost, Extra Tree}
2. For each base model
   – A random subset of features selected from the ILPD dataset.
   – A random selection of a subset of samples from the ILPD dataset with replacement.
   – Train the base model using the selected features and samples.
3. For each patient in the ILPD dataset:
   – Make a prediction using each of the base models.
   – Combine the predictions using a majority vote.
4. Calculate the accuracy of the Bagging algorithm using the combined predictions and the true labels
   from the ILPD dataset.
```

5 PERFORMANCE EVALUATION

The evaluation of performance is a critical stage in scientific studies. In ML, evaluation metrics are used to assess the effectiveness of a trained model on newly unseen data. These metrics are applied to determine how well the model can generalize from the training data to newly unseen examples [23].

Multiple evaluation metrics are utilized in ML, including recall, precision, F1-score, accuracy, specificity False Negative Rate (FNR), False Positive Rate (FPR), and AUC-ROC (Area under the Receiver Operating Characteristic Curve). Moreover, there is the ROC curve, which is a measure of the classifier's performance across all possible discrimination thresholds. Each of these metrics provides a different perspective on the performance of the model.

The metrics represented in equations (2)–(6) for calculating the various metrics are based on the following terms.

- True Positive (TP): means that the model accurately predicts positive values as positive.
- True Negative (TN): means that the model correctly predicts negative values as negative.
- False Positive (FP): means that the model mistakenly predicts negative values as positive.
- False Negative (FN): means that the model erroneously predicts positive values as negative.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Precision\ (P) = \frac{TP}{TP + FP} \tag{3}$$

$$Recall\ (R) = \frac{TP}{TP + FN} \tag{4}$$

$$F1 - score = \frac{2 \times (P \times R)}{P + R} \tag{5}$$

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

6 EXPERIMENTAL RESULT

The study is composed of two types of experiments to determine the most effective ML methods to predict liver disease based on the ILPD dataset. In the first experiment, five ensemble ML methods, including RF, XGBoost, Extra trees, bagging, and stacking method, are evaluated using the entire set of features in the dataset after a preprocessing step. To acquire the optimal feature parameters for the framework using ensemble learning, the Grid search technique is applied with 5-fold cross-validation to optimize the hyperparameters of each model.

In the second series of experiments, the RFE feature selection method chose the optimal features for the models. RFE is a feature selection technique that operates by iteratively removing features and constructing a model using the remaining features. Subsequently, the method assesses the relevance of each feature and ranks them in terms of importance while eliminating the least important features. This process is iterated until the optimal subset of the most important features is reached. The five most relevant characteristics are selected to conduct this experiment according to their importance, namely age, TB, Alkphos, Sgpt, and Sgot. The objectives behind using RFE in this study are to enhance the models' performance by identifying the critical features for predicting liver disease, reducing data volume without losing important and pertinent information, and saving processing costs.

The results of the evaluation of the classification models are based on the evaluation criteria presented in the previous paragraph and were obtained by using the confusion matrix. Table 2 presents the performance of different ensemble ML models, which enables a straightforward comparison of each model's performance in the two sets of experiments outlined earlier.

According to the study, the Extra Trees and RF models were effective in predicting liver disease, outperforming the baseline model when all variables were used, with accuracy rates of 85% and 83%, respectively. This suggests that ensemble ML models can provide stable and reliable predictions for this type of application. However, ML models are sensitive to noise caused by irrelevant or less important variables during the learning process. To tackle this issue, the study implemented a feature selection technique known as RFE. The choice of this method resulted in improved model performance and an increased accuracy rate of the Extra Trees model from 85% to 91%.

However, out of all the models considered, the stacking ensemble model attained the highest accuracy, reaching 93%, when utilizing the RFE feature selection technique. In addition, the stacking model was identified as the best-performing predictive model based on other criteria, such as sensitivity, accuracy, and F1 score. It achieved the highest values for all these measures, consolidating its position as the best model for predicting liver disease as shown in Table 2.

