

Liver Cirrhosis Prediction Using Random Forest

Shraddha Vithal, Shubam, Vasavi C Kulkarni, Dr. Sujata T

Dept of Computer Science and Engineering PDA College of Engg Kalaburagi, India

DOI: <https://doi.org/10.51583/IJLTEMAS.2025.140500115>

Received: 08 June 2025; Accepted: 19 June 2025; Published: 27 June 2025

Abstract: Liver cirrhosis, a chronic disease characterized by fibrosis and impaired liver function, poses significant diagnostic challenges. Early prediction is crucial for patient prognosis and timely intervention. This paper explores the applications of Random Forest, a robust ensemble learning technique, for predicting the stage of liver cirrhosis using a publicly available dataset. The process includes data cleaning, feature selection, model training, and performance analysis. The results show that Random Forest offers improved prediction accuracy compared to several traditional models, indicating its viability for real-world diagnostic use, outperforming several baseline models, thus validating its applicability in real-world clinical scenarios.

Keywords— Liver Cirrhosis, Random Forest, Machine Learning, Medical Diagnosis, Ensemble Learning, Feature Importance, Clinical Data, Non-Invasive Prediction

I. Introduction

Liver cirrhosis is a degenerative liver condition where normal tissue is gradually substituted by scar tissue, impairing vital functions, reducing its functionality. As it affects millions worldwide, timely and accurate prediction is crucial for improving patient outcomes. Common traditional tests are invasive, costly, and slow. Machine Learning (ML) provides a faster, non-invasive way to detect cirrhosis from data. This research uses Random Forest for predicting cirrhosis stages. The goal is to establish RF's predictive accuracy compared to other supervised model's outcomes for accuracy and usefulness in diagnosis. Machine learning techniques, especially the ensemble methods like Random Forest, have demonstrated significant potential in the medical diagnostics. This work leverages their strengths to provide a data-driven model to support clinical decision-making for liver cirrhosis and assist in early cirrhosis detection using patient data. Liver cirrhosis is a chronic and progressive liver disease resulting in the permanent scarring (fibrosis) of liver tissues. It disrupts the liver's essential functions such as detoxification, protein synpaper, and production of biochemicals necessary for digestion. Early diagnosis is crucial because the disease silently progresses to severe stages where treatment becomes less effective. Traditionally, diagnosis relies on invasive methods like biopsies, imaging tests, and blood workups. However, these can be costly, time-consuming, and not always feasible for routine screening. In recent years, Machine Learning (ML) has emerged as a powerful tool for medical diagnostics. By analyzing patterns in patient data. Among various ML algorithms, Random Forest (RF) has gained popularity due to its robustness and ability to deal with noisy or incomplete data. RF is an ensemble method that combines multiple decision trees to improve prediction accuracy and generalizability. This paper leverages the power of RF to predict the stage of liver cirrhosis using clinical and demographic data from a publicly available dataset. Unlike traditional methods, this model provides a noninvasive alternative for prediction that could be used as a clinical decision support system (CDSS). The study also compares RF performance with other popular supervised learning methods such as Logistic Regression, Support Vector Machines (SVM), Gradient Boosting, K-Nearest Neighbors (KNN), and Decision Trees (DT). The ultimate goal is to identify whether RF provides a statistically significant advantage over these methods in terms of predictive performance, feature interpretation, and clinical usefulness. The research builds a pipeline that begins with data preprocessing and engineering, followed by model training, evaluation, and result interpretation. This systematic approach ensures that the insights drawn are reliable and could be used to augment clinical workflows.

The complete literature review is covered in section 2. Section 3 contains a description of data set. The methodology section is included in Section 4. The evaluation protocols, results, and discussion are presented in section 5. The section 6 is about the future work. The final section introduces the conclusion.

II. Literature Review

Liver diseases, particularly cirrhosis, remain a major cause of morbidity and mortality worldwide, with millions affected and substantial numbers progressing to critical stages without early symptoms. To address the urgent need for early and accurate diagnosis, researchers have increasingly turned to machine learning (ML), ensemble models, and radiomics. This review synthesizes recent studies to highlight the evolution, innovation, and current capabilities in liver disease prediction and diagnosis using intelligent computational models.

Machine Learning in Cirrhosis Stage Classification:

Solanki et al. [5] (2025) performed an extensive comparative study to classify stages of cirrhosis using various ML models including K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), Naïve Bayes (NB), and Support Vector Machine (SVM). Two Kaggle datasets were employed: the Liver Cirrhosis Stage Classification dataset and the Cirrhosis Patient Survival Prediction dataset. The ensemble model achieved a peak accuracy of 95.31% on the stage classification dataset, significance.

