Journal of Information Systems and Telecommunication
Vol.13, No.3, July-September 2025, 165-176
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

Predicting Primary Biliary Cholangitis Stages Using Machine Learning with Automated Hyperparameter Optimization and Recursive Feature Elimination

Arman Rezasoltani[1], Amir Mohammad Khani[1], Ali Husseinzadeh Kashan[2], Shahram Agah[3*], Fatemeh Agah[4]

[1].Department of Industrial Management, Faculty of Management, University of Tehran, Tehran, Iran.
[2].Department of Industrial Engineering, Faculty of Industrial and Systems Engineering, Tarbiat Modares University, Tehran, Iran.
[3].Department of Gastroenterology and Hepatolog, Colorectal Research Center, Iran University of Medical Sciences, Tehran, Iran.
[4].The University of Adelaide, Discipline of Medicine, Adelaide, South Australia, Australia. Fatemeh.

Received: 30 Jan 2025/ Revised: 04 Sep 2025/ Accepted: 05 Oct 2025

Abstract

This research used modern machine learning ways to predict the stages of primary biliary cholangitis using data from the Mayo Clinic trial. The research aims to obtain high prediction accuracy while representing balanced evaluation metrics. Important techniques include automated hyperparameters optimization with Optuna and Recursive Feature Elimination to improve model performance. Pre-processing included handling missing values, encoding of categorical features, and addressing class imbalances using SMOTE. A total of twelve machine learning algorithms are evaluated with ensemble-based models such as CatBoost and Extra Trees producing much better results. Evaluation metrics take into account all model predictions, including accuracy, precision, recall, F1 score, and ROC-AUC for performing balanced and interpretative evaluations of performances critical for imbalanced datasets. This endeavor includes clinical and laboratory information illustrating the prospect of machine learning in advancing therapeutic diagnosis, emphasizing the rigor and robustness in evaluation laid groundwork for future research to encompass even more generalizable and robust diagnostic tools.

Keywords: Primary Biliary Cholangitis; Machine Learning; Recursive Feature Elimination; Optuna, Imbalanced Data.

1- Introduction

Primary Biliary Cholangitis (PBC), formerly known as primary biliary cirrhosis, is a chronic autoimmune liver disease. It is characterized by the gradual and progressive destruction of the liver's small bile ducts, leading to the accumulation of bile and other toxins within the liver, a condition known as cholestasis. Over time, this persistent damage can result in scarring, fibrosis, and ultimately cirrhosis. Cirrhosis is a late-stage liver disease that occurs when scar tissue replaces healthy liver tissue. The underlying pathologies that may cause this disease include viral hepatitis, chronic alcoholism, and NAFLD (non-alcoholic fatty liver disease) (Konerman et al., 2019). Chronic alcohol consumption leads to advanced forms of liver damage, which eventually result in cirrhosis and subsequent liver failure (Topcu et al., 2024). In the primary stages, the disease is asymptomatic, and awareness is typically raised only in the advanced stages. Cirrhosis may lead to liver failure, liver cancer, and, ultimately, death (Tapper & Parikh, 2023). There is a strong need for the most accurate and least invasive methods to predict the progression of cirrhosis, given the critical importance of diagnosing and managing such diseases optimally. Although traditional methods, such as liver biopsy, provide accurate results, these procedures are invasive and may lead to complications (Wei et al., 2018). Chronic alcohol consumption is one of the main causes of this disease and, in the long term, can lead to advanced stages such as cirrhosis, ultimately culminating in complete liver failure

✉ **Shahram Agah**
Agah.sh@iums.ac.ir

(Topcu et al., 2024). Previous studies have established that cirrhosis of the liver progresses through four stages. The first stage, Steatosis, is characterized by inflammation of either the liver or the bile ducts, and immediate treatment at this juncture can control the disease. The second stage, Fibrosis, involves the development of scar tissue that cuts off normal blood flow to the liver and impairs its function; however, medical treatment can halt the progression of the disease. In the third stage, Cirrhosis, healthy liver tissue is replaced by scar tissue, and swelling may occur in the spleen. Finally, the fourth stage, Liver Failure, is characterized by complete liver failure. At this stage, patients transition from normal health to a comatose state and require emergency treatment by medical professionals (Wei et al., 2018).

The subtlety of its early symptoms permits the diagnosis of cirrhosis only at advanced stages; if mismanaged, the disease can inevitably culminate in liver failure or cancer. Recent studies have highlighted the significance of early detection and management. An SEAL screening algorithm study demonstrated a remarkable 59% higher rate of early cirrhosis detection compared to routine care, thereby advocating for the role of structured programs in identifying asymptomatic cases (Labenz et al., 2022). In addition, top-down proteomics identified the proteoform signatures in plasma that correlate with the progression of cirrhosis, forming the template for a biomarker-driven risk stratification (Forte et al., 2024). Another paper emphasized the role of miRNA-gene regulatory axes in monitoring and diagnosing cirrhosis and hepatocellular carcinoma and proposed new diagnostic targets (Premnath & Shanthi, 2024). Asymptomatic superior mesenteric vein thrombosis (SMVT), however, has not been proven to significantly impact cirrhosis outcomes, unlike the risks posed by portal thrombosis (PT) (Wang et al., 2022). These collective findings emphasize the crucial role of early, target-oriented interventions and the potentially significant role of additional biomarkers in preventing the progression of asymptomatic cirrhosis. Prior studies discussed the notable success of various machine-learning-based approaches like Random Forest, Gradient Boosting, Ensemble Learning, and others in increasing the accuracy with which the stages of disease progression are predicted. For example, the LivMarX model achieved an accuracy of up to 86% for predicting different stages of cirrhosis based on a combination of biomarkers and optimization techniques (Kamath et al., 2024). Other models suggested that longitudinal models outperformed other cross-sectional models in accurately detecting disease progression (Hanif et al., 2022).