Additionally, the AUC criterion is a crucial measure used to assess the effectiveness of a classification model [24]. The ROC curve is a visual depiction that illustrates the association between the sensitivity (true positive rate) and specificity's complement (false positive rate) in a graphical form; whereas, the AUC measures the area under this curve. It is worth noting that a higher AUC value indicates better model performance. Once the ROC curves for the models are displayed, as depicted in Figure 5, we found that most ensemble ML models have a relatively good ROC-AUC value, which shows that ensemble methods are more suitable for this type of prediction. It is also apparent that the stacking ensemble model outperforms the other models, as demonstrated by its superior ROC-AUC value of 93%.

After comparing the performance of our framework to existing studies that used the same dataset and evaluation criteria, we found that the result of our model based on a stacking ensemble model classifier that combined RF, XGBoost, and Extra Trees algorithms, using RFE feature selection and a fine preprocessing step, showed the best results followed by the bagging ensemble model. The proposed framework demonstrated superior performance compared to various other research works, most of which utilized basic ML techniques. For instance, Amin et al. [7] attained an accuracy of 91.40% by the ensemble classification algorithm. Additionally, the authors of [9] obtained an accuracy of 73.07% using only RF. In contrast, in a recent study of [10] after having used SMOTE with 10-fold cross-validation, the researchers applied the Voting classification method which outperformed other methods, achieving an accuracy of 80.1%, recall of 80.1%, F-measure of 80.1%, precision of 80.4%, and an AUC of 88.4%. More importantly, the authors of [22] achieved an accuracy of 91.8% using all features. However, the proposed framework in this study surpassed most of these results, with the stacking ensemble model classifier using the RFE feature selection showing the best performance while achieving an AUC of 93% followed by the bagging ensemble model with an AUC of 91%.

Table 2. Performance evaluation of ensemble ML models

Models	Accuracy		Precision		Recall		F1-Score		Specificity	
	All	RFE	All	RFE	All	RFE	All	RFE	All	RFE
RF	0.83	0.88	0.79	0.81	0.91	0.96	0.85	0.88	0.83	0.89
XGBOOST	0.81	0.86	0.81	0.77	0.90	0.96	0.83	0.86	0.81	0.79
EXTRA TREE	0.85	0.90	0.82	0.88	0.89	0.89	0.85	0.89	0.88	0.90
STACKING	0.86	0.93	0.82	0.96	0.92	0.89	0.87	0.93	0.85	0.93
BAGGING	0.84	0.90	0.79	0.85	0.91	0.96	0.85	0.90	0.81	0.89

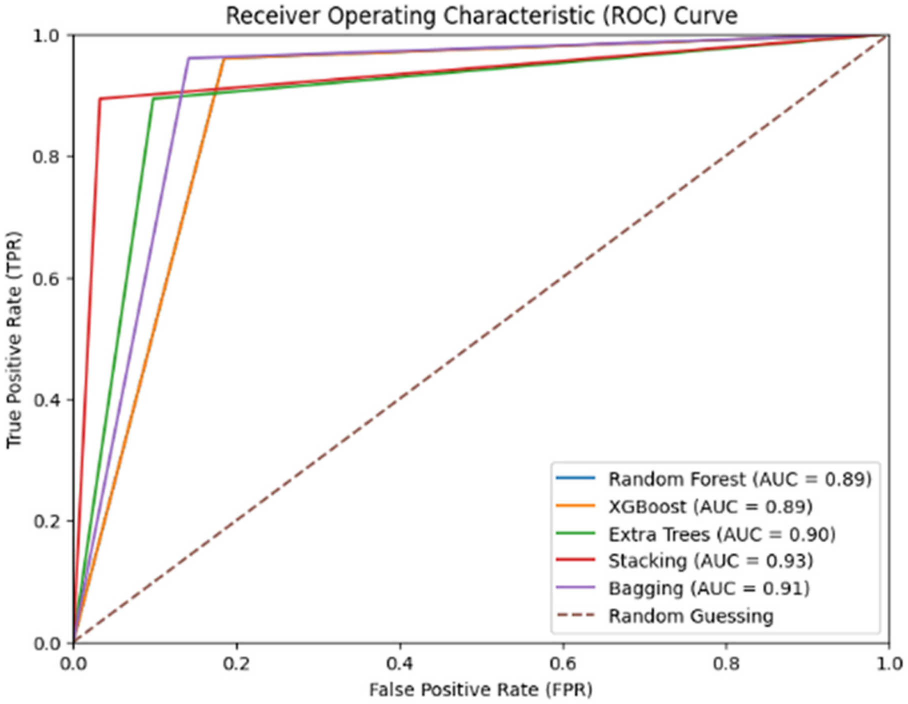

Fig. 5. ROC curve of models

7 CONCLUSION

Liver disease has been on the rise in populations around the world. Early diagnosis can save lives and help clinicians provide the right treatment. To deal with this issue, several ensemble learning models were tested, and their effectiveness was compared based on multiple criteria. This study investigated the performance of five ensemble learning models, including RF, XGBoost, Extra trees, the bagging method, and the stacking method, to predict liver disease based on the ILPD dataset. The ultimate objective of the current work is to solve the problems of bias and over