Predicting Liver Cirrhosis Stages Using Extra Trees, Random Forest, and SVM with Data Mining Techniques

Duaa S. Ali, Maalim A. Aljabery
[1]Faculty of Computer Science & Information Technology, Computer Science Dept, University of Basrah, Basrah, Iraq
E-mail: duaasabeeh1@gmail.com and maalim.aljabery@uobasrah.edu.iq

Keywords: data mining, machine learning, extra trees, cirrhosis, feature selection

Received: July 21, 2024

Liver cirrhosis often occurs as a result of the lengthy and persistent progression of chronic liver disorders. It is a key crucial cause of death on a global scale. Early diagnosis and identification of cirrhosis are essential for preventing the disease's progression and the complete devastation of liver tissue. This paper aims to build an intelligent automated system that can predict the stages of cirrhosis employing Machine Learning (ML) algorithms, including Random Forest (RF), Extra Trees (ET), and Support Vector Machine (SVM). The dataset used in this research is sourced from the Zenodo website and linked to the GitHub website. This was our initial use of the data, which is publicly accessible and consists of 70 features and 10,000 records. In addition, data mining techniques were used to analyze the data before predicting the outcome. This involved data balancing due to the significant imbalance in the dataset's classes. To address this, we employed the Synthetic Minority Oversampling Technique (SMOTE) to mitigate a bias problem in a machine learning model. Then, feature selection techniques were applied, such as Chi-Square, Mutual Information (MI), and Recursive Feature Elimination and Cross-Validation (RFECV) based on classifiers RF and SVM (RF-RFECV, SVM-RFECV) to select relevant features. Lastly, the experimental findings showed that the Extra-Trees model with the Chi-square feature selection method (ET-Chi-Square) achieved the maximum level of accuracy of 93.87%. Additionally, it obtained recall, F1-score, and precision values of 94% each and an Area Under Curve (AUC) of 99%. Our method exhibited exceptional performance as compared to previous relevant research.

Povzetek: Razvit je nov pristop k napovedovanje stadijev jetrne ciroze z uporabo metod strojnega učenja, kot so Extra Trees, Random Forest in SVM, ter s tem izboljšuje medicinsko diagnostiko.

1 Introduction

Cirrhosis is a response to persistent liver injury. It is defined by the formation of regenerating nodules inside the liver tissue, which are encircled by fibrous bands [1]. The 11th most prevalent cause of mortality worldwide is liver cirrhosis. From 1990 to 2017, the incidence of deaths attributed to cirrhosis rose from 899,000 to 1,320,000, representing 2.4% of the total reported global mortality. Decompensated liver cirrhosis has a cumulative mortality rate of approximately 40% over one year [2]. The main etiologies of cirrhosis are Alcohol-Associated Liver Disease (ALD), autoimmune illnesses, Non-Alcoholic Fatty Liver Disease (NAFLD), and hepatitis B and C virus [1]. Symptoms may be mild in the early stages. The symptoms exacerbate as the liver's damage increases. In the initial phase, the most prevalent symptoms include fatigue, lethargy, nausea, weight loss, and loss of appetite. These results are expected to deteriorate in the future. Jaundice during this process, edema in the legs, rapid skin bruising, excessive itching, abdominal swelling, a propensity to hemorrhage, muscle atrophy, and Water retention in the body are all observed [3]. Liver biopsy is the established method for diagnosing cirrhosis. However, there are disadvantages to this intrusive method. A liver biopsy is a medical operation that involves the removal of a small piece of liver tissue using a local anesthetic. This tissue is then examined by a pathologist. Several factors affect the accuracy of cirrhosis staging. For illustration, the diagnostic precision of a biopsy specimen is affected by its length [1].

1.1 Data mining (DM)

Medical data has grown unparalleled over the years due to the exponential daily volume of transactions. Indeed, due to the vast volume of data generated, Data Mining (DM) has emerged to transform this data into practical and significant knowledge for hospitals [4]. Within the framework of Knowledge Discovery in a Database (KDD), Data Mining (DM) includes the processes of collation, transformation, preparation, mining of data, model interpretation or evaluation, and the application of the obtained information [5]. This study will employ data mining techniques to perform preprocessing for data, which is a crucial step in the DM process before using models to predict the stage of cirrhosis.

1.2 Machine learning (ML)

ML is a branch of AI that concentrates on accurately predicting changes in clinical conditions. Machine learning approaches are highly effective in the medical

field because they can extensively address a wide range of problems in medical data analysis. Various liver ailments have been accurately predicted through the application of ML [4][6]. A timely diagnosis of cirrhosis, facilitated by AI and ML, enables the implementation of medical interventions and lifestyle modifications. At the same time, this mitigates the likelihood of complications, including hepatic encephalopathy and variceal hemorrhage. The selection of appropriate treatment modalities is also facilitated, which reduces the necessity for advanced and expensive interventions such as liver transplantation [7].

This research is segmented into the subsequent sections: Section 2 comprises a collection of relevant studies and research. Detailed descriptions of the methodology implemented in the proposed system are provided in Section 3. Section 4 presents the details of an experimental findings account acquired from the suggested system. Section 5 provides an analysis and interpretation of the findings. Section 6 comprises the conclusions.

2 Related works

This section presents a selection of recent research focused on predicting liver cirrhosis using automated AI techniques. ***Table 1*** provides a concise overview of the studies related to the identification of liver cirrhosis.

In 2024, Sudiksha et al. [7] introduced LivMarX, a new model that seeks to stage liver cirrhosis by utilizing biomarkers rather than relying on photographs; it utilized advanced ML methods such as Gradient Boosting Classifier, CatBoost Classifier, Decision Tree Classifier (DT), and RF within an interdisciplinary framework. With an accuracy of 86%, LivMarX's RF Classifier was the most effective model.

In 2023, Greeshma et al. [1] created a new framework for noninvasive early detection of liver cirrhosis. An accuracy of 90.5% is achieved using the Extreme Gradient Boosting (XGboost) classifier. Has suggested a way to learn about the features that go into making predictions, which can aid doctors in making accurate diagnoses with the Explainable-AI algorithm.

In 2023, Oguzhan et al. [8] used the following classification algorithms to determine if the patient has liver cirrhosis: Multilayer Perceptron-Artificial Neural Networks, RF, NB, Logistic Regression (LR), K-Nearest Neighborhood (KNN), SVM, and DT. According to the findings, the DT algorithm achieves the utmost accuracy of (87.75%) compared to all other approaches.

In 2022, Anıl Utku [3] proposed a deep learning model utilizing multilayer perceptron (MLPs) to predict the probability of cirrhosis. The model's effectiveness was evaluated by comparing it to other approaches, including NB, KNN, LR, RF, SVM, and DT. The effectiveness of the generated model was examined. The suggested model exhibited better results compared to the methods in experimental studies. Based on the experimental results, the tested model achieved a precision, F1-score, recall of 85.71%, and an accuracy of 80.48%.

