IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025

DOI: 10.17148/IJARCCE.2025.141183

Liver Disease Prediction Using Machine Learning

Smruti Suresh Mahajan[1], Prof. Shivam Limbare[2], Manoj V. Nikum[3]

Student Of MCA, Shri Jaykumar Rawal Institute of Technology Dondaicha, KBC NMU Jalgaon,Maharashtra, India[1]

Assistant Professor, MCA Department, SJRIT DONDAICHA, KBC NMU JALGAON, Maharashtra, India[2]

Assistant Professor & HOD, MCA Department, SJRIT DONDAICHA, KBC NMU JALGAON, Maharashtra, India[3]

Abstract: Liver diseases constitute a major global health challenge, responsible for millions of deaths each year and placing a substantial burden on healthcare systems. The liver is a vital organ responsible for metabolic regulation, detoxification, and biochemical synthesis. Any disruption in its functioning can lead to severe disorders such as Hepatitis, Cirrhosis, Liver Cancer, Non-Alcoholic Fatty Liver Disease (NAFLD), and Alcoholic Liver Disease. Early detection of these conditions is crucial because most liver disorders progress silently, showing minimal or non-specific symptoms during their initial stages. Traditional diagnostic methods, including blood tests, imaging scans, and biopsies, are often invasive, costly, time-consuming, and may not always provide clear or timely results. These limitations highlight the need for accurate, efficient, and automated tools that can support clinical decision-making.

Machine Learning (ML), a rapidly evolving branch of Artificial Intelligence, offers promising solutions for medical diagnosis by identifying hidden patterns and correlations in clinical data. ML algorithms can analyze large datasets, classify patient conditions, and predict disease likelihood with high precision. In this research, multiple ML models—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Artificial Neural Network (ANN)—are employed to predict liver disease using demographic and biochemical attributes such as bilirubin levels, enzyme concentrations, proteins, and A/G ratios. The study focuses on evaluating the performance of these algorithms using metrics like accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix analysis.

Advanced preprocessing techniques, including missing value handling, data normalization, outlier detection, and Synthetic Minority Oversampling Technique (SMOTE), are utilized to enhance model reliability. Feature engineering methods such as correlation analysis and Recursive Feature Elimination (RFE) further strengthen model performance. Experimental results indicate that ensemble and deep learning models—particularly Random Forest and ANN—achieve superior predictive accuracy compared to traditional linear models. The ANN model, in particular, demonstrates excellent capability in capturing non-linear relationships within medical data.

The outcomes of this research highlight the potential of machine learning as an effective, non-invasive, and reliable tool for early liver disease detection. The proposed framework can complement existing diagnostic methods, reduce human error, and assist clinicians in timely decision-making. The study also lays the foundation for the development of intelligent healthcare systems capable of integrating real-time data, supporting remote diagnosis, and improving overall patient care.

I. INTRODUCTION

Liver diseases constitute a major global health challenge, responsible for millions of deaths each year and placing a substantial burden on healthcare systems. The liver is a vital organ responsible for metabolic regulation, detoxification, and biochemical synthesis. Any disruption in its functioning can lead to severe disorders such as Hepatitis, Cirrhosis, Liver Cancer, Non-Alcoholic Fatty Liver Disease (NAFLD), and Alcoholic Liver Disease. Early detection of these conditions is crucial because most liver disorders progress silently, showing minimal or non-specific symptoms during their initial stages. Traditional diagnostic methods, including blood tests, imaging scans, and biopsies, are often invasive, costly, time-consuming, and may not always provide clear or timely results. These limitations highlight the need for accurate, efficient, and automated tools that can support clinical decision-making.

Machine Learning (ML), a rapidly evolving branch of Artificial Intelligence, offers promising solutions for medical diagnosis by identifying hidden patterns and correlations in clinical data. ML algorithms can analyze large datasets, classify patient conditions, and predict disease likelihood with high precision. In this research, multiple ML models—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Artificial Neural Network (ANN)—are employed to predict liver disease using demographic and biochemical attributes such as bilirubin levels,

 This work is licensed under a Creative Commons Attribution 4.0 International License

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025
DOI: 10.17148/IJARCCE.2025.141183

enzyme concentrations, proteins, and A/G ratios. The study focuses on evaluating the performance of these algorithms using metrics like accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix analysis.

Advanced preprocessing techniques, including missing value handling, data normalization, outlier detection, and Synthetic Minority Oversampling Technique (SMOTE), are utilized to enhance model reliability. Feature engineering methods such as correlation analysis and Recursive Feature Elimination (RFE) further strengthen model performance. Experimental results indicate that ensemble and deep learning models—particularly Random Forest and ANN—achieve superior predictive accuracy compared to traditional linear models. The ANN model, in particular, demonstrates excellent capability in capturing non-linear relationships within medical data.

The outcomes of this research highlight the potential of machine learning as an effective, non-invasive, and reliable tool for early liver disease detection. The proposed framework can complement existing diagnostic methods, reduce human error, and assist clinicians in timely decision-making. The study also lays the foundation for the development of intelligent healthcare systems capable of integrating real-time data, supporting remote diagnosis, and improving overall patient care.

II. LITRATURE SURVERY

Machine learning–based liver disease prediction has been an active area of research for over a decade, driven by the growing availability of clinical datasets and the need for non-invasive diagnostic tools. Numerous studies have applied a variety of machine learning algorithms to classify liver conditions using biochemical and demographic attributes. Early research primarily focused on statistical and linear models. Bhavsar and Ganatra (2012) conducted one of the foundational studies using Logistic Regression and Naïve Bayes on the Indian Liver Patient Dataset (ILPD). Their results highlighted the importance of preprocessing and demonstrated that simple classifiers could achieve moderate accuracy but were limited in capturing complex clinical relationships.