Millions live with cirrhosis worldwide, and it remains a leading cause of death every year. The effects on patients" quality of life following late diagnosis of cirrhosis can be dire and place a huge burden on the health sector. Furthermore, improper management of the disease may lead to serious complications, such as advanced liver failure, liver cancer, and other comorbidities (Hanif et al., 2022). New artificial intelligence and machine-aided processes enable much finer accuracy in determining the stage of the disease and are immensely beneficial in reducing complications, promoting early diagnosis, and improving patient management. The ability of this technology to offer a serious advancement in the management of cirrhosis is most felt in areas where modern imaging methods are seldom available (Topcu et al., 2024). This research aims to develop an efficient and accurate model for predicting early liver cirrhosis by employing advanced machine learning algorithms. It seeks to improve prediction accuracy by combining intelligent feature selection and model optimization approaches to create models that are not only highly efficient but also practical for implementation in real clinical settings. The major aim of the study is to devise a model for prediction of stage of PBC that is accurate, generalizable, and efficient using advanced techniques of machine learning. Some cutting-edge work presented therein involves, but is not restricted to, tuning of model hyperparameters via advanced optimization methods of Optuna, feature selection algorithms, such as RFECV to identify crucial disease progress variables. A further significant aspect in the study includes the use of rich and varied data composed of clinical and laboratory data drawn from credible sources. The evaluation of model performance metrics such as accuracy, precision, recall, F1-score, and AUC is performed in a very detailed way so as to allow transparency in the evaluation of the quality of predictions. This paper is organized as follows: the first part introduces the research and its various objectives; the second part broaches the research background and pinpoints the weaknesses of previous studies; the third part describes the research methodology regarding the dataset, preprocessing techniques and machine learning algorithms used; the penultimate section conveys all the experimental results and critically evaluates the performance of various models; and finally, the last part deliberates and draws its conclusions in respect of the findings obtained, drawing comparisons with previous studies, scrutinizing the implications of the results, providing an overview of the contributions made, and suggesting future areas of research. In this study, such a constructive approach enhances the efforts toward improving the prediction of cirrhotic liver disease risk while further enhancing the development of AI in aiding diagnostic medicine.

2- Theoretical Foundations and Research Background

In very recent times, prognosis and evaluation of liver diseases have made remarkable advancements. Cirrhosis often deteriorates into liver failure, requiring transplants in

many cases, often due to chronic liver insult. Making an accurate diagnosis of the stage of liver cirrhosis and tracking the patients' progress remains among the greatest challenges of medicine. Addressing these difficulties straightaway impacts treatment strategies and the potency of medical involvement. In the past years, machine learning methods have emerged as a contemporary remedy for prognosticating the diverse phases of liver cirrhosis. These algorithms identify clinically pertinent traits that describe singular patient characteristics through exhaustive data examination. Table 1 briefly summarizes related research on predicting the stages of liver cirrhosis and contrasts assorted methods. This table comprises the titles of the reports, aims, datasets, machine learning algorithms, and key outcomes of each analysis. An inspection of this background reveals that machine learning designs such as Random Forest, Support Vector Machine, and amalgamated tactics, exploiting an assortment of datasets and sundry optimization techniques, have been successfully applied and have achieved meaningful accuracy in prognosticating the phases of liver cirrhosis. This data furnishes worthwhile insights into the strengths and shortcomings of preceding studies and helps pinpoint existing research gaps.

Table 1. Research background

Authors	***Article Title***	***Goals***	***Model used***	***Dataset***	***Conclusion***
Konerman et al. (2019)	Machine learning models to predict disease progression among veterans with hepatitis C virus	Predict cirrhosis progression in CHC patients	Cox models and boosted-survival-tree model	Veterans' Health Administration (72,683 individuals)	The longitudinal boosted survival tree model achieved superior concordance (0/774) and AuROC in prediction compared to cross-sectional models, demonstrating higher reliability in long-term forecasts.
Topcu et al. (2024)	Machine Learning-Based Analysis and Prediction of Liver Cirrhosis	Early detection of liver cirrhosis	Random Forest, Logistic Regression, AdaBoost, k-Nearest Neighbors	Open-access liver cirrhosis dataset	The Random Forest model achieved high accuracy (~98%), demonstrating superior performance in early cirrhosis prediction. Precision, recall, and F1-score were not explicitly reported.
Bhardwaj et al. (2024)	Improving Prognostic Prediction of Cirrhosis Using an Optimized Ensemble Machine Learning Approach	Enhance prediction of cirrhosis prognosis	Ensemble model integrating Gradient Boosting, Random Forest, and Decision Trees	Multisource liver disease datasets	The ensemble models improved prediction accuracy and generalizability, making significant advances in reliability and forecasting. While specific metrics such as accuracy, precision, and recall were not directly reported, overall improvements were observed.
K et al. (2024)	Stage Prediction of Liver Cirrhosis Disease using Machine Learning	Determine stages of liver cirrhosis	Support Vector Machine, Random Forest, Gradient Boosting	Dataset with 418 records and 20 attributes	Random Forest was among the models with the highest accuracy (~97%), achieved through feature engineering and cross-validation. Precision, recall, and F1-score for the Random Forest model are not specified.
Kamath et al. (2024)	LivMarX: An Optimized Low-Cost Predictive Model Using Biomarkers for Interpretable Liver Cirrhosis Stage Classification	Stage liver cirrhosis using biomarkers	Random Forest (optimized with Genetic Algorithm and GridSearchCV)	Comprehensive dataset of 424 patients	LivMarX achieved over 86% accuracy after optimization, with an AUC of 0/95. The model demonstrated high cost-effectiveness for accurately staging cirrhosis in the absence of imaging. Precision, recall, and F1-score were not reported.
(Elmasinejad and Golabpour, 2024)	Predicting Liver Fibrosis Severity Using Machine Learning Models	Development of a machine learning model for diagnosing fatty liver using demographic information and hematology tests	Support Vector Machine (SVM) with Radial Kernel	Data from 1,078 patients referred to Imam Reza Hospital	The model achieved 93/55% accuracy on the training data and 78/62% on the test data, outperforming six comparable algorithms.
Jamadar et al. (2023)	Cirrhosis Disease Prediction Using Machine Learning	Using machine learning methods to	Different machine	Data from patients with	The proposed model demonstrated high accuracy in predicting the stages of cirrhosis.

		predict liver cirrhosis	learning algorithms	physiological characteristics associated with cirrhosis	
Hanif and Khan (2022)	Liver Cirrhosis Prediction Using Machine Learning Approaches	Predict liver cirrhosis stages	Support Vector Machine, Decision Tree, Random Forest	Liver Cirrhosis dataset (418 records)	Random Forest achieved an accuracy of ~97%, demonstrating reliability and robustness in phase-wise predictions of liver cirrhosis. Precision, recall, and F1-score were not reported.
Sidana et al. (2022)	Liver Cirrhosis Stage Prediction Using Machine Learning: Multiclass Classification	Predicting the stage of liver cirrhosis in patients using machine learning algorithms	Artificial Neural Network, Random Forest , Logistic Regression, Support Vector Machine , KNN, Decision Tree , Naive Bayes	Data from patients with liver cirrhosis	The Artificial Neural Network (ANN) demonstrated the best performance with high accuracy, while the RF+MI feature selection method showed a slight improvement over the standard Random Forest (RF) model.