In 2022, Ke Chena et al. [9] they constructed a prognostic model utilizing ML to anticipate the happening of liver cirrhosis in persons diagnosed with Wilson disease (WD). Out of 346 patients with WD who were analyzed in this study, 246 were found to be free of liver cirrhosis. The model (XGBoost) exhibited exceptional accuracy, with a testing set accuracy of 0.76 and an AUC of 0.78.

Table 1: A summary of related works.

Ref.	Year	Classifiers for prediction of liver cirrhosis	Dataset Size	Key Findings	Accuracy
[7]	2024	Gradient Boosting, CatBoost, DT, RF	424 patients	Presented an approach for predicting the stage of liver cirrhosis by using blood tests. utilized several ML methods, and the RF model achieved a remarkable accuracy of 86%, surpassing comparable models.	86%
[1]	2023	Logistic regression (LR) and Extreme Gradient Boosting (XGBoost)	424 patients	Diagnostic biomarkers were employed to detect early-stage liver cirrhosis using the LR and XGBoost classifiers; the XGBoost achieved an accuracy of 90.5%.	90.5%
[8]	2023	MLP-ANN, SVM, NB, DT, RF, K-NN, LR	2000 patients	identified whether the patient has liver cirrhosis or not. The results showed that the DT algorithm outperforms all other methods used.	87.75%
[3]	2022	MLPs, KNN, NB, DT, LR RF, and SVM	418 patients	A deep learning model based on MLP has been created to forecast cirrhosis from blood testing. The MLP achieved an accuracy of 80.48%.	80.48%
[9]	2022	Extreme Gradient Boosting (XGboost)	346 patients	Using clinical information, analyzed 346 patients with WD, out of which 246 did not have cirrhosis; the XGboost model demonstrated superior accuracy in predicting cirrhosis in WD.	76%

3 Methodology

Figure 1 illustrates the principal steps of the proposed system for predicting the stages of cirrhosis. This research procedure includes dataset encoding, data scale, handling the imbalance of the dataset, features selection, splitting the dataset, and analyzing the classifier's performance.

3.1 Dataset description

Tests were administered on a publicly available dataset obtained from the Zenodo website, an open-access repository that supports research in many fields; a link is indicated on the GitHub website [10]. This study used the dataset for the first time to forecast cirrhosis; it comprises comprehensive clinical data intended for analyzing and forecasting health outcomes associated with liver diseases. Below, describe its characteristics and clinical relevance:

- **Demographics and sample size:** The dataset consists of 70 categorical and continuous features and 10,000 records with no missing values. It included 40.1% males and 59.9% females aged between 15 and 83 years.
- **Demographic variables:** Include age and gender. These variables are considered necessary for comprehending patient profiles.
- **Clinical symptoms:** Include features such as fatigue, itching, upper pain, spiders, bleeding, anorexia, nausea, edema, ascites, hepatomegaly, encephalopathy, etc. The dataset provides values for these variables (absent or present). These symptoms are indicators of the patient's health status and are essential for diagnosing and monitoring the progression of cirrhosis.
- **Laboratory results:** These results, including liver function tests that reflect liver damage, such as (Alanine aminotransferase (ALT), aspartate aminotransferase (AST), Gamma-glutamyl Transferase (GGT), etc.), provide a thorough assessment of the patient's liver health.
- **Clinical features other:** injections, transfusions, gallstones, choledocholithotomy, fibrosis, diabetes, blood pressure, Primary biliary cholangitis (PBC), obesity, steatosis, and Hepatitis. Their feature's relevance helps in developing complications of liver cirrhosis and its stages and treatment approach direction.
- **Target available:** is cirrhosis feature with three cases (absent, compensated, decompensated).

3.2 Data preprocessing

Data preparation involves converting raw data into a logical and understandable format. Data cleaning is an essential and obligatory process in preparing data for utilization in ML models and enhancing the precision of diagnostic findings. The data processing stages in our suggested system consist of the following steps:

3.2.1 Data encoding

Transforming categorical variables, which are represented as text strings, into numerical values is an essential step in facilitating ML models to calculate the correlation between them and generate precise predictions. This is because most ML models are designed to interpret numerical data rather than text [11].

A label coding technique was employed in this study to convert categorical data into numerical values. Where a distinct integer is allotted to every categorical value in the features. Our dataset contained 55 categorical attributes; the "Cirrhosis" feature, for example, is characterized by the presence of (absent, compensated, decompensated) variables. The encoding of the variables will be as follows (absent = 0, compensated = 1, decompensated = 2).

3.2.2 Feature selection

In the fields of pattern recognition and ML, feature selection is a critical data preprocessing utility. The minimally sized feature subset from the original set that is optimal for the objective is selected through the feature selection process [12]. Feature selection's advantages include enhanced data quality through an efficient data collection process, enhanced predictive performance, and reduced computational time required for the prediction model. [13]. In this study, we used two methods of feature selection: Filter and wrapper methods.

a) **Filter methods:** Select features based on feature ranking as the evaluation metric. For the most part, features are evaluated according to their scores within variety of statistical tests that evaluate their correlation with the class. Features that receive a score below a specific threshold are eliminated, while those that receive a score above it are preferred [14]. Without incorporating any learning classifier algorithm, it selects the feature based on its integral features. Compared to the wrapper technique, this method yields results more rapidly [15]. Chi-square is a widely recognized filter method.

- **Chi-Square:** A score of $\chi 2$ is allocated to every feature as well as the target of feature selection. "$\chi 2$ score" calculation is based on the reasoning that a characteristic with a low "$\chi 2$ score" is unrelated to a target class, suggesting that is not suitable for classifying samples of data [16]. It is mathematically represented using Equation 2.

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} \qquad (1)$$

Where E_i is the expected value, X^2 is Chi-Square, and O_i is the observed value.

- **Mutual Information (MI):** The MI classification concept is founded on entropy, a metric that quantifies the level of uncertainty in random data, yielding a numerical value ranging from 0 to 1. MI quantifies the extent of shared information between two random variables. This approach examines the features based on their interdependencies with the class and their duplication with each other [17]. It is mathematically represented using Equation 3.

$$I(F;T) = H(F) - H(F|T) \qquad (2)$$

Where I (F; T) is the MI for F and T, H(F) is the entropy of the F and T, and H (F | T) is the conditional entropy for F given T.

b) **Wrapper methods:** Irrespective of the ML technique used, use the performance of the chosen classifier algorithm as a metric to aid in selecting the best subset of features [14]. Variable subsets are evaluated and selected by employing the predictor's performance as the objective function, which is regarded as a black box. RFECV is a widely recognized form of wrapper method [18].