fitting using various pre-processing techniques and significantly clean up and managing imbalanced data. The results of the study underline the importance of using the RFE feature selection method, which allowed the use of only the most relevant features for the model, which may have improved the accuracy and efficiency of the model. The findings demonstrated that the proposed model, which incorporated an improved pre-processing approach and a stacking classifier, achieved the highest testing accuracy of 93% in comparison with 90% for the Extra Tree model and the bagging method. Future research can explore ways to improve the proposed method. Increasing the volume of data by integrating different datasets for liver disease classification will certainly help to further enhance the model's accuracy, which will enable the use of more advanced feature selection techniques.

8 REFERENCES

[1] "World Hepatitis Summit 2022 statement," Available online: https://www.who.int/news/item/10-06-2022-world-hepatitis-summit-2022-statement

[2] D. Nam, J. Chapiro, V. Paradis, T. P. Seraphin, and J. N. Kather, "Artificial Intelligence in Liver Diseases: Improving Diagnostics, Prognostics and Response Prediction," *JHEP Reports*, vol. 100443, 2022. https://doi.org/10.1016/j.jhepr.2022.100443

[3] S. Cheemerla and M. Balakrishnan, "Global Epidemiology of Chronic Liver Disease," *Clin Liver Dis (Hoboken)*, vol. 17, no. 5, pp. 365–370, 2021. https://doi.org/10.1002/cld.1061

[4] L. G. Ambika, "Chronic Liver Disease Prediction Analysis Based on the Impact of Life Quality Attributes," *Int. J. Recent Technol. Eng.*, vol. 7, no. 7, 2023.

[5] K. Gupta, N. Jiwani, N. Afreen, and D. D, "Liver Disease Prediction using Machine Learning Classification Techniques," In *Proc. 2022 8th International Conference on Computer Science, Networks and Information Technology (ICCSNIT)*, pp. 221–226, 2022. https://doi.org/10.1109/CSNT54456.2022.9787574

[6] B. V. Ramana, M. S. Babu, and N. B. Venkateswarlu, "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis," *Int. J. Database Manage. Syst.*, vol. 3, pp. 101–114, 2011. https://doi.org/10.5121/ijdms.2011.3207

[7] R. Amin, R. Yasmin, S. Ruhi, M. H. Rahman, and M. S. Reza, "Prediction of Chronic Liver Disease Patients using Integrated Projection Based Statistical Feature Extraction with Machine Learning Algorithms," *Informatics in Medicine Unlocked*, vol. 36, p. 101155, 2023. https://doi.org/10.1016/j.imu.2022.101155

[8] Y. Kumar and G. Sahoo, "Prediction of different types of liver diseases using rule based classification model," *Technol. Health Care*, vol. 21, no. 5, pp. 417–432, 2013. https://doi.org/10.3233/THC-130742

[9] S. Tokala et al., "Liver Disease Prediction and Classification using Machine Learning Techniques," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 2, pp. 299–306, 2023. https://doi.org/10.14569/IJACSA.2023.0140299

[10] E. Dritsas and M. Trigka, "Supervised Machine Learning Models for Liver Disease Risk Prediction," *Computers*, vol. 12, no. 1, p. 19, 2023. https://doi.org/10.3390/computers12010019

[11] A. S. Abdalrada et al., "A Predictive model for liver disease progression based on logistic regression algorithm," *Periodicals of Engineering and Natural Sciences (PEN)*, 2019.