The studies discussed in Table 1 delineate just some of the many advances in the use of machine learning algorithms in predicting the stages of liver cirrhosis. However, one of the main gaps identified there was the significant delay in consideration of imbalanced data sets and excessive focus on a single performance metric, such as accuracy, for model evaluation. The studies by Bhardwaj et ub. and Sidana et ub., while dealing with random forest or SVM, do not appease the challenge of imbalanced dataset(s), and they wholly rely on a single evaluation criterion, such as accuracy, thus not completely evaluating models one through other proper performance criteria such as Precision, Recall, and F1 Score. Such excessive focus on accuracy alone results in a very skewed perspective on their prediction capabilities, since such models often guarantee high-performance measures yet produce very poor results on overweighted classes. Another very important limitation discussed in Table 1 is their use of unoptimized models and poorly defined feature sets. For example, models like Random Forests and SVM have been applied, ill as the studies by Hanif and Khan, and Jamadar et al., did not apply state-of-the-art optimization techniques that would potentially improve model performance, structure feature selections, and reduce the framework of their studies, thus precluding meaningful generalization and accuracy of their interpretations. In the contrary, the current paper uses a rather spirited approach by using advanced machine learning algorithms guaranteeing accuracy in predictions and correcting the data imbalance, with the models being subjected to various acute evaluations by areas such as accuracy, precision, recall, F1 score, and ROC-AUC, which is possible to ascertain an appropriate and transparent evaluation of the models' performances addressing fundamental gaps in prior research and leading the investigation towards more reliable and generalized results.

Moreover, a large number of studies will focus only on one model, with limited analysis of the effects of combinations of algorithms or full comparisons between the efficiency of techniques. The novel methodology presented in this paper serves as an ensemble framework to enrich predictive technology, apply advanced feature selection techniques, optimize model computational costs, and improve the implementation of models openly in the real world, all of which are overly venturous in previous studies, such as the LivMarX (Kamath et al., 2024). Finally, this research makes a significant contribution to advancing existing methods by focusing on early-stage liver cirrhosis prediction, presenting a comprehensive optimization framework, thoroughly analyzing model performance indicators, and utilizing diverse and extensive datasets. Through the articulation of emerging and current research gaps, as well as the modest input of novelties, this will provide a further route for an exhaustive yet accurate approach to be developed in this area.

3- Research Method

The goal of this study was to use machine learning algorithms to predict the stage of primary biliary cholangitis (PBC) in patients. The main objective is to use the machine learning model to accurately predict the stage of the disease using medical and laboratory data. The dataset used in this study was derived from a clinical investigation of PBC patients conducted at the Mayo Clinic and supplemented by a publicly available dataset released on the Kaggle platform, which included numerous original features. After data analysis and feature selection, key variables were identified using recursive feature elimination with cross-validation (Priyatno Widiyaningtyas, 2024). During the preprocessing stages, correlation analysis was performed, and the SMOTE

method was applied to address class imbalance. Additional steps included handling missing values, and encoding categorical features (Khan & Hoque, 2020). Twelve machine learning algorithms were evaluated for modeling purposes: Decision Tree, Random Forest, Extra Tree, Gradient Boosting, AdaBoost, XGBoost, LightGBM, Logistic Regression, Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Naive Bayes, and CatBoost. The Optuna optimization framework was used to fine-tune the hyperparameters of all models in such a way as to provide the best performance (Jeganathan et al., 2024). The performance of the models was assessed against four main metrics: accuracy, precision, recall, and F1-score (Fazel & Foing, 2024). In addition, the ROC curve and AUC values are used for more details regarding the model performance. All other steps of this study were done using the Python programming language with its corresponding libraries.

3-1- Data Source

The data set used in this study was extracted from the Cirrhosis Prediction Dataset, which is publicly available on the Kaggle platform. It includes information of patients with PBC, collected over ten years in a clinical study carried out at the Mayo Clinic. In this study, 420 patients diagnosed with PBC were identified as eligible to participate in a randomized, controlled trial of the drug D-penicillamine. Of these, 312 patients obtained consent to participate in the randomized clinical trial, their records had a minimal loss. There were also 112 other eligible patients who were not trial participants, who did allow for basic information and survival follow-ups to be recorded; 6 out of these 112 patients were lost from follow-up soon after diagnosis, so data on 106 remained. Thus, the total number of patients entered in the dataset is 418 (Fedesoriano,2021).

3-2- Dataset Features

The data used in this study include comprehensive information from patients with PBC. The dataset initially comprised 20 features, which are presented in Table 2.

Table 2. Variables Description

Feature Name	***Description***	***Type***	***Values/Unit***
ID	Unique identifier for each patient	Categorical	Numeric
N_Days	Number of days between registration and the earlier of death, transplantation, or study analysis time	Numeric	Days
Status	Status of the patient	Categorical	C (Censored), CL (Censored due to liver tx), D (Death)
Drug	Type of drug administered	Categorical	D-penicillamine, Placebo
Age	Age of the patient	Numeric	Days
Sex	Gender of the patient	Categorical	M (Male), F (Female)
Ascites	Presence of ascites	Categorical (Binary)	N (No), Y (Yes)
Hepatomegaly	Presence of hepatomegaly	Categorical (Binary)	N (No), Y (Yes)
Spiders	Presence of spiders	Categorical (Binary)	N (No), Y (Yes)
Edema	Presence of edema	Categorical	N, S, Y
Bilirubin	Serum bilirubin	Numeric	mg/dl
Cholesterol	Serum cholesterol	Numeric	mg/dl
Albumin	Serum albumin	Numeric	gm/dl
Copper	Urine copper	Numeric	µg/day
Alk_Phos	Alkaline phosphatase	Numeric	U/liter
SGOT	SGOT (serum glutamic-oxaloacetic transaminase)	Numeric	U/ml
Triglycerides	Serum triglycerides	Numeric	mg/dl
Platelets	Platelet count	Numeric	per cubic ml/1000
Prothrombin	Prothrombin time	Numeric	Seconds (s)
Stage	Histologic stage of the disease	Categorical (Ordinal)	1, 2, 3, 4

In this study, the target variable was defined as Stage, representing a disease stage that ranges from 1 to 4. The aim is to model the Stage variable in relationship to the other features in the data set. The ID column was ruled out of the analysis simply because it works as a patient identifier and provides no substantial contribution to prediction.