- **RFECV:** This method is a wrapper that selects features using an ML algorithm, ensuring that the most relevant attributes are selected. RFECV combines cross-validation and recursive feature elimination To determine the best features that improve model performance [19]. RFE is a procedure that evaluates the significance of features in a model through an iterative process. Remove the feature with the lowest significance in each iteration. The procedure entails the assessment of the validation errors of all potential feature subsets as well as the selection of therefore determining the best feature subset with the least error rate [20]. Algorithm 1 illustrates the RFECV feature selection steps:

Algorithm 1: RFECV for feature selection

Input: Dataset D

Output: feature subset

For each feature in D **do**
  For K-fold= 1 to 5 **do**
Cross-validation:
    D: Randomly subdivided into **5** subsets
    1 subset for validation data, and **4** for the training set;
    Train model (**RF** or **SVM**) using the training set;
    Compute the accuracy (ACC) using the testing set;
    Acquire the importance of each feature using the model;
    Updated the train set and removed/least weighted features;
  End for
  Acquire feature subset *(FS)* with the best ACC;
  If the *FS*-ACC is the best ACC, **then:**
    Features Selected = *FS*;
  End if
End for
Return Features Selected.

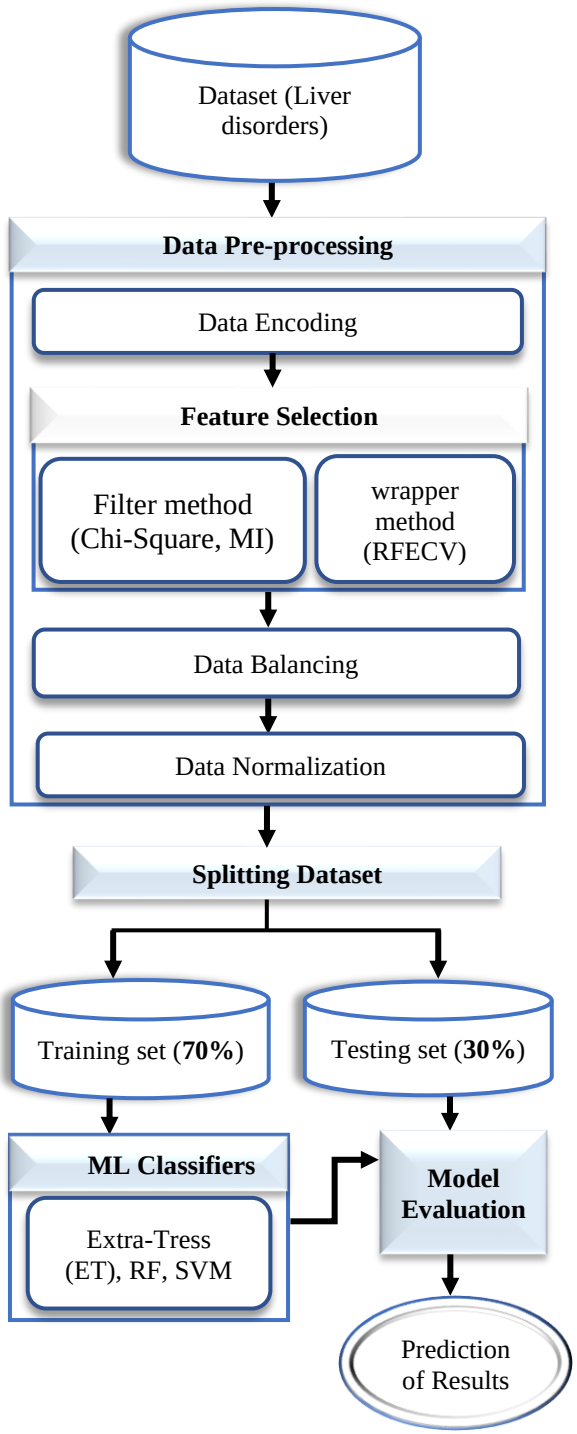

Figure 1: The proposed system for prediction of liver cirrhosis.

3.2.1 Balanced data

The dataset used in this paper for liver cirrhosis prediction is highly asymmetrical, which influences the result prediction. When there is a considerable disparity in the data between two classes, the classifiers will show a bias towards the class with the bigger amount of data. This phenomenon is commonly known as "imbalanced data" [21]. We employed SMOTE, a method that seeks to expand the samples of the smaller class by generating synthetic samples to prevent all overfitting problems. The presence of overfitting causes the model to exhibit

excellent performance on the training set but inadequate fit on the testing set [22]. The SMOTE steps are as follows [23]:

1. Preparation of quantity of synthetic minority class samples.
2. Randomly select a sample of a minority class.
3. Employs the KNN algorithm to obtain associate neighbors of the chosen sample [11].
4. Utilizes random interpolation to produce a new synthesis by combining minority and selected adjacent class instances.
5. The process of replicating steps 2 and 4 is continued until the desired quantity is achieved.

To guarantee the efficiency of SMOTE, it is essential to precisely adjust its setup by choosing the ideal value of **K**. In this work, after experimenting with several values, including **1,5, 7**, and **9**, we identified **5** as the final k-value to enhance the data properties of the randomly produced sample and minimize excessive computation, which may not yield improvements for the algorithm.

In this research, we utilized 1000 samples from the referred to above dataset because the dominant category in these data is the absence of cirrhosis in a large percentage, distributed throughout each stage from class as follows: absent (200), compensated (238), and decompensated (562). Subsequently, the entries underwent the Oversampling technique, leading to the data being balanced, thus, an equal distribution of samples across each stage: absent (562), compensated (562), and decompensated (562). ***Figure 2 (a, b)*** shows the count plot before and after applying the SMOTE technique. ***Table 2*** displays the total number of samples for each stage of cirrhosis in the dataset before and after applying the SMOTE approach.

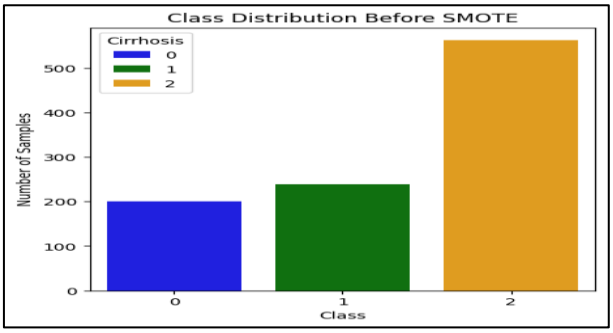

Figure 2-a: Before using the SMOTE Count Plot.

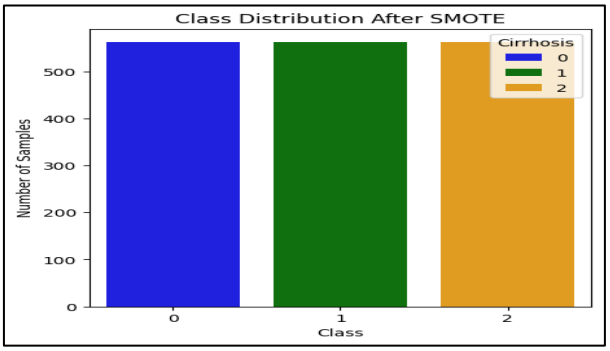

Figure 2-b: After using the SMOTE Count Plot.

Table 2: Details the number of class samples before and after data balancing.

