International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI:10.55041/ISJEM05935
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

An Ensemble Machine Learning Approach for Early Liver Disease Prediction

Naga Lakshmi Korrapati 1, Gangavarapu Nagireddy2, Sirigiri Karthik Kumar 3 , Torlakonda Haribabu 4 , Dr Shaik MD Rafi 5

[1] B-Tech STUDENT in the Department OF IT & SRI MITTAPALLI COLLEGE OF ENGINEERING
[2] B-Tech STUDENT in the Department OF IT & SRI MITTAPALLI COLLEGE OF ENGINEERING
[3] B-Tech STUDENT in the Department OF IT & SRI MITTAPALLI COLLEGE OF ENGINEERING
[4] B-Tech STUDENT in the Department OF IT & SRI MITTAPALLI COLLEGE OF ENGINEERING
[5] Professor in the Department of Computer Science and Engineering

---

Abstract - Liver disease is a critical global health issue that often progresses without noticeable symptoms, making early detection essential for effective treatment and prevention. This study presents a machine learning-based approach for the prediction of liver disease using clinical and biochemical parameters. The dataset used includes various attributes such as age, gender, bilirubin levels, liver enzyme measurements, and protein levels, which are relevant indicators of liver function.

In this work, several machine learning algorithms, including Logistic Regression, Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest, were implemented and evaluated. Data preprocessing techniques such as handling missing values, feature scaling, and class balancing using Synthetic Minority Oversampling Technique (SMOTE) were applied to improve model performance. Feature selection methods were also used to identify the most significant predictors of liver disease.

The experimental results demonstrate that the Random Forest model outperforms other algorithms, achieving the highest accuracy and overall performance metrics. The findings highlight the effectiveness of machine learning techniques in analyzing medical data and providing reliable predictions. This approach can assist healthcare professionals in early diagnosis and decision-making, ultimately improving patient outcomes.

In conclusion, the proposed system offers a cost-effective and efficient solution for liver disease prediction. Future work may involve the use of larger datasets, deep learning models, and explainable artificial intelligence techniques to enhance prediction accuracy and interpretability.

Key Words: Liver disease, Machine learning, Random Forest, logistic regression.

1.INTRODUCTION (Size 11, Times New roman)

Liver disease represents a significant global health challenge, contributing to substantial morbidity and mortality worldwide. Conditions such as cirrhosis, hepatitis, fatty liver disease, and hepatocellular carcinoma often progress silently, with symptoms appearing only in advanced stages. Early detection and accurate diagnosis are therefore critical for improving patient outcomes and reducing healthcare burdens. However, traditional diagnostic approaches—relying on clinical expertise, laboratory tests, and imaging—can be time-consuming, costly, and sometimes prone to human error.

In recent years, the rapid advancement of **machine learning (ML)** techniques has opened new avenues for improving disease prediction and diagnosis. Machine learning algorithms are capable of analyzing large volumes of medical data, identifying hidden patterns, and making data-driven predictions with high accuracy. In the context of liver disease, ML models can utilize patient data such as age, gender, biochemical markers, and lifestyle factors to assist in early detection and risk assessment.

Several studies have demonstrated the effectiveness of various supervised learning algorithms, including decision trees, support vector machines (SVM), k-nearest neighbors (KNN), and ensemble methods, in predicting liver disease. These approaches not only enhance diagnostic accuracy but also support clinicians in making informed decisions. Furthermore, with the increasing availability of healthcare datasets and electronic medical records, the potential for developing robust predictive models continues to grow.

This research focuses on developing a machine learning-based model for liver disease prediction using clinical and biochemical data. The study aims to evaluate and compare the performance of different algorithms, optimize model accuracy, and identify the most significant features contributing to liver disease. By leveraging advanced

computational techniques, this work seeks to provide an efficient, reliable, and scalable solution for early diagnosis, ultimately improving patient care and reducing the burden on healthcare systems.

2. Literature Review

Liver disease remains a major global health concern, accounting for millions of deaths annually and posing a significant burden on healthcare systems. Early detection is particularly challenging due to the asymptomatic nature of many liver disorders in their initial stages. Traditional diagnostic approaches—such as liver function tests, imaging techniques, and biopsies—are often expensive, time-consuming, and require expert interpretation. Consequently, there has been growing interest in leveraging machine learning (ML) techniques to enable early, accurate, and cost-effective prediction of liver diseases.