3-3- Data Cleaning

The cohort included 424 patients with PBC data collected as part of a Mayo Clinic clinical trial. Of those, the final analysis was based on 312 samples. In the first step of cleaning the data, the ID column, which was judged not relevant to the target variable, was deleted as it would not contribute to prediction. In addition, missing values in features with limited incompleteness were substituted with the mean value for less impact on the modeling. Out of the 424 data points, 112 pertained to patients who did not participate in the randomized tests and had incomplete information. Out of these, six samples were excluded shortly after data collection due to critical missing information. According to strict sampling standards, the information from the remaining 112 non-participating patients had to be rejected because of poor quality. This left 312 samples that were complete and of good quality for analysis. Data cleaning allowed such preparation, producing better quality data for the predictions.

3-4- Correlation Analysis

Correlation analysis was conducted to identify linear relationships between variables in the dataset. The primary

purpose of this analysis was to determine variables with a significant impact on the target variable and to eliminate those with redundant or weak associations with other variables. In this study, a correlation matrix, visualized using a heatmap, was employed to illustrate the relationships between variables.

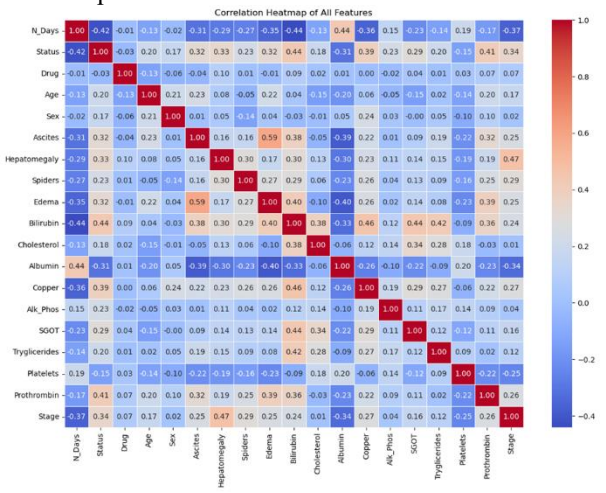

Figure 1. Correlation Heatmap

From the correlation analysis, no variables exhibited high correlation with other variables (greater than 0/8 or less than -0/8). The highest positive correlation found is between the Copper and Bilirubin (about 0/46), indicating no removal of features for redundancy because of excessive correlation. Furthermore, it is found that the independent variable (Stage) correlates positively with Hepatomegaly (about 0/47), thus this variable is important in predicting the stage of the disease. In this regard, all the features were retained for modeling since they provide independent and informative information. Such independence can be expected to add strength to model value.

3-5- Feature Selection

Therefore, feature selection becomes a big step for preprocessing data to enhance the performances of machine learning classifiers and reduce computational complexity. The dataset initially had many primary features, but some of them had bad correlations with the target variable or brought more noisy and redundant information. To extract important features, RFECV was used. RFECV is a very efficient recursive feature elimination mechanism (Thambawita et al., 2020) that starts by training the model with all features available, estimates the importance of each individual feature in terms of importance score such as those derived from feature importance or model coefficients, and then removes one feature at a time, retraining the model at each iteration. The process continues until all possible combinations of features have been tried. It implements cross-validation to find the best set of features. The other applications of cross-validation are to make the dataset as many segments as needed, then evaluate the model performances for each feature combination. Finally, RFECV was used to optimize feature selection based on model performance during cross-validation. In addition to evaluating model performance, this technique effectively eliminates irrelevant features, selecting the minimum number of features necessary to make accurate predictions. In this study, a total of 14 features were identified as the most informative from the initial set: N_days, status, drug, age, bilirubin, cholesterol, albumin, copper, alk_phos, sgot, triglycerides, platelets, and prothrombin. These selected features were found to significantly contribute to the prediction of disease stages. The removal of non-essential features reduced model complexity while improving model estimation accuracy and computational efficiency. Figure 2 illustrates the significance of these features in this study.

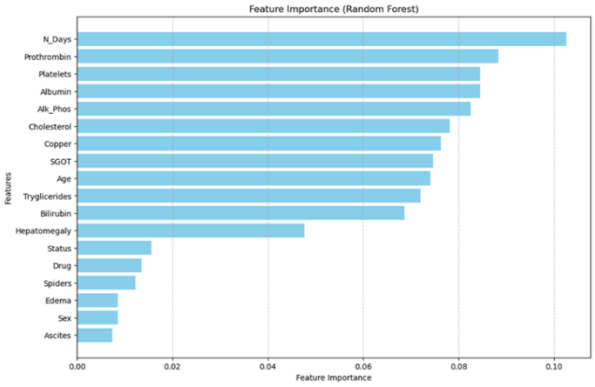

Figure 2. Feature Importance

3-6- Data Normalization

The MinMaxScaler is used to scale data for SVM (Support Vector Machine) and KNN (K-Nearest Neighbors) algorithms (Ali, 2022). This choice is made because these algorithms are generally sensitive to feature scaling. For SVM algorithms, to determine the separating hyperplane, the feature values are being used; whereas KNN uses feature values to compute distances amongst samples. Thus, features in varied scales could significantly affect the models' performance. The MinMaxScaler scales every feature to a fixed-range value, usually ranging between 0 and 1, on an equivalent scale. The formula for MinMaxScaler is:

$$x\_scaled=(x-x\_min)/(x\_max-x\_min) \quad (1)$$

In this formula:
xscaled is the normalized (scaled) feature value.
x is the original value of the feature.
xmin is the smallest value of the feature in the dataset.
xmax is the largest value of the feature in the dataset.

3-7- Data Balance

One of the major challenges outlined in this study was the distribution of samples into the different classes with unequal frequency. From the data distribution, it has been noted that there were only 16 samples at Stage I, while there were 97 samples at Stage II, 109 samples at Stage III, and more than to bring the order at the top. This imbalance causes the machine learning algorithms to converge toward the large classes, thus reducing any learning focused on the smaller classes, like stage I. This will probably have the effect that the model identifies the classes having more samples correctly, while disregarding or misclassifying the classes that have very few samples.

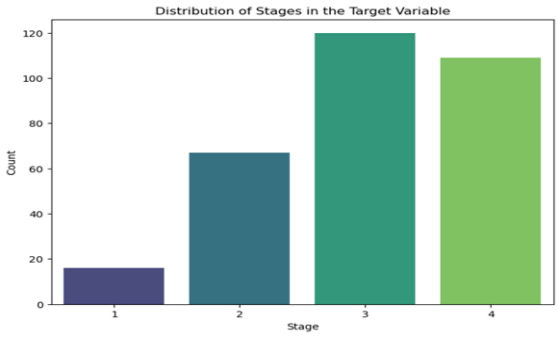

Figure 3. Distribution of Stages

The SMOTE method was used to increase the number of samples belonging to the minority class in the data set to remove imbalance namely synthetic minority over-sampling technique. It constructs synthetic instances and follows the following steps:

1. A random sample from the minority class is chosen as a reference sample.
2. Using the KNN algorithm (usually with K = 5) several nearest neighbors from the same minority class, are identified.
3. SMOTE generates new synthetic examples in feature space. This is achieved by selecting at random one of the nearest neighbors and by creating a new sample at a point in-between the reference sample and the chosen neighbor.