Before SMOTE		After SMOTE	
Class	Sample count	Class	Sample count
absent	200	absent	562
compensate	238	compensate	562
decompensate	562	decompensate	562
Total	1000	Total	1686

3.2.2 Data scaling

Scaling, also known as normalization, ensures that features with higher numerical ranges do not supersede those that vary within a lower range. Thus, dataset scaling can alleviate this phenomenon and, as a result, enhance classification performance [24]. In this study, data scaling was implemented within the standard-scaler function. Consequently, it is a frequently employed data transformation technique in the field of ML to standardize the scale of numerical variables or features. The data is scaled using the standard deviation and the mean. The resulting features represent the distribution with zero mean and unit variance. The standard-scaler was mathematically represented using Equation 1.

$$Z = \frac{X - \mu}{\sigma} \qquad (3)$$

Where $\boldsymbol{x}$ is an original value, $\boldsymbol{\sigma}$ is a standard deviation, $\boldsymbol{\mu}$ is the mean.

3.2.3 Splitting dataset

The dataset undergoes division into a pair of distinct sets; training set and test set. Employing a training set, classifiers are trained, which consists of 70% of the data (1180 samples). The model's accuracy and effectiveness on unseen data are explicitly assessed using the testing dataset, which comprises 506 samples (30% of the data).

3.3 Classification models

The classifiers ETs, SVM, and RF were selected and implemented after the preceding phases. Classification is an essential process of supervised learning in which classifiers apply their knowledge to the testing dataset to identify the target attribute by learning from the training dataset. The classification techniques employed in this research are detailed below.

3.3.1 Extra-Trees (ET)

ET is an ensemble learning method that is based on DTs. ET employs a random selection process to choose particular judgments and subsets of data to avoid excessive learning and fitting. Multitudes of trees are constructed, and the nodes are divided into groups according to randomly selected subsets of features. Data bootstrapping is not employed to introduce randomization into Extra-Trees; rather, all observations are randomly divided [25].

3.3.2 Support vector machine (SVM)

SVM is the potent model of the realm of supervised ML methods and is particularly effective in addressing binary classification and multiclass classification issues [26]. SVM is a method for identifying a hyperplane in a space with N dimensions. Data points could be categorized into distinct categories or predicted as continuous values using a variety of hyperplanes. Hyperplanes serve as decision boundaries, enabling the classification of data points. Furthermore, by expanding the margin distance, it is possible to enhance the accuracy of classifying future data points [27].

3.3.3 Random forest

RF refers to the compilation of DTs randomly selected from the input feature set. It acquires input data, trains numerous models, accumulates each model's prediction, and subsequently implements a voting mechanism to choose the optimal solution [28]. In clinical data analysis, the RF achieves the balance between robustness and accuracy. The capacity to manage intricate interactions and nonlinear relationships among variables can improve the diagnostic and prognostic accuracy of diseases [29].

3.4 Performance metrics

To assess the models' efficacy in this study, numerous metrics were implemented, including F1-score, precision, confusion matrix, recall, accuracy, and AUC, as illustrated in the following details:

- **Confusion matrix:** Was employed to assess the effectiveness of ML models. It consists of the classification keys, which are illustrated below:
  - **TP** (**True positive):** Refers to an output classed as positive and is correctly predicted.
  - **TN (True negative):** This is a negative output, ensuring the predicted result is correctly classified.
  - **FP (False positive):** signifies the output is positive, resulting in an incorrect classification of the predicted result
  - **FN (False negative):** signifies the output is negative, resulting in an incorrect classification of the predicted result.

Table 3 displays the description of this method.

- **Accuracy:** the proportion of right predictions out of all forecasts made. Equation 3 is used to represent it mathematically.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \quad (4)$$

- **Recall:** It indicates the "true positive rate (TPR)": It is calculated by dividing correctly predicted positive values by the entire of positive values. Equation 4 is used to represent it mathematically.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (5)$$

- **Precision:** It determines the proportion of the correctly detected samples to all detected samples. Equation 4 is used to represent it mathematically.

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \quad (6)$$

- **F1-score:** signifies the equilibrium between precision and recall. Mathematically, it is represented using Equation 6.

$$\text{F1-score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recal}} \quad (7)$$

- **AUC:** Utilizing the AUC of the ROC curve (Receiver Operating Characteristic). An ideal classifier achieves a maximum value of AUC which is 1, while a random classifier achieves a value of AUC of 0.5. A greater AUC indicates an enhanced ability to distinguish between excellent and negative occurrences [30].

Table 3: The general layout of a confusion matrix for a classification with three classes.

Confusion matrix		Actual class		
		Stage 1	**Stage 2**	**Stage 3**
Predicted class	**Stage 1**	TP	FP	FP
	Stage 2	FN	TN	TN
	Stage 3	FN	TN	TN

4 Experimental results

This part will present the findings for all the tests we conducted using RF, SVM, and ET classification models with feature selection techniques such as Chi-Square, MI, and RFECV based on (RF and SVM). Furthermore, the models' efficacy and effectiveness on a testing set were evaluated by employing a variety of metrics, such as recall, confusion matrix, precision, F1-score, and accuracy. Additionally, the ROC plots in our research were chosen based on the models' greatest AUC scores. To evaluate the efficacy of the models in the proposed study, we implemented a sequence of experiments. These experiments are referred to as Experiment I and Experiment II.

4.1 Experiment I

In this experiment, we processed the data so that it was appropriate for the models to make predictions. Before data balancing, we trained the models using the original dataset, which consisted of 1000 samples with all features (70). The ET classifier had superior performance compared to the other classifiers, as evidenced by its accuracy of 84%, precision of 78%, recall of 87.66%, and F1-score of 79.66%. Due to the asymmetrical nature of our dataset, there is a notable imbalance in the number of stages. Specifically, there are far more instances of the "absent" stage compared to the "compensate" and "decompensate" stages. To address this problem, we applied the SMOTE technique to equalize the dataset and then divided it into two separate datasets. The first category consists of (1180) training samples, representing 70% of the total, and the test set, which consists of 506 samples, accounts for 30% of the overall dataset, which contains 1686 samples of the total samples. The result of the train–test split method on the balanced dataset with all features (70), the ET model obtained the best findings in

the accuracy of (88.54%), F1-score, recall, and precision of (88.33%) respectively. The outcomes of the algorithms before and following the implementation of the SMOTE approach are illustrated in ***Table 4***.

Table 4: Performances of ML models on imbalanced and balanced data with all features.

Before SMOTE with all features			
Evaluation Metrics	**RF**	**SVM**	**ET**
Accuracy	0.83	0.80	0.84
Precision	0.88	0.79	0.78
F1-score	0.78	0.78	79.66%
Recall	0.76	0.78	87.66%
After SMOTE with all features			
Evaluation Metrics	**RF**	**SVM**	**ET**
Accuracy	0.87	0.83	88.54%
Precision	0.87	82.33%	88.33%
F1-score	86.33%	82.66%	88.33%
Recall	86.66%	82.66%	88.33%

4.2 Experiment II

In this experiment, we employed feature selection techniques in a sequence of experiments conducted independently of one another; the sequence is as follows:

First, we implemented the Chi-Square method, in which we conducted numerous tests to determine the optimal number of pertinent and critical features for the target. 40 features were chosen due to their superior performance with the proposed models. Compared to other experiments in this study, the ET model yielded the highest precision, F1-score of (0.94), and recall and accuracy of (93.87%).