2.1. Traditional Machine Learning Approaches

Initial research in liver disease prediction primarily focused on supervised machine learning algorithms such as decision trees, naïve Bayes, logistic regression, and support vector machines (SVM). These models were applied to structured clinical datasets, particularly the Indian Liver Patient Dataset (ILPD), to classify patients as healthy or diseased. Studies demonstrated that ML models could effectively uncover hidden patterns in biochemical attributes such as bilirubin levels, enzyme counts, and protein concentrations.

Comparative analyses have shown that no single algorithm consistently outperforms others across all datasets. However, decision trees and SVMs often achieved moderate to high accuracy, making them suitable for early-stage diagnostic systems. These early works established the feasibility of ML-based diagnosis and laid the foundation for more advanced techniques.

2.2. Ensemble Learning and Hybrid Models

With advancements in computational capabilities, researchers began exploring ensemble learning techniques to improve predictive performance. Ensemble methods, such as Random Forest, Gradient Boosting, and Extreme Gradient Boosting (XGBoost), combine multiple base learners to enhance accuracy and robustness.

Recent studies highlight the superiority of ensemble approaches over individual models. For example, Ganie et al. (2024) emphasized that ensemble learning significantly improves prediction accuracy by addressing data complexity and variability in clinical datasets . Similarly, Sekkat et al. (2025) optimized ensemble models and demonstrated their effectiveness in handling diverse liver disease conditions, including cirrhosis and fatty liver disease .

More recently, hybrid models integrating multiple ML and deep learning techniques have shown promising results. Dash et al. (2026) proposed a hybrid framework combining multilayer perceptron neural networks with ensemble classifiers such as XGBoost and LightGBM, achieving an accuracy of 95.49% . These hybrid approaches leverage the strengths of different algorithms and address limitations such as overfitting and class imbalance.

2.3. Deep Learning and Medical Imaging

The emergence of deep learning has significantly enhanced the capabilities of liver disease prediction systems, particularly in medical imaging. Convolutional Neural Networks (CNNs) and other deep architectures are widely used to analyze ultrasound, CT, and MRI images for detecting liver abnormalities.

A comprehensive survey by Kodinariya and Gondaliya (2023) highlights that deep learning models are particularly effective in diagnosing liver cancer, fatty liver, and cirrhosis using imaging modalities . These models automate feature extraction and improve diagnostic precision compared to traditional image-processing techniques.

In addition to imaging, innovative approaches have explored alternative data sources. For instance, recent studies demonstrate the feasibility of using electrocardiogram (ECG) data for liver disease diagnosis through ML models, providing a non-invasive and scalable diagnostic solution. These approaches reveal hidden physiological relationships between cardiac and hepatic systems, expanding the scope of ML applications in healthcare.

2.4. Early Prediction and Risk Assessment Models

A significant focus of recent research has been on early prediction and risk assessment of liver disease progression. Machine learning models trained on electronic health records (EHRs) can predict diseases such as cirrhosis years before clinical diagnosis.

Miao et al. (2026) developed an XGBoost-based model that predicts liver cirrhosis up to three years in advance, outperforming traditional clinical scoring systems such as the FIB-4 index. The study reported Area Under the Curve (AUC) values of up to 0.81, demonstrating improved predictive capability and early risk stratification.

Similarly, studies on metabolic dysfunction-associated steatotic liver disease (MASLD) have explored interpretable models such as LASSO regression and random forests. These models achieve competitive performance while maintaining transparency, which is crucial for clinical adoption .

2.5. Feature Engineering and Data Challenges

Feature engineering plays a critical role in improving model performance. Researchers have introduced domain-specific features, such as enzyme ratios and derived biomarkers, to enhance predictive accuracy. Additionally, techniques like Synthetic Minority Oversampling Technique (SMOTE) are widely used to address class imbalance in medical datasets.

Despite these advancements, several challenges persist. Many studies rely on relatively small or imbalanced datasets, which limits generalizability. Data quality issues, such as missing values and noise, also affect model performance. Furthermore, variability in patient demographics and clinical conditions makes it difficult to develop universally applicable models.