The formula used to compute the interpolation between the two samples is expressed as:

$$X\_new = X\_sample + gap \times (X\_neighbor - X\_sample) \quad (2)$$

Here, Xsample stands for the reference sample, Xneighbor for one of the nearest neighbors, and Gap for some random number in the range (0, 1). The dataset in this research was divided into two parts: training 70% of the data and using 30% for the encoding models' performance evaluation.

3-8- Machine Learning Algorithms

For predicting the stage of PBC in this study twelve different machine learning algorithms were used. These algorithms were used to identify the best-performing model that would predict the disease stages with the highest accuracy. The hyperparameters of each algorithm were optimized using the Optuna tool. Optuna is a dynamically designed hyperparameter optimization tool to automatically find the best values for model parameters (Akiba et al., 2019). Like others, efficiently finds the best hyperparameter configurations with advanced search techniques like Tree-structured Parzen Estimator (TPE) and Random Search. By running several tests and comparing how models perform, this tool minimizes the time to gain optimality. The table below provides the list of 12 machine learning algorithms, operational mechanisms, and the optimized values achieved using Optuna:

Table 3. Machine learning algorithms used and optimized hyperparameter values

Algorithm	Method	Optimal hyperparameters
Decision Tree	The algorithm applies successive splitting of the data into either two or more subsets. At every stage, one feature which works best for data splitting is selected according to certain criteria, some of which are Gini Index and Entropy(Mienye & Jere, 2024).	max_depth=32, min_samples_split=8
Random Forest	This algorithm, using a combination of multiple decision trees to reduce data variance, trains each tree on a random subset of the data and obtains its final output by following the majority voting rule in the case of classification, or averaging in the case of regression (Schonlau & Zou, 2020).	n_estimators=331, max_depth=8
Extra Trees	It operates similarly to Random Forest but uses random values instead of optimal values for node splitting. This approach reduces variance and results in faster model training (Geurts et al., 2006).	n_estimators=373, max_depth=14
Gradient Boosting	To build weak models (decision trees) one after the other, correcting the mistakes done by the previous model. The aim is to gradually minimize model errors and boost performance with each step (Biau & Cadre, 2017).	n_estimators=191, learning_rate=0/02662

AdaBoost	This algorithm iteratively trains weak models (small decision trees) and assigns greater weight to misclassified samples at each step to create a stronger final model (Ding et al., 2022).	n_estimators=162, learning_rate=0/54684
XGBoost	An optimized version of Gradient Boosting that reconciles the conflicts between solving the execution speed and the execution accuracy by analyzing operations in parallel and using more efficient algorithms. This optimization method can address large amounts of information and diversity (Bentéjac et al., 2020).	n_estimators=162, learning_rate=0/54684
LightGBM	An optimized Boosting algorithm that grows leaves instead of levels. This method is suitable for large-scale, high-dimensional data and provides faster performance compared to other Boosting algorithms (Ke et al., 2017).	n_estimators=329, num_leaves=210, learning_rate=0/1247
CatBoost	A fast and efficient Boosting algorithm optimized for categorical data, which automatically encodes categorical values. This method requires fewer parameter adjustments compared to other Boosting algorithms (Dorogush et al., 2018)	iterations=435, depth=9, learning_rate=0/2872
Logistic Regression	A method for data classification using a linear model computes the probability of the data belonging to different classes using the logistic (sigmoid) function. It is well suited to low-dimensional datasets (Starbuck, 2023).	C=0/1228
Support Vector Machine	This algorithm finds an optimal hyperplane to separate classes in the feature space. Using the RBF kernel, it maps data to a higher-dimensional space, enabling nonlinear separation (Shmilovici, 2023).	C=459/87, gamma=0/0573, kernel='rbf'
K-Nearest Neighbors	The prediction takes into account the distance of the other instances from the input data. The majority class among the k nearest neighbors is considered for predicting the class of the novel sample (Halder et al., 2024).	n_neighbors=3
Naive Bayes	A probabilistic model based on Bayes' theorem. This algorithm assumes complete independence between features and is well-suited for low-dimensional and categorical data (Pajila et al., 2023).	Lacks suitable hyperparameters for optimization.

To evaluate the performance of machine learning models in this study, five key metrics were used: accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (ROC-AUC). These metrics are defined based on the concepts of True Positive (TP) and True Negative (TN) for correct predictions, and False Positive (FP) and False Negative (FN) for incorrect predictions.

Table 4. Evaluation indicators for machine learning models

index	definition	Formula
Accuracy	The ratio of correct predictions (both positive and negative) to the total number of samples.	(TP+TN)/(TP+FP+FN+TN)
Precision	The ratio of correctly predicted instances for a class to all instances predicted as that class.	TP/(TP+FP)
Recall	The ratio of correctly predicted instances for a class to all actual instances of that class.	TP/(TP+FN)
F1 Score	The harmonic mean of Precision and Recall, balancing the trade-off between the two metrics.	(2×Precision×Recall)/(Precision+Recall)

The ROC-AUC metric measures the performance of a classification model at all threshold levels and illustrates how well the model is at distinguishing between classes; thus, it shows how well the model can predict the different stages of the disease. The ROC curve is created by plotting the value of false positive rate (FPR) vs true positive rate (TPR) for different thresholds and area under this curve is known as the AUC. AUC can be understood as the higher the better: The closer the AUC value is to 1, the better. In order to test the generalizability of the model and verify that it performed successfully regardless of the dataset with 5-Fold Cross-Validation was performed. In this method, the data set is split into five equal parts. At each iteration, one of its sections is considered as test data, while the other four sections are used as training data. This is done five times to guarantee that each batch is tested once. Finally, the overall performance of the model is reported as the mean values of all the evaluation metrics across all iterations.

4- Results

In this section, the results of the machine learning models are presented and analyzed. The Python programming language was utilized for this study, and all models were executed on a system equipped with an Intel Core i7-13700H processor, 16GB of RAM, and Python version 3/12.

The following outlines the performance results of the models.

