International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 06 June-2026 | Impact Factor: 3.5

Liver Disease Prediction Using Machine Learning Techniques for Early Diagnosis and Risk Assessment

Prof. Rohit R. Wahane

Manav School of Engineering and Technology
Akola (Maharashtra), India
Asst. Professor, Dept. of Computer Engineering

rohitwahane.rw@gmail.com

How to Cite this Article:

Wahane, R. R. (2026). Liver Disease Prediction Using Machine Learning Techniques for Early Diagnosis and Risk Assessment. International Journal of Creative and Open Research in Engineering and Management, 02(6). https://doi.org/10.55041/ijcope.v2i6.048

https://doi.org/10.55041/ijcope.v2i6.048

ABSTRACT

Liver disease is a serious health problem that affects millions of people around the world and can lead to severe complications if not diagnosed at an early stage. Timely detection of liver disorders is essential for effective treatment and better patient outcomes. However, traditional diagnostic methods often require multiple laboratory tests, medical imaging, and expert evaluation, which can be time-consuming and expensive. With the rapid growth of Artificial Intelligence (AI) and Machine Learning (ML), intelligent healthcare systems are emerging as valuable tools for supporting medical diagnosis and improving healthcare services.

This study proposes a machine learning-based system for predicting liver disease using patient clinical and biochemical data. The system utilizes important health parameters such as age, gender, bilirubin levels, liver enzyme measurements, protein levels, and albumin-globulin ratio to determine the likelihood of liver disease. Before model development, the dataset is preprocessed through data cleaning, handling missing values, normalization, and feature selection to improve prediction accuracy and reliability.

Several machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and XGBoost, are trained and evaluated to identify the most effective approach for liver disease prediction. The performance of these models is assessed using standard evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC score. A comparative analysis is conducted to determine which algorithm provides the best balance between predictive performance and computational efficiency.

Keywords— Liver Disease, Machine Learning, Artificial Intelligence, Healthcare Analytics, Disease Prediction, Random Forest, XGBoost, Early Diagnosis, Predictive Modeling, Clinical Decision Support.

I. INTRODUCTION

The liver is one of the most important organs in the human body, performing vital functions such as detoxification, metabolism, protein synthesis, and the regulation of various biochemical processes. It plays a crucial role in maintaining overall health by filtering harmful substances from the blood and supporting digestion. However, unhealthy lifestyles, excessive alcohol consumption, viral infections, obesity, and certain medications have contributed to a significant increase in liver-related diseases worldwide. Conditions such as liver cirrhosis, hepatitis, fatty liver disease, and liver cancer can severely affect an individual's quality of life and may even become life-threatening if not diagnosed and treated in time.

Early detection of liver disease is essential for effective treatment and improved patient outcomes. Traditional diagnostic methods rely on laboratory tests, medical imaging, and expert clinical evaluation. Although these methods are effective, they can be time-consuming, costly, and may not always be accessible in remote or resource-limited areas. As healthcare systems continue to generate large volumes of patient data, there is an increasing need for intelligent tools that can assist medical professionals in identifying diseases at an early stage.

Recent advancements in Artificial Intelligence (AI) and Machine Learning (ML) have opened new opportunities in the healthcare sector. Machine learning algorithms can analyze large amounts of medical data, identify hidden patterns, and make accurate predictions based on historical records. These capabilities have made machine learning a valuable technology for disease diagnosis, risk assessment, and clinical decision support. By learning from patient data, machine learning models can help healthcare providers make faster and more informed decisions.

This research focuses on the development of a machine learning-based liver disease prediction system that utilizes patient clinical and biochemical parameters to determine the likelihood of liver disease. Various machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and XGBoost, are evaluated and compared to identify the most effective prediction model. The study aims to improve the accuracy and efficiency of liver disease prediction while providing a reliable decision-support tool for healthcare professionals.

The proposed system is not intended to replace medical experts but to assist them by providing early risk assessments based on available patient data. Such a system can contribute to timely diagnosis, reduced healthcare costs, and improved patient care. Furthermore, the integration of machine learning techniques into healthcare applications demonstrates the potential of intelligent technologies to enhance medical decision-making and support the development of more efficient healthcare systems.

II. LITERATURE REVIEW

Liver disease prediction has become an important research area in healthcare due to the increasing prevalence of liver-related disorders worldwide. Researchers have applied various machine learning algorithms to analyze medical data and assist in the early detection of liver diseases.