Secondly, in the series of experiments we conducted, we used the RFECV feature selection technique based on the RF classifier (RF-RFECV). The technique automatically selected 38 features as the best features. The results indicated that the RF was characterized by superior accuracy (90.32%), precision, F1-score (89.66%), and recall (90.33%) compared to other models (SVM, ET).

Third, in this experience, we employed the same procedure (RFECV), as previously mentioned to select features, but we relied on the SVM classifier. This technique selected a reduced number of features, in comparison to RF-RFECV, specifically 15 features. The RF model achieved the highest outcomes in the terms of recall, accuracy (91%), F1 score, and precision (90.66%).

Fourth and last, MI is used to evaluate the dependency between the target variable and each feature. This method identified the top 39 features that provide the most information about the target variable. However, as shown in Table 5, MI did not enhance the total results of our models in the selection process.

Table 5 illustrates the outcomes of the models after using feature selection techniques.

Table 5: Performances of ML models on balanced data with feature selection methods.

Feature Methods	Number of Features	Evaluation Metrics	Classifiers		
			RF	**SVM**	**ET**
Chi-Square	40	**Accuracy**	93.08%	91.50%	**93.87%**
		Precision	0.93	91.33%	**0.94**
		F1-score	0.93	0.91	**0.94**
		Recall	0.93	91.33%	**0.94**
RF-RFECV	38		**RF**	**SVM**	**ET**
		Accuracy	**90.32%**	85.17%	90.12%
		Precision	**89.66%**	84.33%	89.33%
		F1-score	**89.66%**	84.33%	89.33%
		Recall	**90.33%**	84.66%	0.90
SVM-RFECV	15		**RF**	**SVM**	**ET**
		Accuracy	**0.91**	89.13%	90.32%
		Precision	**90.66%**	0.89	89.66%
		F1-score	**90.66%**	0.89	0.90
		Recall	**0.91**	89.33%	89.66%
Mutual Information (MI)	39		**RF**	**SVM**	**ET**
		Accuracy	84.78%	79.24%	84.78%
		Precision	84.33%	0.79	84.66%
		F1-score	84.33%	78.33%	84.66%
		Recall	84.66%	78.66%	84.66%

Figure 3(a, b, c, and d) displays the ROC curves of the algorithms utilizing feature selection techniques. The (ET+ Chi-Square) model achieved 0.99 highest AUC by comparison with all experiments, as shown in ***Figure 3 (a)***. Besides, when we used the (RF-RFECV) method, the (ET+RF-RFECV) and (RF+RF-RFECV) achieved 0.97 AUC, as ***Figure 3 (b)*** illustrates. The (RF+SVM-RFECV) achieved 0.98 AUC, as ***Figure 3 (c)*** shows using the (SVM-RFECV) method, in which the (ET+SVM-RFECV) achieved 0.98 AUC.

Figure 4(a, b, c,d) displays the performance results of the models (ET, RF, SVM) with feature selection methods (Chi-Square, RFECV, MI) based on the confusion matrix.

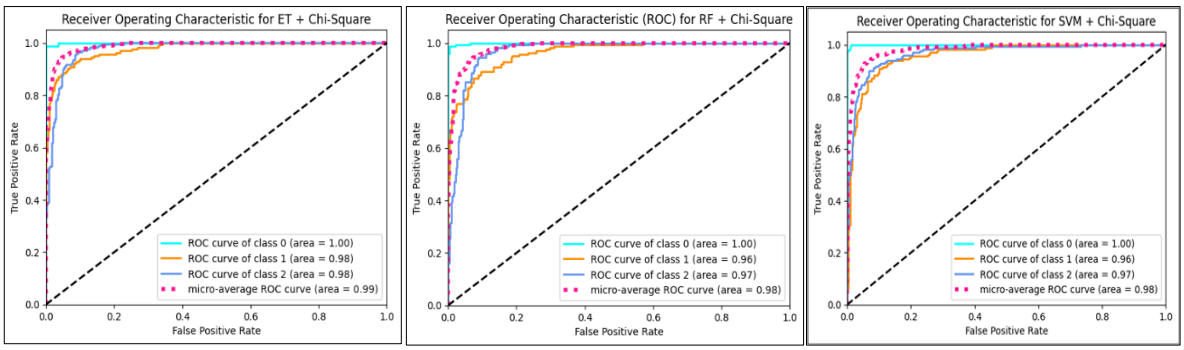

Figure 3-a: ROC-AUC curves of (ET, RF, SVM) with (Chi-Square).

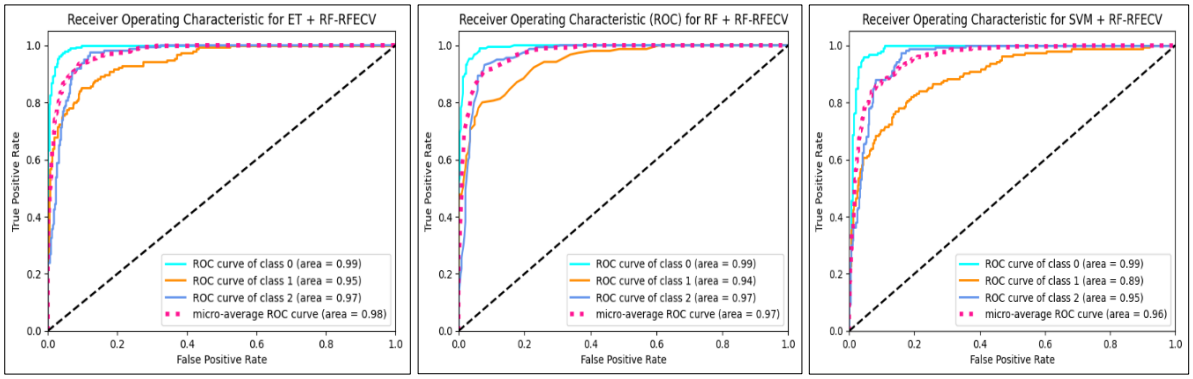

Figure 3-b: ROC-AUC curves of (ET, RF, SVM) with (RF-RFECV).

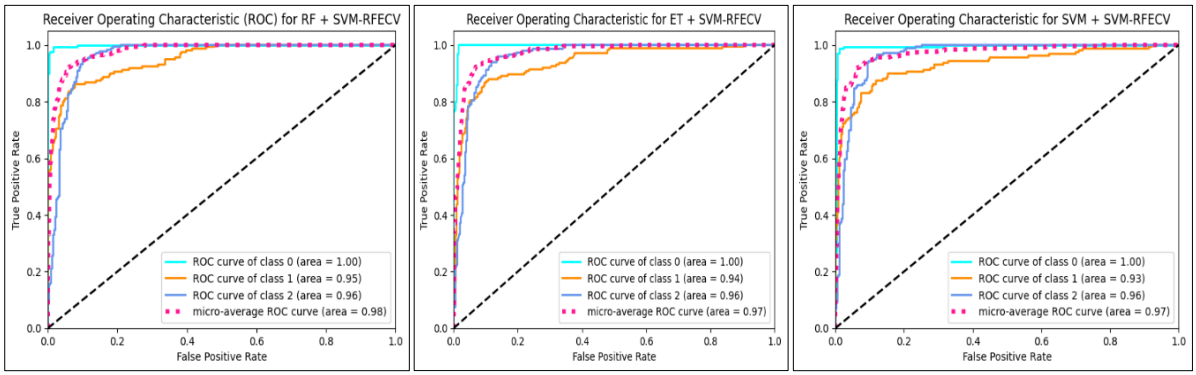

Figure 3-c: ROC-AUC curves of (ET, RF, SVM) with (SVM-RFECV).