Table 5. Comparison of results

Model	Accuracy	Precision	Recall	F1 Score
CatBoost	0/7708	0/7688	0/7708	0/7519
Extra Trees	0/7569	0/7636	0/7569	0/7400
LightGBM	0/7292	0/7182	0/7292	0/7126
Random Forest	0/7222	0/7146	0/7222	0/7085
Gradient Boosting	0/7153	0/7057	0/7153	0/7017
XGBoost	0/7083	0/6973	0/7083	0/6993
Support Vector Machine	0/7014	0/6895	0/7014	0/6847
K-Nearest Neighbors	0/6667	0/6569	0/6667	0/6531
Decision Tree	0/6319	0/6222	0/6319	0/6252
AdaBoost	0/5972	0/5961	0/5972	0/5949
Logistic Regression	0/5139	0/5131	0/5139	0/5094
Naive Bayes	0/5347	0/5129	0/5347	0/5083

Evaluation Results of the Machine Learning Models. From all the above models, the CatBoost model presented the best performance results with an accuracy equal to 0.7708, precision equal to 0.7688, recall equal to 0.7708 and F1-score equal to 0/7519. These results show that CatBoost not only predicts accurately, but have a good mean for all metrics. This is because of its strong architecture for processing categorical data and its automatic hyperparameter tuning. Second only to CatBoost, the Extra Trees model achieved an accuracy score of 0/7569 and an F1-score of 0/7400. Through a series of randomized decision trees, this model provided a somewhat good performance and outperformed other models, such as LightGBM, Random Forest. Similarly, LightGBM also performed well but produced an accuracy of 0/7292 and an F1-score of 0/7126, highlighting its ability to process complex and high-dimensional data. Random Forest and Gradient Boosting ranked next, achieving accuracies of 0/7222 and 0/7153, respectively. The two models presented balanced trade-off between all metrics but were not able to beat CatBoost and Extra Trees. The XGBoost model followed closely, with an accuracy of 0.7083 and an F1-score of 0.6993, highlighting the competitive nature of Boosting-based algorithms. On the other hand, SVM (accuracy = 0/7014) and KNN (accuracy = 0/6667) exhibited less accuracy in predicting disease stages and hence this concludes their lower efficiency in dealing with complex data processing compared to the Boosting models. Relative to simpler models like Decision Tree and AdaBoost, these models exhibited moderate performance. The Decision Tree performed with an accuracy of 0.6319. Standard decision trees are underfitting models, and their performance is less than ensemble trees (i.e. Random Forest, Extra Trees). The AdaBoost model also performed relatively weakly, with 0/5972 accuracy. Logistic Regression and Naive Bayes performed the worst, respectively. As a result of Logistic Regression (accuracy of 0/5139) and Naive Bayes (accuracy of 0/5347), we could claim that these simple models do not provide the ability to process and predict complex, multidimensional data effectively in this study.

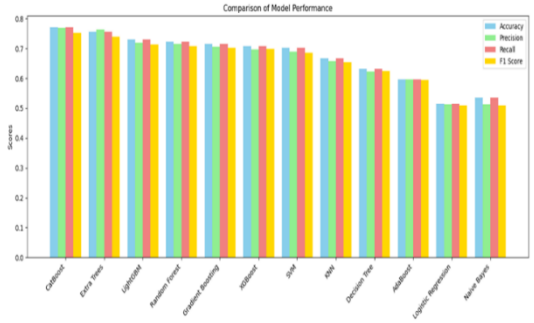

Figure 4. Performance of various machine learning models

In the figure 4, we can see the comparison of various machine learning models by accuracy, precision, recall, and F1-score. Overall, ensemble learning based models like CatBoost, Extra Trees and LightGBM performed the best. The outcomes show that advanced models based on Boosting and ensemble approaches using decision trees excel in performing accurate prediction of disease phases whilst preserving an optimal equilibrium among evaluation metrics compared with alternative models.

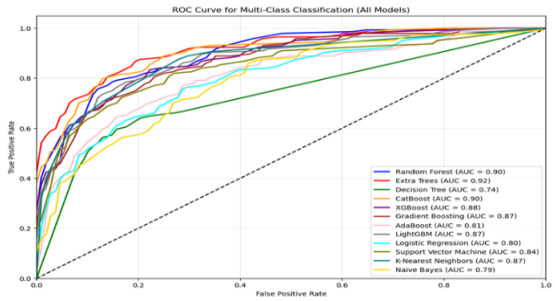

Figure 5. ROC curve

Figure 5 shows ROC curves and AUC for PBC prediction. The performance of models in separating classes is visualized using the ROC curve, whereas the AUC is another robust measure of model performance. If we observe the graph, it is clear that Extra Trees model gave the highest AUC 0/92. The CatBoost and Random Forest both gave AUC 0/90. Gradient Boosting, LightGBM, and SVM also performed distinctively well, attaining AUC values ranging over 0/87 and 0/88. Conversely, simpler models like Decision Tree and Naive Bayes had lower

performance, with AUC of 0/74 and 0/79, respectively. From the results collectively, we see that ensemble-based models, specifically Extra Trees and CatBoost perform better than simple models in class separation. This shows that implementing complex algorithms in highly intricate medical problems, like predicting the progression of diseases, increases the performance of models significantly.

5- Discussion and Conclusion

Results from our study indicate that with the application of modern machine learning algorithms, like CatBoost and Extra Trees, it is possible to obtain accurate predictions of PBC stages. CatBoost was found to be the best of all models achieved, having produced an accuracy of 0/7708 and AUC of 0/90.) Extra Trees also performed well in classifying complex datasets, reaching an AUC of 0/92. These findings underscore the significance of ensemble-based methods in achieving superior predictive accuracy compared to simpler models . This research represents significant advances in machine learning techniques as compared to previous studies. A notable limitation in earlier studies was the use of unoptimized models with poorly defined feature sets. For example, while Hanif and Khan (2022) and Jamadar et al. (2023) employed algorithms such as Random Forest and SVM, they did not utilize advanced optimization techniques to enhance model performance or implement robust feature selection methods.This poor optimization restricted the generalizability and accuracy of their results. As a result, the present study led to stable prediction performance across all metrics by using an automated hyperparameter optimization method (Optuna) and an advanced feature selection method (Recursive Feature Elimination with Cross-Validation). Another key difference in prior studies is their inadequate consideration of imbalanced datasets. When models are evaluated in such manner, it may lead to misleading results because the model can easily predict the majority class while performing poorly on minority classes. For example, Bhardwaj et al. (2024) and Sidana et al. (2022), which did not evaluate models properly and did not point out that a better evaluation is characterized by the reporting of important imbalanced evaluation metrics such as precision, recall, F1-score, etc. This contrast with this study, which used standard performance metrics to give transparent and comprehensive evaluation of model quality. SMOTE process was applied to supporter model to solve imbalance class, while RFECV was used to find out 14 essential features to both reduce model complexity and improve quality. These developments make this study unique compared to previous studies that did not properly resolve dataset imbalance or attempted basic feature selection methodology. Here, we showcase the possibilities of advanced machine learning models and structured optimization techniques in predicting medical health outcomes. Ensemble methods like CatBoost and Extra Trees are better suited for these medical datasets with high-dimension characteristics due to their superior performances compared to simple methods Logistic Regression and Naive Bayes. Such findings provide a direction for future research using larger and diverse data sets having imaging data to create models more accurate with clinical relevance.