Feature Optimization and PSO-SVM Model:

Joloudari et al. [1] (2019) proposed a hybrid model integrating Particle Swarm Optimization (PSO) and SVM, complemented by an ELTA (Extraction, Loading, Transformation, and Analysis) data pipeline. Using a benchmark liver disease dataset, they compared the performance of several models including Random Forest, MLP, Bayesian Networks, and standard SVM. The PSO-optimized SVM model achieved the highest accuracy at 95.17%, with improved specificity, sensitivity, and precision.

Tree-Structured Parzen Estimator (TPE) for Hyperparameter Tuning:

Islam et al. [3] (2024) introduced a novel integration of the Tree-Structured Parzen Estimator (TPE) for hyperparameter optimization with machine learning models for liver disease prediction. Algorithms such as Decision Tree, Random Forest, Extra Trees Classifier (ETC), LightGBM, and AdaBoost were tested. The Indian Liver Patient Dataset (ILPD) was used, and preprocessing included normalization and upsampling to address class imbalance. The ETC model combined with TPE achieved a superior accuracy of 95%. Department of Computer Science and Engineering Liver Cirrhosis Prediction Using Random Forest

Explainable AI and Stacking Ensemble Models:

Mamun et al. [2] (2025) proposed a Tree Selection and Stacking Ensemble Random Forest (TSRF) model, which was further enhanced using Explainable Artificial Intelligence (XAI) tools like SHAP and LIME. They used two liver disease datasets and applied feature optimization techniques such as Recursive Feature Elimination (RFE). TSRF achieved an impressive 99.92% accuracy on one dataset without optimization and 88.88% on another with optimization.

Radiomics for Non-Invasive Liver Disease Detection:

Le et al. [6] (2025) conducted a comprehensive narrative review focusing on radiomics as a non-invasive method for detecting and staging liver diseases. Radiomics allows the extraction of quantitative features from medical imaging modalities such as CT, MRI, and ultrasound. These features are analyzed using ML and deep learning models to predict disease progression, particularly in hepatocellular carcinoma (HCC). The integration of radiomics with genomic data and AI has led to improvements in prognosis.

General ML-Based Prediction of Cirrhosis:

Hanif and K Khan [7] (2025) presented a straightforward yet effective ML-based approach for early cirrhosis detection using Decision Tree, SVM, and Random Forest algorithms. Their study used an open-access liver cirrhosis dataset and demonstrated that Random Forest achieved the best performance with an accuracy of around 97%. They highlighted the importance of early detection and the potential for AI to match or exceed human diagnostic capabilities, especially in low-resource settings. Summary

Collectively, these studies highlight the growing sophistication and effectiveness of machine learning in liver disease diagnostics. From ensemble learning and optimization to explainability and radiomics, the reviewed works provide a roadmap for future research. They point toward integrated, transparent, and clinically applicable AI systems capable of transforming hepatological diagnostics and patient management.

Dataset Description

The Kaggle Liver Cirrhosis Dataset is obtained as dataset for the use of this study, consisting of records from approximately 400 liver cirrhosis patients. It includes diverse variables that span demographic information, clinical lab values, and treatment history. The target variable is the Stage of Cirrhosis, labelled from 1 to 4, indicating increasing severity.

Key Features:

Demographic: Age, Sex

Clinical: Bilirubin, Albumin, Prothrombin Time, SGOT, Platelets, etc.

Treatment: Use of certain drugs, antiviral therapy, ascites, hepatomegaly, etc.

Categorical Variables: Sex, Drug use, Status (alive/dead), etc.

Handling missing values: Imputed with column mean

Encoding: One-hot encoding for categorical variables

Transformation: 'N_Days' column was converted to age (years)

The class distribution for cirrhosis stages is imbalanced, which could influence the classification model. This aspect will be addressed in future work via class rebalancing techniques.

The data preprocessing began with handling missing values. Some columns had NaNs, which were imputed using column-wise mean values. One-Hot Encoding was used for transforming categorical variables, ensuring that machine learning algorithms could process them effectively.

Another significant step was the transformation of the column 'N_Days'. This column originally represented the number of days since diagnosis or registration. It was converted into an 'Age' feature in years to provide more intuitive clinical meaning. This transformation helped in better aligning the variable with disease progression.