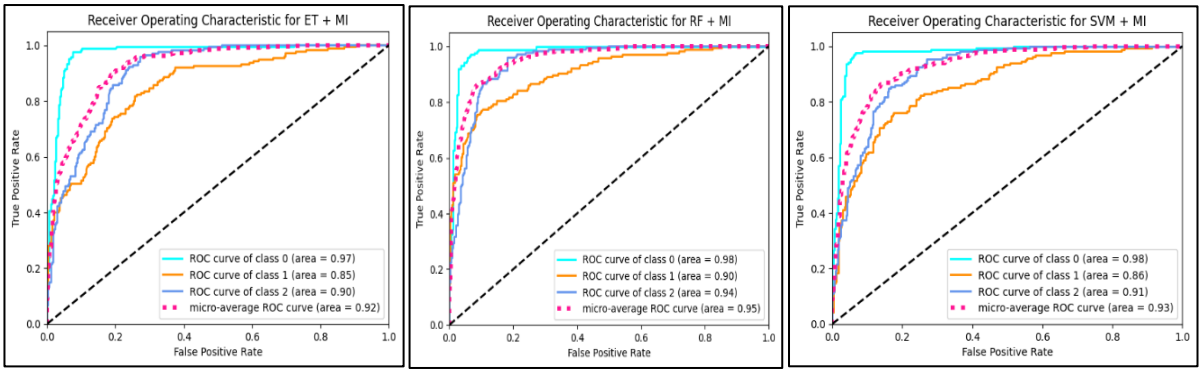

Figure 3-d: ROC-AUC curves of (ET, RF, SVM) with (Mutual Information).

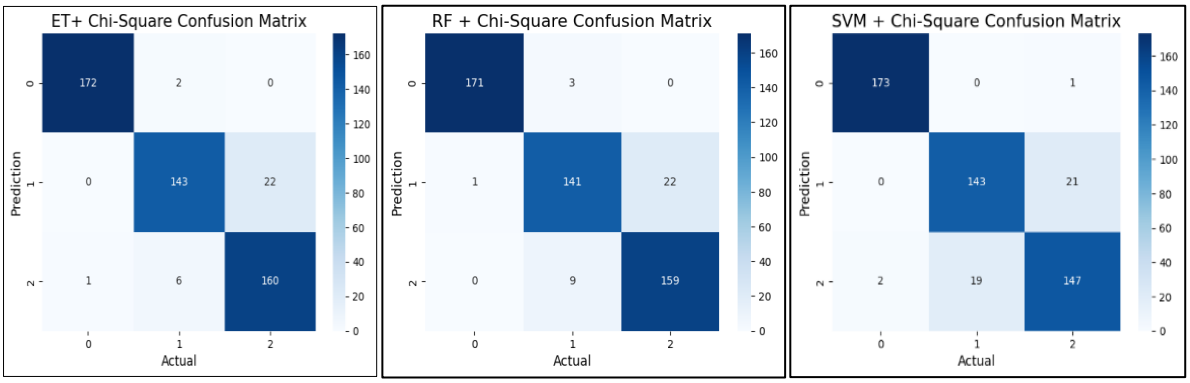

Figure 4-a: Confusion matrix of (ET, RF, SVM) with Chi-Square.

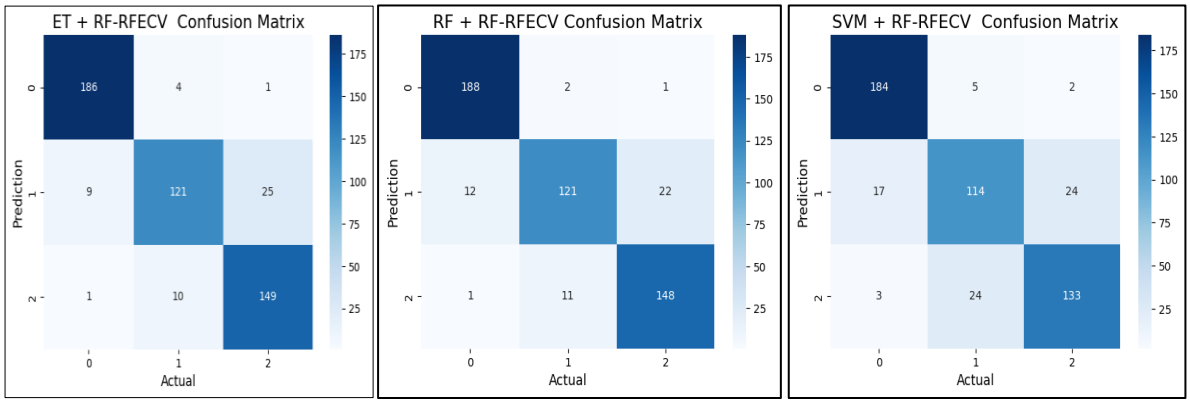

Figure 4-b: Confusion matrix of (ET, RF, SVM) with RF-RFECV.

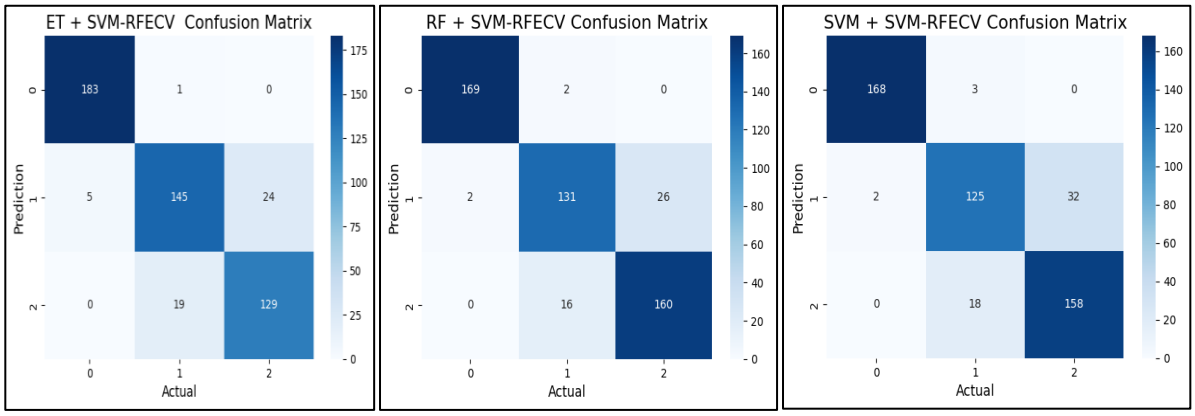

Figure 4-c: Confusion matrix of (ET, RF, SVM) with SVM-RFECV.