[12] M. Alghobiri, H. U. Khan, and A. Mahmood, "An Empirical Comparative Analysis Using Machine Learning Techniques for Liver Disease Prediction," *Int. J. Heal. Inf. Syst. Informatics*, vol. 16, pp. 1–12, 2021. https://doi.org/10.4018/IJHISI.20211001.oa10

[13] J. H. Joloudari et al., "Computer-Aided Decision-Making for Predicting Liver Disease using PSO-Based Optimized SVM with Feature Selection," *Informatics in Medicine Unlocked*, 2019. https://doi.org/10.1016/j.imu.2019.100255

[14] "UCI Machine Learning Repository: ILPD (Indian Liver Patient Dataset) Data Set," Uci.edu. Available online: https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)

[15] H. EL Hamdaoui, S. Boujraf, N. E. H. Chaoui, B. Alami, and M. Maaroufi, “Improving Heart Disease Prediction Using Random Forest and AdaBoost Algorithms”, *Int. J. Onl. Eng.*, vol. 17, no. 11, pp. 60–75, 2021. https://doi.org/10.3991/ijoe.v17i11.24781

[16] N. Liu et al., “A Novel Ensemble Learning Paradigm for Medical Diagnosis with Imbalanced Data,” *IEEE Access*, vol. 8, pp. 171263–171280, 2020. https://doi.org/10.1109/ACCESS.2020.3014362

[17] N. Chitra, S. Safinaz, and K. Bhanu Rekha, “Divergence Based Feature Selection for Pattern Recognizing of the Performance of Intrusion Detection in Mobile Communications Merged with the Computer Communication Networks,” *Int. J. Interact. Mob. Technol.*, vol. 17, no. 04, pp. 75–88, 2023. https://doi.org/10.3991/ijim.v17i04.37733

[18] N. Pudjihartono et al., “A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction,” *Front. Bioinform.*, vol. 2, p. 927312, 2022. https://doi.org/10.3389/fbinf.2022.927312

[19] M. Torabi et al., “A Review on Feature Selection and Ensemble Techniques for Intrusion Detection System,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 5, 2021. https://doi.org/10.14569/IJACSA.2021.0120566

[20] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016. https://doi.org/10.1145/2939672.2939785

[21] J. Yu, “Academic Performance Prediction Method of Online Education using Random Forest Algorithm and Artificial Intelligence Methods,” *Int. J. Emerg. Technol. Learn.*, vol. 16, no. 05, pp. 45–57, 2021. https://doi.org/10.3991/ijet.v16i05.20297

[22] A. Q. Md et al., “Enhanced Preprocessing Approach using Ensemble Machine Learning Algorithms for Detecting Liver Disease,” *Biomedicines*, vol. 11, no. 1, 2023. https://doi.org/10.3390/biomedicines11020581

[23] E. Valuations, “A Review on Evaluation Metrics for Data Classification Evaluations,” 2015.

[24] A. P. Bradley, “The use of the area under the ROC curve in the evaluation of machine learning algorithms,” *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997. https://doi.org/10.1016/S0031-3203(96)00142-2

9 AUTHORS

Soufiane Ardchir is a Professor of computer sciences at the National School of Business and Management, University Hassan II, Casablanca, Morocco. His area of expertise include big data and machine learning algorithms.

Youssef Ouassit is a Professor of preparatory classes for engineering schools. He completed PhD in computer science from Hassan II University, Casablanca, Morocco. His area of expertise include big data and machine learning algorithms.

Soumaya Ounacer is a Professor of computer engineering at the Hassan II University, Casablanca, Morocco. His area of expertise include big data and machine learning algorithms.

Mohammed Yassine El Ghoumari is a Professor of computer sciences at the National School of Business and Management, Hassan II University, Casablanca, Morocco. His area of expertise include big data and machine learning algorithms.

Mohamed Azzouazi is a professor of computer engineering at the Hassan II University, Casablanca, Morocco. His area of expertise is big data and machine learning algorithms. Dr. Azzouazi is a contributing author for more than 50 journal papers and conference papers.