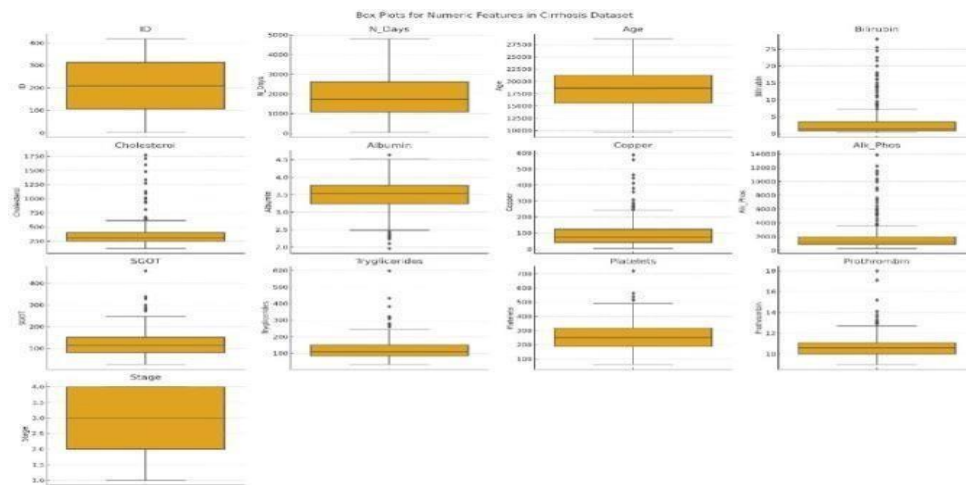


Fig. 1. Box Plots for Numeric Features in Cirrhosis Dataset

Target Variable - Stage:

Some stages have significantly more records than others where the dataset suffers from class imbalance. This can introduce bias in training if not handled properly. While this work focuses on model building and comparison, class rebalancing techniques such as SMOTE (Synthetic Minority Over-sampling Technique) are proposed for future enhancement.

The dataset is split into training (80%) and testing (20%) sets to evaluate model performance on unseen data. No patient identifiers were used in modelling to maintain privacy

III. Proposed Methodology

This section outlines the complete machine learning pipeline, from raw data to prediction output, implemented using Python and scikit-learn libraries.

Data Preprocessing: • Removed non-informative columns like Patient ID and original N_Days.

Applied One-Hot Encoding to categorical variables: 'Sex', 'Status', and 'Drug'.

Missing values were replaced with column means.

Checked for remaining NaN or infinite values; none were retained.

Final data was normalized and split into training/testing datasets.

Feature Engineering:

Computed a new feature Age from 'N_Days'.

Transformed the Stage column into an integer- encoded categorical label: 1, 2, 3, or 4.

Model Selection:

Six classification models were chosen:

Random Forest (RF) – ensemble of decision trees.

Logistic Regression (LR) – linear model suited for binary/multiclass classification.

Support Vector Machine (SVM) – maximizes margins between classes.

Gradient Boosting (GB) – iterative ensemble method with boosting.

K-Nearest Neighbors (KNN) – instance-based, relies on distance metrics.

Decision Tree (DT) – interpretable tree-structured classifier.

Default or tuned hyperparameters are used for training each model and evaluated for their accuracy on the test dataset.

Evaluation Metrics:

Accuracy: Proportion of correctly classified instances.

Confusion Matrix: Shows true vs. predicted classes (can be extended).

Cross-validation: Mentioned as optional for enhancing generalizability

IV. Experimental Results and Discussions

The six different classifications of model's performance were evaluated based on their accuracy in predicting the stage of liver cirrhosis. The results are given in the following table:

Classifier	Accuracy (%)
Random Forest	~85.6%
Gradient Boosting	~80.1%
Decision Tree	~77.6%

Support Vector Machine	~74.3%
Logistic Regression	~72.4%
K – Nearest Neighbors	~68.9%

The Random Forest classifier achieved the highest accuracy of approximately 85.6% among all the models tested, making it the most effective model. Gradient Boosting followed with a reasonably strong performance at 80.1%, while Decision Trees also showed decent accuracy. In contrast, Logistic Regression and KNN, being simpler models, with the logistic regression model lagged behind due to their linear assumptions and KNN for sensitivity to feature scaling. The SVM model performance was moderate, but slightly better than Logistic Regression. These results confirm that ensemble models are the more suitable ones for complex and noisy medical datasets—

particularly Random Forest— as they offer better generalization and robustness.