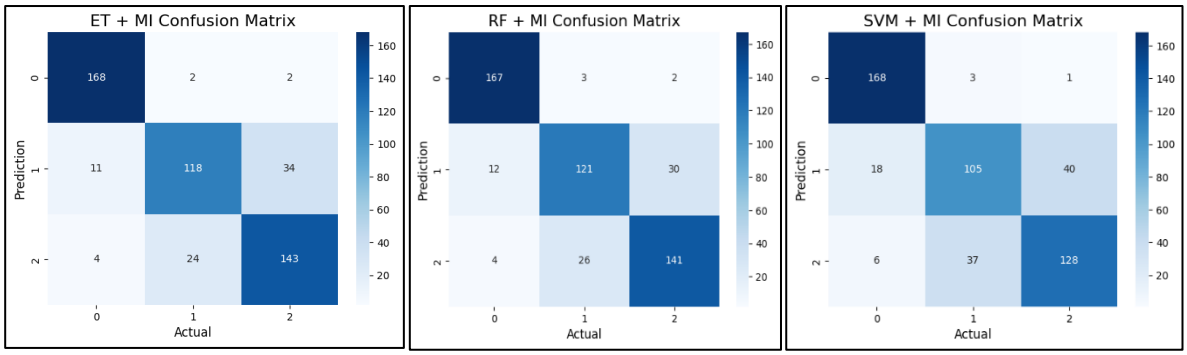

Figure 4-d: Confusion matrix of (ET, RF, SVM) with Mutual Information.

5 Discussion

Our proposal for this system comes as a result of a notable lack of studies investigating the application of ML in predicting cirrhosis. Compared to the state-of-the-art (SOTA) listed in ***Table 1***, our suggested system exhibited exceptional performance, achieving an accuracy of 93.87%. This is due to its unique contributions to the field of liver cirrhosis staging, including the dataset size, where the studies in [1], [3], [7], and [9] depend on small or heavily imbalanced datasets and the dataset contains many missing restrictions, which can bias performance metrics like accuracy and AUC. Conversely, the dataset employed by our system consists of a sufficiently extensive collection of clinical characteristics and very big samples. These factors are crucial for precisely determining the stage of liver cirrhosis, thereby enhancing the results' realism.

Regarding the process of feature selection. In [1], [7] use SHAP (SHapley Additive exPlanations). In [8], Boruta is used as a feature selection method. However, these methods may not be as effective in handling the specific dataset characteristics used in this study, especially with high-dimensional clinical data. The study in [3] did not use any feature selection methods, Where irrelevant features can lead to model bias and complicate interpretation, especially in clinical settings where understanding the relationship between features and outcomes is critical. Consequently, this may result in diminished performance.

Furthermore, the incorporation of ET as a classifier with Chi-Square as a feature selection method is an innovative methodology that has not been extensively explored in existing scholarly works. This combination has a substantial effect on the model's recall and AUC, especially in accurately identifying the precise patterns within clinical characteristics that indicate various stages of liver cirrhosis.

Another distinguishing characteristic is the selection of ET as the classifier. With its capacity to handle high-dimensional data and mitigate overfitting by employing ensemble learning, ET provides a more robust and accurate classification compared to single decision tree-based models or the linear models used by other state-of-the-art (SOTA) methods.

Regarding the dataset balance used in this study, our primary objective was to oversample rather than under-sample to prevent the loss of important information, so we did not use under-sampling techniques. Within the realm of oversampling approaches, a viable approach exists to augment the quantity of data through basic randomization. Conversely, SMOTE produces data using algorithms, thus reducing the probability of overfitting compared to basic random approaches. Consequently, our system achieved optimal performance.

The main constraint of this study is that the oversampling technique can artificially boost the number of minority-class instances by creating new ones based on their resemblance to existing minority examples. This raises apprehensions about the potential occurrence of overfitting during the ML procedure.

6 Conclusion

This paper proposed an automated system to predict the stages of cirrhosis employing ML methods. The performance comparison baseline was established by implementing three ML classifiers including ET, RF, and SVM. It used data mining techniques to pre-process the data, enhancing its quality and rendering it appropriate for analysis and predictive modeling. This research aims to aid medical professionals in diagnosing and managing this intricate disease, to limit the progression of the disease, and to reduce effort, time, and cost by predicting the early phases of cirrhosis as well. The proposed system goes through several stages. In the first stage, we converted the categorical data to digital. The second stage involved the utilization of feature selection methods to identify a subset of important and relevant features to reduce model training time; in addition to improving system performance, the outputs of this stage were balanced by applying the SMOTE method to be converted into a format more suitable for ML models. In the third stage, we partitioned the dataset into training and testing and subsequently employed a standard scaler to standardize the feature measurements. The system's performance was assessed by employing a variety of metrics, such as AUC-ROC, recall, precision, accuracy, F1-score, and confusion matrix. The (ET-Chi-Square) demonstrated a 93.87% accuracy rate, surpassing all other methodologies in terms of all metrics.

The current study's concentration on a single dataset allows for a thorough analysis within the defined scope. However, future work could broaden the scope by testing the model on other datasets related to liver cirrhosis, employing cross-validation methods, and using supplementary feature selection techniques to strengthen the reliability of the results and improve the predictive accuracy. Additionally, it is worth considering the use of domain adaptation strategies to improve the model's ability to adapt to different data sources, thereby ensuring broader applicability and greater clinical relevance.

References

[1] G. Arya, A. Bagwari, H. Saini, P. Thakur, C. Rodriguez, and P. Lezama, “Explainable AI for Enhanced Interpretation of Liver Cirrhosis Biomarkers,” *IEEE Access*, vol. 11, no. September, pp. 123729–123741, 2023, doi: 10.1109/ACCESS.2023.3329759.

[2] J. Pan *et al.*, “Epidemiology of portal vein thrombosis in liver cirrhosis: A systematic review and meta-analysis,” Oct. 01, 2022, *Elsevier B.V.* doi: 10.1016/j.ejim.2022.05.032.

[3] A. Utku, “DEEP LEARNING BASED CIRRHOSIS DETECTION,” *Oper. Res. Eng. Sci. Theory Appl.*, vol. 6, pp. 95–114, 2023, doi: 10.31181/oresta/060105.

[4] D. Jamil, S. Palaniappan, A. Lokman, M. Naseem, and S. S. Zia, "Diagnosis of Gastric Cancer Using Machine Learning Techniques in Healthcare Sector: A Survey," *Inform.*, vol. 45, no. 7, pp. 147–166, 2021, doi: 10.31449/inf.v45i7.3633.