Based on the findings of this review, several recommendations are made to enhance and direct future research. The first improvement could be using more and diverse data to provide machine learning models capable of getting generalized. The combination of data from multiple clinical sources with covariate data available in existing datasets could provide more robust results. Secondly, it is proposed that some of the more sophisticated preprocessing methods such as feature engineering and nonlinear transformations might reveal hidden patterns in the data that could improve the model's performance. In future works, DNN (Deep Neural Networks) or LSTM (Long-Short Term Memory) could potentially replace GBDTs with a better prediction performance for the disease stages. More sophisticated ensemble techniques (hybrid Voting and Stacking) are additionally likely to enhance the prediction capabilities due to the synergy of the respective standalone models. On the clinical side, a more detailed analysis of the importance and sensitivity of the model features must facilitate the identification of pertinent biomarkers associated with the prediction of disease stage; each of the findings will assist clinical applications. Finally, validating the above machine learning models against clinical data from hospitals and clinics would make various algorithms appropriate for use as well as more reliable. Initiating these efforts may lead to the development of more accurate and reliable models of timely diagnostics and improved care of patients.

References

[1] M. A. Konerman et al., “Machine learning models to predict disease progression among veterans with hepatitis C virus,” PLOS ONE, vol. 14, no. 1, p. e0208141, Jan. 2019, doi: https://doi.org/10.1371/journal.pone.0208141.

[2] Ahmet Ercan Topcu, Ersin Elbasi, and Yehia Ibrahim Alzoubi, “Machine Learning-Based Analysis and Prediction of Liver Cirrhosis,”Jul.2024,doi:https://doi.org/10.1109/tsp63128.2024.10605929.

[3] E. B. Tapper and N. D. Parikh, “Diagnosis and Management of Cirrhosis and Its Complications: A Review,” JAMA, vol. 329, no. 18, pp. 1589–1602, May 2023, doi: https://doi.org/10.1001/jama.2023.5997.

[4] R. Wei et al., “Clinical prediction of HBV and HCV related hepatic fibrosis using machine learning,” vol. 35, pp. 124–132, Sep. 2018, doi: https://doi.org/10.1016/j.ebiom.2018.07.041.

[5] C. Labenz et al., "Structured Early detection of Asymptomatic Liver Cirrhosis: Results of the population-based liver screening program SEAL," Journal of Hepatology, vol. 77, no. 3,pp.695–701,Sep.2022,doi: https://doi.org/10.1016/j.jhep.2022.04.009.

[6] E. Forte et al., "Top-Down Proteomics Identifies Plasma Proteoform Signatures of Liver Cirrhosis Progression," Molecular & Cellular Proteomics, pp. 100876–100876, Nov. 2024, doi: https://doi.org/10.1016/j.mcpro.2024.100876.

[7] Varshni Premnath and Shanthi Veerappapillai, "Unveiling miRNA–Gene Regulatory Axes as Promising Biomarkers for Liver Cirrhosis and Hepatocellular Carcinoma," ACS Omega, vol. 9, no. 44, pp. 44507–44521, Oct. 2024, doi: https://doi.org/10.1021/acsomega.4c06551.

[8] L. Wang et al., "Impact of Asymptomatic Superior Mesenteric Vein Thrombosis on the Outcomes of Patients with Liver Cirrhosis," Thrombosis and Haemostasis, vol. 122, no. 12, pp. 2019–2029, Sep. 2022, doi: https://doi.org/10.1055/s-0042-1756648.

[9] Md. Nahid Hasan, T. Ahmed, Md. Ashik, Md. Jahid Hasan, Tahaziba Azmin, and J. Uddin, "An Analysis of Covid-19 Pandemic Outbreak on Economy using Neural Network and Random Forest," Journal of Information Systems and Telecommunication (JIST), vol. 11, no. 42, pp. 163–175, Jun. 2023, doi: https://doi.org/10.52547/jist.34246.11.42.163.

[10] Sudiksha Kottachery Kamath, Sanjeev Kushal Pendekanti, and D. Rao, "LivMarX: An Optimized Low-Cost Predictive Model Using Biomarkers for Interpretable Liver Cirrhosis Stage Classification," IEEE Access, vol. 12, pp. 92506–92522,Jan.2024,doi:https://doi.org/10.1109/access.2024.3422451.

[11] I. Hanif and M. M. Khan, "Liver Cirrhosis Prediction using Machine Learning Approaches," 2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), Oct. 2022, doi: https://doi.org/10.1109/uemcon54665.2022.9965718.

[12] D. Bhardwaj, G. Kaur, and G. L. Babu, "Improving Prognostic Prediction of Cirrhosis Using an Optimized Ensemble Machine Learning Approach," pp. 1–6, Aug. 2024, doi: https://doi.org/10.1109/ciscon62171.2024.10695979.

[13] Bhanu Prakash K, Vennela D, Dhana Lakshmi N, and Siva Priyanka S, "Stage Prediction of Liver Cirrhosis Disease using Machine Learning," pp. 1–6, Aug. 2024, doi: https://doi.org/10.1109/icecsp61809.2024.10698096.

[14] Rauf Jamadar, Harsh Uike, and Vaishali Jabade, "Cirrhosis Disease Prediction Using Machine Learning," pp. 515–520, Dec.2023,doi:https://doi.org/10.1109/icacctech61146.2023.00090.

[15] Tejasv Singh Sidana, S. Singhal, S. Gupta, and R. Goel, "Liver Cirrhosis Stage Prediction Using Machine Learning: Multiclass Classification," Lecture notes in networks and systems, pp. 109–129, Nov. 2022, doi: https://doi.org/10.1007/978-981-19-3679-1_9.