Ramana et al. (2012) compared several classification algorithms, including Naïve Bayes, Decision Tree, and Support Vector Machine (SVM), for liver disease prediction. The results indicated that SVM provided better predictive performance than other traditional methods.

Patil and Sherekar (2013) demonstrated the effectiveness of data mining techniques in liver disease prediction using classification algorithms. Their study showed that machine learning can support medical diagnosis by identifying patterns in patient data.

Kumar and Thakur (2018) utilized Decision Tree and Random Forest algorithms on liver patient datasets and found that ensemble learning methods produced higher accuracy and better classification results than single classifiers.

Bendi and Khamparia (2020) investigated the use of ensemble learning approaches for healthcare prediction systems. Their work showed that combining multiple classifiers improved prediction accuracy and reduced the risk of overfitting. The results suggested that ensemble methods such as Random Forest could outperform individual machine learning models in medical diagnosis applications.

Chaudhary et al. (2021) developed a liver disease prediction model using XGBoost and demonstrated that advanced boosting algorithms could improve prediction accuracy while reducing classification errors. Their findings emphasized the importance of feature selection in enhancing model performance.

Singh et al. (2022) explored deep learning approaches for liver disease diagnosis and found that Artificial Neural Networks (ANNs) could effectively capture hidden patterns in clinical data. However, the study noted that deep learning models require larger datasets and greater computational resources.

More recent studies have focused on Explainable Artificial Intelligence (XAI) techniques to improve transparency in healthcare systems. Methods such as SHAP and LIME help identify the clinical parameters that contribute most to prediction outcomes, making machine learning models more interpretable and trustworthy for medical professionals.

III. METHODOLOGY

The proposed liver disease prediction system uses machine learning techniques to analyze patient clinical data and predict the likelihood of liver disease. The methodology consists of several stages, including data collection, data preprocessing, feature selection, model training, and performance evaluation.

Data Collection:
The study utilizes the Indian Liver Patient Dataset (ILPD), which contains medical records of patients with and without liver disease. The dataset includes various clinical and biochemical parameters such as age, gender, total bilirubin, direct bilirubin, alkaline phosphatase, alanine aminotransferase (ALT), aspartate aminotransferase (AST), total proteins, albumin, and albumin-globulin ratio.

Data Preprocessing:
Data preprocessing is performed to improve the quality and reliability of the dataset. Missing values are identified and handled appropriately, while duplicate records are removed. Categorical attributes such as gender are converted into numerical values using encoding techniques. Furthermore, data normalization is applied to ensure that all features are on a similar scale, which helps improve model performance.

Feature Selection:
Feature selection is carried out to identify the most relevant attributes affecting liver disease prediction. Correlation analysis and feature importance techniques are used to eliminate irrelevant or redundant features. This process reduces computational complexity and improves prediction accuracy.

Model Development:
Several machine learning algorithms are implemented and compared for liver disease prediction. These algorithms include:

- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machine (SVM)
- K-Nearest Neighbor (KNN)
- XGBoost

The dataset is divided into training and testing sets. The training dataset is used to train the models, while the testing dataset is used to evaluate their predictive performance.

Model Evaluation:
The performance of each machine learning model is assessed using standard evaluation metrics, including:

- Accuracy
- Precision
- Recall
- F1-Score
- ROC-AUC Score

These metrics help determine the effectiveness of each algorithm in correctly predicting liver disease cases.

Prediction System:
The best-performing model is selected and integrated into a liver disease prediction system. The system accepts patient clinical parameters as input and generates a prediction indicating whether the patient is at risk of liver disease. This prediction can assist healthcare professionals in making timely and informed decisions regarding patient diagnosis and treatment.