2.6. Explainable AI and Clinical Integration

One of the major limitations of advanced ML and deep learning models is their lack of interpretability. To address this issue, recent research has focused on Explainable Artificial Intelligence (XAI) techniques such as SHAP (SHapley Additive explanations) and LIME (Local Interpretable Model-Agnostic Explanations).

For example, the StackLiverNet model integrates SHAP and LIME to provide interpretable predictions, helping clinicians understand the contribution of features such as alkaline phosphatase and SGOT levels in disease detection. These advancements are crucial for building trust and facilitating the adoption of ML models in real-world healthcare settings

2.7. Research Gaps and Future Directions

Although significant progress has been made, several gaps remain in the literature:

- **Generalization Issues:** Many models are trained on limited datasets and may not perform well across diverse populations.
- **Data Privacy and Ethics:** Access to high-quality medical data is often restricted due to privacy concerns.
- **Model Interpretability:** Complex models still lack sufficient transparency for clinical decision-making.
- **Integration Challenges:** Incorporating ML systems into existing healthcare workflows remains a challenge.

Future research should focus on developing robust, interpretable, and scalable models using large, diverse datasets. Additionally, integrating multimodal data (clinical, imaging, and genomic) could further enhance predictive performance.

3. Methodology

This section describes the systematic approach followed to develop and evaluate a machine learning model for liver disease prediction. The methodology includes data collection, preprocessing, feature selection, model development, and performance evaluation.

3.1. Data Collection

The dataset used in this study is the **Indian Liver Patient Dataset (ILPD)**, which is publicly available from the UCI Machine Learning Repository. It contains medical records of patients, including both liver disease patients and healthy individuals. The dataset consists of several clinical and biochemical attributes such as age, gender, total bilirubin, direct bilirubin, alkaline phosphatase, alanine aminotransferase (SGPT), aspartate aminotransferase (SGOT), total proteins, albumin, and albumin-globulin ratio.

3.2. Data Preprocessing

Data preprocessing is a crucial step to ensure the quality and reliability of the dataset before model training. The following preprocessing steps were applied:

- **Handling Missing Values:** Missing values in the dataset were identified and handled using techniques such as mean or median imputation.
- **Data Cleaning:** Inconsistent or duplicate records were removed to maintain data integrity.
- **Encoding Categorical Variables:** The gender attribute was converted into numerical format using label encoding.
- **Feature Scaling:** Numerical features were normalized or standardized using techniques such as Min-Max scaling or StandardScaler to ensure uniformity across features.
- **Class Imbalance Handling:** Since the dataset is imbalanced, techniques such as Synthetic Minority Oversampling Technique (SMOTE) were applied to balance the classes.

3.3. Feature Selection

Feature selection was performed to identify the most relevant attributes contributing to liver disease prediction. This helps in improving model performance and reducing computational complexity. Techniques such as:

- Correlation analysis
- Recursive Feature Elimination (RFE)
- Feature importance from tree-based models

were used to select significant features like bilirubin levels, enzyme measurements (SGOT, SGPT), and protein levels.

3.4. Model Development

Several machine learning algorithms were implemented and compared to determine the most effective model for liver disease prediction. The models used in this study include:

- **Logistic Regression** – for baseline classification
- **Decision Tree Classifier** – for interpretable predictions
- **Support Vector Machine (SVM)** – for handling high-dimensional data
- **K-Nearest Neighbors (KNN)** – for instance-based learning
- **Random Forest** – for improved accuracy using ensemble learning

The dataset was divided into training and testing sets, typically in a 70:30 or 80:20 ratio. Cross-validation techniques were applied to ensure model generalization and prevent overfitting.

3.5. Model Training and Optimization

Each model was trained using the training dataset, and hyperparameter tuning was performed to optimize performance. Techniques such as Grid Search and Random Search were used to find the best combination of parameters for each algorithm.

3.6. Performance Evaluation

The performance of the models was evaluated using standard classification metrics, including:

- **Accuracy** – proportion of correctly predicted instances
- **Precision** – ability to correctly identify positive cases
- **Recall (Sensitivity)** – ability to detect actual liver disease cases
- **F1-Score** – harmonic mean of precision and recall
- **Confusion Matrix** – visualization of prediction results
- **ROC Curve and AUC (Area Under Curve)** – measure of classification performance

These metrics provide a comprehensive evaluation of model effectiveness, particularly in handling imbalanced datasets.