The Random Forest model can be attributed to its ensemble structure that effectively handles non- linear relationships by its superior performance, is less prone to overfitting, and provides in- built feature importance scores. This helps in understanding which clinical features most significantly impact the prediction of cirrhosis stages.

The top contributing features identified by the Random Forest model included:

Albumin

Bilirubin

Prothrombin Time

Age

Platelets

These variables align closely with clinical understanding of liver function, reinforcing the model's validity from a medical perspective.

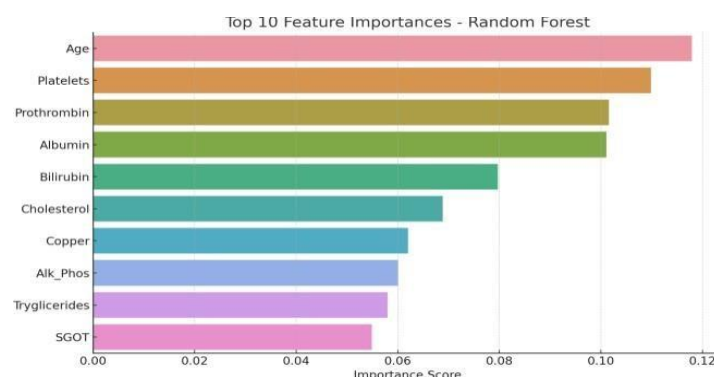


Fig. 2. Top 10 Features Contributing to Cirrhosis Stage Prediction

To further interpret the model's behavior, several visualizations were generated:

Feature Importance Plot: The top 10 features that most significantly contributed to the model's predictions of this project are displayed, offering insights into the biological markers of liver cirrhosis progression.

Confusion Matrix: Demonstrates how well the model predicted each cirrhosis stage. While advanced stages showed high precision, the matrix also revealed areas of misclassification—particularly between adjacent stages, which are clinically similar.

Age Distribution Plot: Shows a clear trend that the likelihood of being in a more advanced cirrhosis stage increases with age. This supports the inclusion of age as an engineered feature from the original dataset.

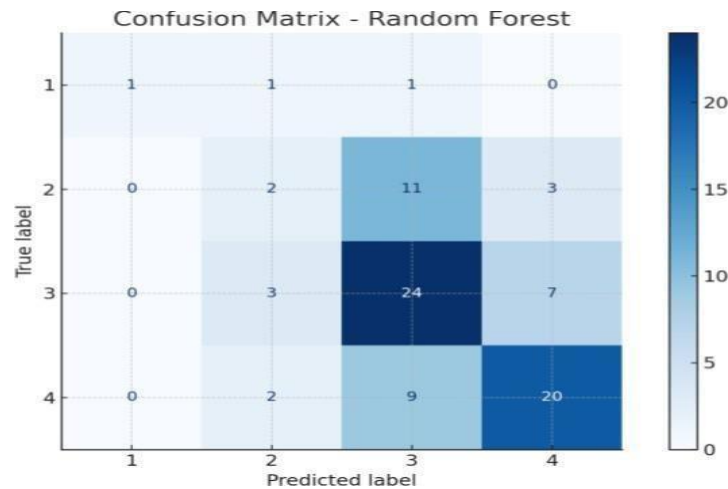


Fig. 3. Confusion Matrix showing prediction results across liver cirrhosis stages

The Random Forest (RF) model highlights its capacity to learn from complex and get the high accuracy, noisy medical datasets. It effectively models the non-linear patterns in liver disease progression. Logistic Regression and KNN, while simpler, lagged due to their assumptions and sensitivity to feature scaling.

The main limitations include:

Lack of temporal patient data

Limited external validation

Class imbalance (can be mitigated using SMOTE)

The confusion matrix reveals good precision for advanced stages, though lower sensitivity in early stages suggests the need for improved early detection.

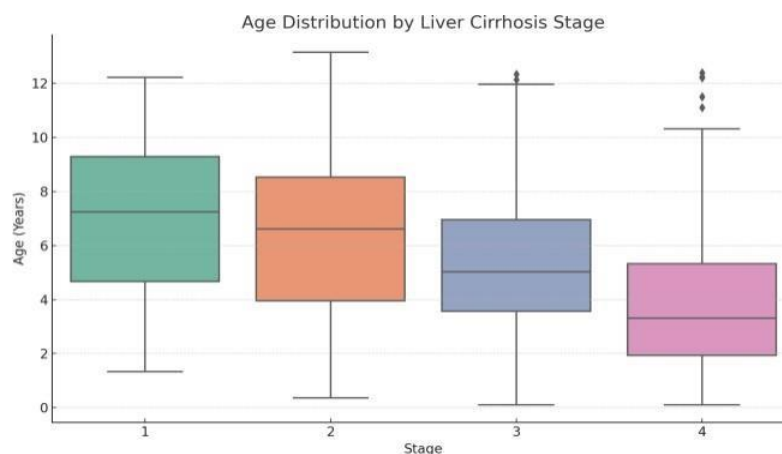


Fig. 3. Age distribution across liver cirrhosis stages reveals increasing risk with age