[5] S. M. Birjandi and S. H. Khasteh, "A survey on data mining techniques used in medicine," *J. Diabetes Metab. Disord.*, vol. 20, no. 2, pp. 2055–2071, 2021, doi: 10.1007/s40200-021-00884-2.

[6] Q. Wang, X. Tang, W. Qiao, L. Sun, and H. Shi, "Machine learning-based characterization of the gut microbiome associated with the progression of primary biliary cholangitis to cirrhosis," *Microbes Infect.*, no. xxxx, p. 105368, 2024, doi: 10.1016/j.micinf.2024.105368.

[7] S. K. Kamath, S. K. Pendekanti, and D. Rao, "LivMarX : An Optimized Low-Cost Predictive Model Using Biomarkers for Interpretable Liver Cirrhosis Stage Classification," *IEEE Access*, vol. 12, no. June, pp. 92506–92522, 2024, doi: 10.1109/ACCESS.2024.3422451.

[8] O. M. Güneş, P. Kasap, and B. S. Çorba Zorlu, "The comparison of machine learning classification algorithms used to diagnose liver cirrhosis disease and a brief review," *Concurr. Comput. Pract. Exp.*, vol. 35, no. 8, Apr. 2023, doi: 10.1002/cpe.7628.

[9] K. Chen, Y. Wan, J. Mao, Y. Lai, G. Zhuo-Ma, and P. Hong, "Liver cirrhosis prediction for patients with Wilson disease based on machine learning: A case-control study from southwest China," *Eur. J. Gastroenterol. Hepatol.*, vol. 34, no. 10, pp. 1067–1073, Oct. 2022, doi: 10.1097/MEG.0000000000002424.

[10] "Liver disorders ." Accessed: Feb. 26, 2024. [Online]. Available: https://zenodo.org/records/6726768#.Y1vxci8Rpz8

[11] T. Al-shehari and R. A. Alsowail, "An insider data leakage detection using one-hot encoding, synthetic minority oversampling and machine learning techniques," *Entropy*, vol. 23, no. 10, Oct. 2021, doi: 10.3390/e23101258.

[12] X. A. Ma, H. Xu, and C. Ju, "Class-specific feature selection via maximal dynamic correlation change and minimal redundancy," *Expert Syst. Appl.*, vol. 229, Nov. 2023, doi: 10.1016/j.eswa.2023.120455.

[13] M. S. Pathan, A. Nag, M. M. Pathan, and S. Dev, "Analyzing the impact of feature selection on the accuracy of heart disease prediction," *Healthc. Anal.*, vol. 2, Nov. 2022, doi: 10.1016/j.health.2022.100060.

[14] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction," *Front. Bioinforma.*, vol. 2, no. June, pp. 1–17, 2022, doi: 10.3389/fbinf.2022.927312.

[15] P. Chittora *et al.*, "Prediction of Chronic Kidney Disease - A Machine Learning Perspective," 2021, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2021.3053763.

[16] A. Abdo, R. Mostafa, and L. Abdel-Hamid, "An Optimized Hybrid Approach for Feature Selection Based on Chi-Square and Particle Swarm Optimization Algorithms," *Data*, vol. 9, no. 2, 2024, doi: 10.3390/data9020020.

[17] F. H. Yagin *et al.*, "Estimation of Obesity Levels with a Trained Neural Network Approach optimized by the Bayesian Technique," *Appl. Sci.*, vol. 13, no. 6, 2023, doi: 10.3390/app13063875.

[18] N. R. Abid-Althaqafi and H. A. Alsalamah, "The Effect of Feature Selection on the Accuracy of X-Platform User Credibility Detection with Supervised Machine Learning," *Electron.*, vol. 13, no. 1, Jan. 2024, doi: 10.3390/electronics13010205.

[19] M. Awad and S. Fraihat, "Recursive Feature Elimination with Cross-Validation with Decision Tree: Feature Selection Method for Machine Learning-Based Intrusion Detection Systems," *J. Sens. Actuator Networks*, vol. 12, no. 5, Oct. 2023, doi: 10.3390/jsan12050067.

[20] K. Shi, R. Shi, T. Fu, Z. Lu, and J. Zhang, "A Novel Identification Approach Using RFECV–Optuna–XGBoost for Assessing Surrounding Rock Grade of Tunnel Boring Machine Based on Tunneling Parameters," *Appl. Sci.*, vol. 14, no. 6, Mar. 2024, doi: 10.3390/app14062347.

[21] J. Park, S. Kwon, and S. P. Jeong, "A study on improving turnover intention forecasting by solving imbalanced data problems: focusing on SMOTE and generative adversarial networks," *J. Big Data*, vol. 10, no. 1, Dec. 2023, doi: 10.1186/s40537-023-00715-6.

[22] S. Wesolowski and N. Klco, "An Overview of Overfitting and its Solutions An Overview of Overfitting and its Solutions," 2019, doi: 10.1088/1742-6596/1168/2/022022.

[23] M. I. Prasetiyowati, N. U. Maulidevi, and K. Surendro, "The accuracy of Random Forest performance can be improved by conducting a feature selection with a balancing strategy," *PeerJ Comput. Sci.*, vol. 8, 2022, doi: 10.7717/PEERJ-CS.1041.

[24] L. B. V. de Amorim, G. D. C. Cavalcanti, and R. M. O. Cruz, "The choice of scaling technique matters for classification performance," *Appl. Soft*

Comput., vol. 133, Jan. 2023, doi: 10.1016/j.asoc.2022.109924.

[25] S. S. Islam, M. S. Haque, M. S. U. Miah, T. Bin Sarwar, and R. Nugraha, "Application of machine learning algorithms to predict the thyroid disease risk: an experimental comparative study," *PeerJ Comput. Sci.*, vol. 8, 2022, doi: 10.7717/PEERJ-CS.898.

[26] A. Sabir, H. A. Ali, and M. A. Aljabery, "ChatGPT Tweets Sentiment Analysis Using Machine Learning and Data Classification," *Inform.*, vol. 48, no. 7, pp. 103–112, 2024, doi: 10.31449/inf.v48i7.5535.

[27] M. A. Hossain and F. Amenta, "Machine Learning-Based Classification of Parkinson's Disease Patients Using Speech Biomarkers," *J. Parkinsons. Dis.*, vol. 14, no. 1, pp. 95–109, Jan. 2024, doi: 10.3233/JPD-230002.

[28] J. A. Alhijaj and R. S. Khudeyer, "Integration of EfficientNetB0 and Machine Learning for Fingerprint Classification," *Inform.*, vol. 47, no. 5, pp. 49–56, 2023, doi: 10.31449/INF.V47I5.4724.

[29] C. Delrue, S. De Bruyne, and M. M. Speeckaert, "Application of Machine Learning in Chronic Kidney Disease: Current Status and Future Prospects," Mar. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/biomedicines12030568.

[30] A. Rehman, T. Saba, M. Mujahid, F. S. Alamri, and N. ElHakim, "Parkinson's Disease Detection Using Hybrid LSTM-GRU Deep Learning Model," *Electron.*, vol. 12, no. 13, pp. 1–21, 2023, doi: 10.3390/electronics12132856.