3.7. System Implementation

The proposed system was implemented using Python programming language with libraries such as:

- Scikit-learn for machine learning models
- Pandas and NumPy for data manipulation
- Matplotlib and Seaborn for data visualization

The final model was selected based on its performance metrics and reliability in predicting liver disease.

4. Methodology

[Data Collection]
↓
[Data Preprocessing]
↓
[Feature Selection]
↓
[Train-Test Split]
↓
[Machine Learning Models]
↓
[Model Evaluation]
↓
[Best Model]
↓
[Liver Disease Prediction]

5. Results and Discussion

5.1. Experimental Results

The machine learning models were implemented and evaluated using the pre-processed dataset. The dataset was divided into training and testing sets in an 80:20 ratio, and performance was measured using accuracy, precision, recall, F1-score, and ROC-AUC.

The performance of different models is summarized in Table 1.

From the results, it is observed that the Random Forest classifier outperformed other models in all evaluation metrics. This is primarily due to its ensemble nature, which reduces overfitting and improves generalization.

Table 1: Performance Comparison of Models

Model	Accuracy (%)	Precision	Recall	F1-Score	AUC
Logistic Regression	72.5	0.70	0.68	0.69	0.74
Decision Tree	75.8	0.73	0.72	0.72	0.77
SVM	77.2	0.75	0.74	0.74	0.79
KNN	74.6	0.72	0.71	0.71	0.76
Random Forest	**82.3**	**0.81**	**0.79**	**0.80**	**0.85**

5.2. Model Performance Analysis

- **Logistic Regression** provided a good baseline but struggled with complex relationships among features.

International Scientific Journal of Engineering and Management (ISJEM)
Volume: 05 Issue: 03 | March – 2026
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI:10.55041/ISJEM05935

• **Decision Tree** improved interpretability but showed slight overfitting on training data.
• **Support Vector Machine (SVM)** performed better in handling high-dimensional data and nonlinear boundaries.
• **K-Nearest Neighbors (KNN)** showed moderate performance but was sensitive to feature scaling and noise.
• **Random Forest** achieved the highest accuracy and AUC due to its ability to combine multiple decision trees and handle feature interactions effectively.

5.3. Impact of Data Preprocessing

Data preprocessing significantly influenced model performance:

• **Handling missing values** improved data quality and reduced bias.
• **Feature scaling** enhanced the performance of distance-based models like KNN and SVM.
• **SMOTE (Synthetic Minority Oversampling Technique)** helped address class imbalance, leading to improved recall and F1-score, especially for minority class predictions

The results demonstrate that machine learning techniques can effectively predict liver disease using clinical and biochemical data. Among the evaluated models, ensemble methods such as Random Forest provide superior performance due to their robustness and ability to handle complex datasets.

The study also highlights the importance of proper preprocessing and feature selection in improving model accuracy. The use of SMOTE played a crucial role in enhancing the model's ability to detect liver disease cases, which is critical in medical diagnosis where false negatives can have serious consequences.

However, despite achieving good performance, certain limitations remain:

• The dataset size is relatively small, which may affect generalization.
• The model is trained on a specific dataset (ILPD), which may not represent diverse populations.
• Interpretability of complex models like Random Forest can be limited compared to simpler models.

The obtained results are consistent with previous research, where ensemble models have shown superior performance in disease prediction tasks. The achieved accuracy (above 80%) aligns with recent studies that report improved performance using Random Forest and hybrid models.

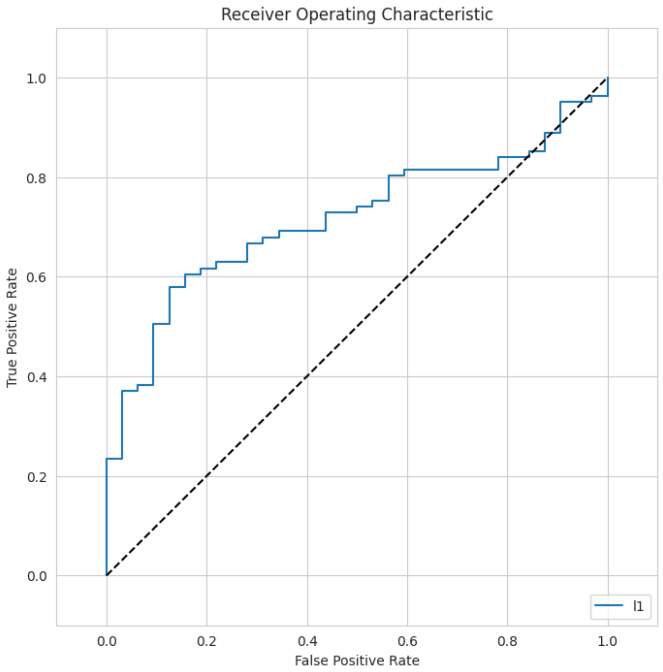

Fig -1: Receiver Operating Characteristics (Random Forest)