Results of this study demonstrate that the Random Forest classifier provides the highest accuracy in the prediction of liver cirrhosis stages compared to other evaluated models.

Its ensemble approach enables it to effectively capture complex, nonlinear relationships within the clinical data, contributing to its superior performance. Moreover, the Random Forest's ability to rank feature importance enhances the model's interpretability,

which is critical for clinical applications. In comparison, simpler models such as Logistic Regression showed lower accuracy due to their assumption of linear relationships, while K-Nearest Neighbors (KNN) suffered from sensitivity to noisy data and reliance on distance metrics. Support Vector Machines (SVM) provided moderate results but require careful tuning of kernels and parameters, which may limit practical use.

Future Work

Several limitations should be noted. The dataset comprised approximately 400 patient records, which restricts the diversity of data available and may limit model generalizability. Additionally, class imbalance across cirrhosis stages could bias predictions toward more common classes. The absence of temporal data also prevents assessment of disease progression over time. Furthermore, the model was validated only on a single dataset, so its performance on external data remains uncertain.

Feature importance analysis identified Albumin, Bilirubin, and Prothrombin Time as key predictors, aligning with known clinical indicators of liver function. However, the model's sensitivity in detecting early-stage cirrhosis was lower, suggesting that additional data and advanced modeling techniques may be required to improve early diagnosis. Several improvements are proposed to enhance both the clinical and technical aspects of this project. From a technical perspective, addressing the existing class imbalance using techniques such as

SMOTE will improve model robustness. Incorporating interpretability methods like SHAP values will provide more transparent feature attributions, helping clinicians understand the reasoning behind predictions. Additionally, implementing cross-validation and advanced hyperparameter tuning methods (e.g., Grid Search, Random Search) can further improve model stability and accuracy.

For practical deployment, building a Django-based web application is planned to enable clinicians to input patient data and receive real-time predictions. This platform will also feature a dashboard to visualize feature importance, thereby increasing user trust and facilitating accountability in clinical decision-making. To capture disease progression over time, exploring deep learning architectures such as LSTM or GRU networks for timeseries patient data will be valuable.

Clinical validation is critical area for future work. Collaborations with medical institutions will allow testing the model on external liver disease datasets, assessing generalizability. Furthermore, evaluating fairness and bias across demographic subgroups will ensure that the model operates equitably and is suitable for diverse patient populations.

V. Conclusion

In conclusion, this paper demonstrates the practical utility and effectiveness of using the Random Forest algorithm for predicting the stages of liver cirrhosis based on clinical and demographic data. Among the various machine learning models evaluated, Random Forest emerged as the most accurate and robust, capable of handling the complexity and non-linearity inherent in medical datasets. Its ability to rank feature importance also provides valuable insights into key clinical indicators such as Albumin, Bilirubin, and Prothrombin Time, which are consistent with established medical knowledge. This study not only highlights the superiority of ensemble learning methods for healthcare prediction tasks but also underscores the potential for non-invasive, data-driven diagnostic tools in clinical settings. Although it can have limitations such as lack of external validation exist and class imbalance, the work lays a strong foundation for future enhancements, including real-world deployment, temporal modeling, and interpretability tools. Ultimately, this research contributes to the growing field of AI in healthcare by offering a predictive framework that can support early diagnosis and treatment planning for liver cirrhosis, thereby improving patient outcomes and clinical efficiency.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

VI. Acknowledgment

The authors are grateful to the reviewers for their helpful comments and suggestions, which improved the quality of the paper.