[16] Arif Mudi Priyatno and Triyanna Widiyaningtyas, "A SYSTEMATIC LITERATURE REVIEW: RECURSIVE FEATURE ELIMINATION ALGORITHMS," JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer), vol. 9, no. 2, pp. 196–207,Feb.2024,doi: https://doi.org/10.33480/jitk.v9i2.5015.

[17] S. I. Khan and A. S. M. L. Hoque, "SICE: an improved missing data imputation technique," Journal of Big Data, vol. 7, no. 1, Jun. 2020, doi: https://doi.org/10.1186/s40537-020-00313-w.

[18] S. Jeganathan, A. R. Lakshminarayanan, S. Parthasarathy, A. Abdul Azeez Khan, and K. J. Sathick, "OptCatB: Optuna Hyperparameter Optimization Model to Forecast the Educational Proficiency of Immigrant Students based on CatBoost Regression," Journal of Internet Services and Information Security, vol. 14, no. 3, pp. 111–132, Aug. 2024, doi: https://doi.org/10.58346/jisis.2024.i2.008.

[19] F. Fazel and B. Foing, "Evaluating Classification Algorithms: Exoplanet Detection using Kepler Time Series Data," arXiv (CornellUniversity),Feb.2024,doi: https://doi.org/10.48550/arxiv.2402.15874.

[20] Fedesoriano, "Cirrhosis Prediction Dataset," www.kaggle.com.https://www.kaggle.com/fedesoriano/cirrhosis-prediction-dataset

[21] V. Thambawita et al., "An Extensive Study on Cross-Dataset Bias and Evaluation Metrics Interpretation for Machine Learning Applied to Gastrointestinal Tract Abnormality Classification," ACM Transactions on Computing for Healthcare, vol. 1, no. 3, pp. 1–29, Jul. 2020, doi: https://doi.org/10.1145/3386295.

[22] P. J. Muhammad Ali, "Investigating the Impact of Min-Max Data Normalization on the Regression Performance of K-Nearest Neighbor with Different Similarity Measurements," ARO-THE SCIENTIFIC JOURNAL OF KOYA UNIVERSITY, vol. 10, no. 1, pp. 85–91, Jun. 2022, doi: https://doi.org/10.14500/aro.10955.

[23] K. K, U. K, S. A, and A. Kumar, "Predicting Student Performance for Early Intervention using Classification Algorithms in Machine Learning," Journal of Information Systems and Telecommunication (JIST), vol. 9, no. 36, pp. 226–235,Oct.2021,doi: https://doi.org/10.52547/jist.9.36.226.

[24] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," arXiv (Cornell University), Jul. 2019, doi: https://doi.org/10.48550/arxiv.1907.10902.

[25] I. D. Mienye and N. Jere, "A Survey of Decision Trees: Concepts, Algorithms, and Applications," IEEE access, pp. 1–1,Jan.2024,doi: https://doi.org/10.1109/access.2024.3416838.

[26] A. Jafarnejad, A. Rezasoltani, and A. M. Khani, "Comparative Analysis of Machine Learning Algorithms in Predicting Jumps in Stock Closing Price: Case Study of Iran Khodro Using NearMiss and SMOTE Approaches," Iranian Journal of Finance, vol. 9, no. 3, pp. 27–54, 2025, doi: 10.30699/ijf.2025.491324.1496.

[27] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," Machine Learning, vol. 63, no. 1, pp. 3–42, Mar. 2006, doi: https://doi.org/10.1007/s10994-006-6226-1.

[28] G. Biau and B. Cadre, "Optimization by gradient boosting," arXiv.org, Jul. 17, 2017. https://arxiv.org/abs/1707.05023 (accessed Apr. 24, 2024).

[29] Y. Ding, H. Zhu, R. Chen, and R. Li, "An Efficient AdaBoost Algorithm with the Multiple Thresholds Classification," Applied Sciences, vol. 12, no. 12, p. 5872, Jun. 2022, doi: https://doi.org/10.3390/app12125872.

[29] C. Starbuck, "Logistic regression," in *Springer eBooks*, pp. 223–238, 2023. doi: 10.1007/978-3-031-28674-2_12.

[30] A. Jafarnejad Chaghoshi, A. Rezasoltani, and A. M. Khani, "Unleashing the Power of Ensemble Learning: Predicting

National Ranks in Iran's University Entrance Examination," Industrial Management Journal, vol. 16, no. 3, pp. 457–481, 2024, doi: 10.22059/imj.2024.381521.1008178.

[31] G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," hal.science, Dec. 04, 2017. https://hal.science/hal-03953007 (accessed Mar. 27, 2023).

[32] A. V. Dorogush, V. Ershov, and A. Gulin, "CatBoost: gradient boosting with categorical features support," arXiv.org, Oct. 24, 2018. https://arxiv.org/abs/1810.11363

[33] Motiei, M., Khani, A. M., & Beyrami, S. (2021). The effect of green supply chain and green human resource management on environmental performance: The mediating role of green innovation. Logistics Thought, 20(77), 165–197. https://doi.org/10.22034/lot.2021.96691

[34] A. Jafarnjad, A. Rezasoltani, and A. M. Khani, "Analyzing and Predicting Hiring Decisions Using Machine Learning and Deep Learning," Journal of Public Administration, vol. 17, no. 2, pp. 295–327, 2025, doi: 10.22059/jipa.2025.390322.3649.

[35] Jafarnejad Chaghoshi, A., Khani, A. M., & Rezasoltani, A. (2024). Risk modeling in banking services for the blind using fuzzy FMEA and graph neural network (GNN). Journal of Industrial Management Perspective, 14(4), 223–255. https://doi.org/10.48308/jimp.14.4.223

[36] P.J.Beslin Pajila, B. Gracelin. Sheena, A. Gayathri, J. Aswini, M. Nalini, and Siva Subramanian R, "A Comprehensive Survey on Naive Bayes Algorithm: Advantages, Limitations and Applications," Sep. 2023, doi: https://doi.org/10.1109/icosec58147.2023.10276274.

[37] J. Kasubi, M. D. Huchaiah, I. Gad, and M. K. Hooshmand, "A Comparison Analysis of Conventional Classifiers and Deep Learning Model for Activity Recognition in Smart Homes based on Multi-label Classification," Journal of Information Systems and Telecommunication (JIST), vol. 12, no. 46, pp. 127–137,Jun.2024,doi: https://doi.org/10.61186/jist.36294.12.46.127.

[38] A. Rezasoltani, A. Jafarnejad, and A. M. Khani, "A voting-based hybrid machine learning model for predicting backorders in the supply chain," Journal of Decisions and Operations Research, vol. 10, no. 1, pp. 194–213, 2025, doi: 10.22105/dmor.2025.511401.1924.