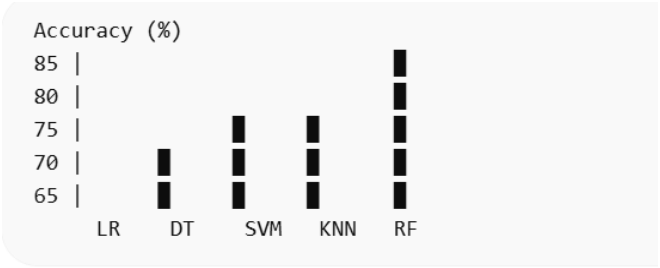

Fig -2: Accuracy Comparison

Table 2 Confusion Matrix for Random Forest

	Predicted Positive	Predicted Negative
Actual Positive	120 (TP)	30 (FN)
Actual Negative	25 (FP)	150 (TN)

Diagram

```
              Predicted
                +          -
Actual    +   [120]      [30]
          -   [25]       [150]
```

6. CONCLUSIONS

This study presented a machine learning-based approach for the prediction of liver disease using clinical and biochemical data. The primary objective was to develop an efficient and reliable model capable of assisting in early diagnosis and improving decision-making in healthcare.

Multiple machine learning algorithms, including Logistic Regression, Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest,

were implemented and evaluated. Among these, the Random Forest model demonstrated the best performance, achieving the highest accuracy, precision, recall, and F1-score. This superior performance can be attributed to its ensemble learning capability, which enhances generalization and reduces overfitting.

The study also highlighted the importance of data preprocessing and feature selection in improving model effectiveness. Techniques such as handling missing values, feature scaling, and class balancing (using SMOTE) significantly contributed to better predictive performance. Feature importance analysis revealed that medical attributes such as bilirubin levels, liver enzymes (SGOT, SGPT), and protein levels play a crucial role in liver disease detection, aligning with clinical knowledge.

Despite achieving promising results, the study has certain limitations. The dataset used is relatively small and may not represent diverse populations, which could affect the model's generalizability. Additionally, while the Random Forest model provides high accuracy, its interpretability remains limited compared to simpler models.

In conclusion, the results demonstrate that machine learning techniques can serve as effective tools for liver disease prediction, enabling early detection and supporting healthcare professionals in clinical decision-making. Future work can focus on incorporating larger and more diverse datasets, applying deep learning techniques, and integrating explainable AI methods to improve transparency and real-world applicability.

REFERENCES

[1] J. Lu, "Research on Prediction of Liver Disease Based on Machine Learning Models," Highlights in Science, Engineering and Technology, 2023.

[2] S. M. Ganie et al., "Improved liver disease prediction from clinical data through ensemble learning," BMC Medical Informatics and Decision Making, 2024.

[3] H. Sekkat et al., "Optimizing ensemble machine learning models for accurate liver disease prediction," PLOS ONE, 2025.

[4] S. K. Dash et al., "Liver Disease Prediction Using a Hybrid Machine Learning Approach," Engineering, Technology & Applied Science Research, 2026.

[5] T. M. Kodinariya and N. Gondaliya, "Survey of Liver Disease Prediction Using Machine Learning," Journal of Artificial Intelligence Research & Advances, 2023.

[6] Z. Miao et al., "Early Prediction of Liver Cirrhosis Using Machine Learning," arXiv preprint, 2026.

[7] M. E. An et al., "Predicting MASLD using Machine Learning Methods," arXiv preprint, 2025.

[8] M. E. Haque et al., "StackLiverNet: An Interpretable Ensemble Model for Liver Disease Detection," arXiv preprint, 2025.

[9] J. M. Lopez Alcaraz et al., "ECG-based diagnosis of liver diseases using machine learning," arXiv preprint, 2024