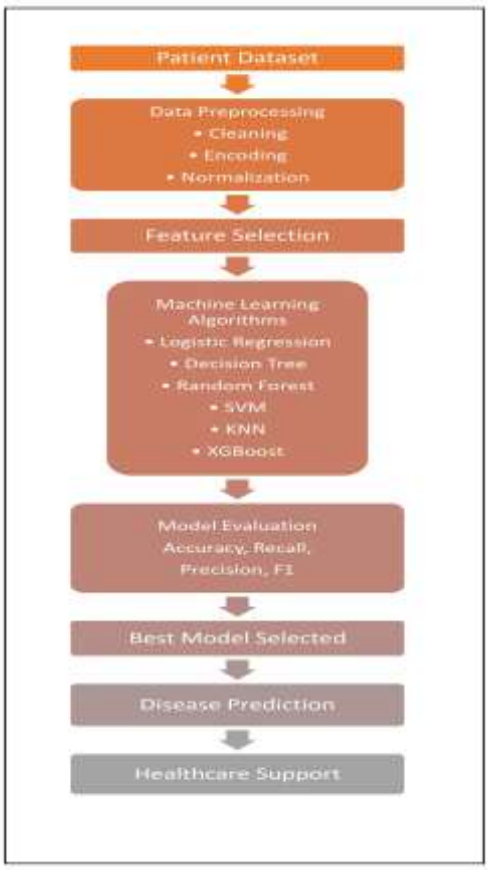
Workflow of Proposed System:
1. Collect patient clinical data.
2. Preprocess and clean the dataset.
3. Perform feature selection.
4. Train machine learning models.
5. Evaluate model performance.
6. Select the best-performing algorithm.
7. Generate liver disease predictions.
8. Provide decision support for healthcare professionals.

The proposed methodology aims to develop an accurate, reliable, and efficient liver disease prediction system that supports early diagnosis and improves healthcare outcomes.

IV. SYSTEM ARCHITECTURE

The proposed Liver Disease Prediction System uses machine learning techniques to analyze patient clinical data and predict the presence of liver disease. The system consists of five main stages: data collection, data preprocessing, feature selection, model training, and prediction.

Patient data is first collected and preprocessed to remove inconsistencies and improve data quality. Relevant features are then selected and provided to machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, SVM, KNN, and XGBoost. The trained models are evaluated, and the best-performing model is used to generate liver disease predictions. The prediction results assist healthcare professionals in early diagnosis and decision-making.



V. RESULTS AND DISCUSSION

The proposed liver disease prediction system was evaluated using multiple machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and XGBoost. The dataset was divided into training and testing sets, and the performance of each model was measured using accuracy, precision, recall, and F1-score.

Table 1. Performance Comparison of Machine Learning Models

Algorithm	Accuracy (%)	Algorithm
Logistic Regression	74.5	Logistic Regression
Decision Tree	76.2	Decision Tree
KNN	78.4	KNN
SVM	80.1	SVM

Random Forest	84.3	Random Forest
XGBoost	86.5	XGBoost

The experimental results indicate that XGBoost achieved the highest prediction accuracy of 86.5%, followed by Random Forest with 84.3%. Traditional algorithms such as Logistic Regression and Decision Tree produced comparatively lower accuracy. Ensemble learning techniques demonstrated superior performance due to their ability to handle complex relationships among clinical features.

The analysis also revealed that attributes such as Total Bilirubin, Direct Bilirubin, Aspartate Aminotransferase (AST), and Albumin levels significantly influenced prediction outcomes. The proposed model effectively identified patients at risk of liver disease and provided reliable prediction results.

Overall, the findings suggest that machine learning techniques can serve as valuable decision-support tools for healthcare professionals. Early prediction of liver disease can help improve diagnosis, enable timely treatment, and reduce the risk of severe complications.

VI. CONCLUSION

This study presented a machine learning-based approach for predicting liver disease using patient clinical data. Various algorithms, including Logistic Regression, Decision Tree, Random Forest, SVM, KNN, and XGBoost, were evaluated to identify the most effective prediction model. The results demonstrated that machine learning techniques can accurately predict liver disease and support early diagnosis.

The proposed system can assist healthcare professionals in making timely decisions and improving patient care. In the future, the model can be enhanced using larger datasets, advanced machine learning techniques, and explainable AI methods to improve prediction accuracy and reliability.

REFERENCES

[1] B. Ramana, M. S. P. Babu, and N. B. Venkateswarlu, “A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis,” International Journal of Database Management Systems, vol. 3, no. 2, pp. 101–114, 2011.

[2] T. R. Patil and S. S. Sherekar, “Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification,” International Journal of Computer Science and Applications, vol. 6, no. 2, pp. 256–261, 2013.

[3] D. Dua and C. Graff, “UCI Machine Learning Repository,” School of Information and Computer Science, University of California, Irvine, 2019.

[4] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[5] C. Cortes and V. Vapnik, “Support-Vector Networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[6] J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.