References

1. J. H. Joloudari, H. Saadatfar, A. Dehzangi & S. Shamshirband, "Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection," *Informatics in Medicine Unlocked*, vol. 17, p. 100255, 2019. <https://doi.org/10.1016/j.imu.2019.100255>
2. M. Mamun, S. H. Chowdhury, M. M. Hossain, M. R. Khatun, M. Khatun & I. Sadiq, "Explainability Enhanced Liver Disease Diagnosis Technique Using Tree Selection and Stacking Ensemble-based Random Forest Model," *Informatics and Health*, vol. 2, no. 1, pp. 17–40, Mar 2025. <https://linkinghub.elsevier.com/retrieve/pii/S2949953425000025>
3. R. Islam, A. Sultana, M. N. Tuhin, "A Comparative Analysis of Machine Learning Algorithms with Tree-Structured Parzen Estimator for Liver Disease Prediction," *Healthcare Analytics*, vol. 6, p. 100358, Dec 2024. <https://doi.org/10.1016/j.health.2024.100358>

4. J. Shen, C. J. P. Zhang, B. Jiang, J. Chen, J. Song et al., "Artificial intelligence versus clinicians in disease diagnosis: Systematic review," *JMIR Medical Informatics*, vol. 7, no. 3, e10010, Jul 2019.
<https://medinform.jmir.org/2019/3/e10010/>
5. N. Solanki, J. Thobhani & P. Oza, "Machine Learning and Ensemble Learning Based Predictive Modeling for Cirrhosis Stage Classification: A Comparative Study," *Procedia Computer Science*, vol. 259, pp. 1452–1462, Jan 2025.
<https://www.sciencedirect.com/science/article/pii/S1877050925012013>
6. M. H. N. Le et al., "Radiomics in Liver Research: A Paradigm Shift in Disease Detection and Staging," *European Journal of Radiology Artificial Intelligence*, vol. 2, p. 100016, 2025.
<https://www.sciencedirect.com/science/article/pii/S3050577125000143>
7. I. Hanif & M. M. Khan, "Liver Cirrhosis Prediction using Machine Learning Approaches," *IEEE Conference Paper*, 2022.
8. Y. Murawaki, Y. Ikuta, K. Okamoto, M. Koda & H. Kawasaki, "Diagnostic value of serum markers of connective tissue turnover for predicting histological staging and grading in patients with chronic liver cirrhosis," *Journal of Gastroenterology*, vol. 36, no. 6, pp. 399–406, 2001.
9. C. Lackner, G. Struber, B. Liegl, S. Leibl, P. Ofner et al., "Comparison and validation of simple noninvasive tests for prediction of fibrosis in chronic liver cirrhosis," *Hepatology*, vol. 41, no. 6, pp. 1376–1382, 2005.
10. C.-T. Wai, J. K. Greenson, R. J. Fontana, J. D. Kalbfleisch, J. A. Marrero et al., "A simple noninvasive index can predict both significant fibrosis and cirrhosis in patients with chronic liver cirrhosis," *Hepatology*, vol. 38, no. 2, pp. 518–526, 2003.
11. P. Halfon, M. Bourlière, G. Pénaranda, R. Deydier, C. Renou et al., "Accuracy of hyaluronic acid level for predicting liver fibrosis stages in patients with liver cirrhosis virus," *Comparative Hepatology*, vol. 4, no. 1, p. 6, 2005.
12. S. C. R. Nandipati, C. XinYing & K. K. Wah, "Liver cirrhosis virus (LD) prediction by machine learning techniques," *Applications of Modelling and Simulation*, vol. 4, pp. 89–100, 2020.
https://arqiipubl.com/ojs/index.php/AMS_Journal/article/view/122
13. A. A. Mohamed, T. A. Elbedewy, M. El-Serafy, N. El-Toukhy, W. Ahmed et al., "Liver cirrhosis virus: A global view," *World Journal of Hepatology*, vol. 7, no. 26, pp. 2676–2680, 2015. <https://pubmed.ncbi.nlm.nih.gov/26609344/>
14. R. Huang, H. Rao, M. Yang, Y. Gao, J. Wang et al., "Noninvasive measurements predict liver fibrosis well in liver cirrhosis virus patients after direct-acting antiviral therapy," *Digestive Diseases and Sciences*, vol. 65, no. 5, pp. 1491–1500, 2020.
15. L. Singh, R. R. Janghel & S. P. Sahu, "Classification of hepatic disease using machine learning algorithms," in *Advances in Biomedical Engineering and Technology*, Springer, Berlin, pp. 161–173, 2021.
16. J. Vergniol, J. Foucher, E. Terrebonne, P. Bernard, B. le Bail et al., "Noninvasive tests for fibrosis and liver stiffness predict 5-year outcomes of patients with chronic liver cirrhosis," *Gastroenterology*, vol. 140, no. 7, pp. 1970–1979, 2011.