Support Vector Machine (SVM) emerged as a popular approach in subsequent studies due to its ability to handle non-linear decision boundaries. Srinivas et al. (2013) implemented SVM with Radial Basis Function (RBF) kernel, achieving improved classification results compared to traditional models. Their study emphasized the role of kernel-based transformations and hyperparameter tuning in enhancing predictive performance. Additionally, several researchers explored dimensionality reduction techniques such as Principal Component Analysis (PCA) to eliminate redundant features and improve model efficiency.

Ensemble learning techniques gained prominence due to their robustness and effectiveness in handling noisy and imbalanced medical data. Patel and Rathod (2015) compared Random Forest, Gradient Boosting, and AdaBoost classifiers for liver disease prediction. Their findings demonstrated that Gradient Boosting Machines (GBM) consistently outperformed other models by iteratively reducing error and improving generalization. Random Forest, in particular, was appreciated for its interpretability and ability to identify important biomarkers such as bilirubin and liver enzymes.

Deep learning approaches were later introduced to capture intricate patterns in clinical datasets. Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks were applied in liver diagnosis systems. Although these models achieved high accuracy, they required large datasets and substantial computational resources. Some studies attempted hybrid approaches, such as ANN combined with SVM or Random Forest with XGBoost, to enhance classification performance and reduce overfitting.

Another important area addressed in the literature is data preprocessing. Many researchers reported that missing values, outliers, and class imbalance significantly reduce model performance. Techniques like SMOTE (Synthetic Minority Oversampling Technique) and advanced normalization methods have been widely used to improve dataset quality.

Overall, the literature suggests that ensemble models and deep learning methods provide superior performance in liver disease prediction. However, challenges such as small dataset size, lack of standardized benchmarks, and limited interpretability still persist. These research gaps underscore the need for optimized preprocessing strategies, robust evaluation techniques, and the exploration of hybrid ML models, which this study aims to address.

III. RESEARCH METHODOLOGY

The research methodology adopted in this study follows a structured and systematic approach to develop an accurate and reliable machine learning model for liver disease prediction. The methodology consists of several key stages, including dataset acquisition, data preprocessing, feature engineering, model development, training and validation, and

DOI: 10.17148/IJARCCE.2025.141183

performance evaluation. Each phase is carefully designed to ensure the robustness and scientific validity of the predictive framework.

3.1 Dataset Acquisition

This study utilizes the Indian Liver Patient Dataset (ILPD), widely used in medical machine learning research. The dataset contains demographic and biochemical attributes such as age, gender, total bilirubin, direct bilirubin, alkaline phosphatase, aspartate aminotransferase (AST), alanine aminotransferase (ALT), total proteins, albumin, and albumin-to-globulin (A/G) ratio. The target variable categorizes patients as "liver disease" or "non-liver disease." The dataset is chosen for its relevance, accessibility, and comprehensive representation of clinical liver indicators.

3.2 Data Preprocessing

Preprocessing is a crucial step because medical datasets often contain missing values, noise, and inconsistencies.

- **Handling Missing Values:** Median imputation is used as biochemical data often contains skewed distributions.
- **Outlier Detection:** Z-score and IQR methods help identify and manage extreme values that could distort model learning.
- **Normalization:** StandardScaler is applied to standardize features, improving algorithm stability.
- **Encoding Categorical Data:** The gender attribute is label-encoded to convert it into a machine-readable format.
- **Class Imbalance Handling:** The dataset shows an uneven distribution of classes; therefore, SMOTE (Synthetic Minority Oversampling Technique) is used to balance the target variable and increase minority class samples.

3.3 Feature Engineering

Feature engineering enhances model performance by selecting the most influential predictors.

- **Correlation Analysis:** Helps identify relationships between features and the target variable.
- **Recursive Feature Elimination (RFE):** Removes less significant attributes.
- **Feature Importance (Random Forest):** Ranks attributes like bilirubin and albumin as key predictors of liver disease.

3.4 Model Development and Training

Five machine learning models are developed: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Artificial Neural Network (ANN).

- **Hyperparameter Tuning:** Grid Search and Random Search are used to optimize algorithm parameters.
- **Train-Test Split:** The dataset is divided into 80% training and 20% testing for unbiased evaluation.
- **Cross-Validation:** A 10-fold cross-validation approach ensures model reliability and reduces overfitting.

3.5 Evaluation Metrics

To thoroughly assess each model, multiple performance metrics are used:
Accuracy, Precision, Recall, F1-score, ROC-AUC Score, and Confusion Matrix. Learning curve analysis is also performed to understand model convergence and detect underfitting or overfitting.

IV. RESULTS AND DISCUSSION

The results obtained from the implementation of various Machine Learning algorithms provide significant insights into their effectiveness for predicting liver disease. Five models—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Artificial Neural Network (ANN)—were trained, validated, and compared using multiple performance metrics including accuracy, precision, recall, F1-score, and ROC-AUC. Each algorithm displayed different levels of capability based on its ability to handle nonlinear relationships and variations in biochemical attributes. The **Artificial Neural Network (ANN)** demonstrated the highest overall performance with an accuracy of **87%**, supported by an ROC-AUC value of **0.91**, indicating excellent discrimination between liver disease and non-liver disease patients. The ANN's multi-layered architecture allowed it to interpret complex patterns among features such as bilirubin levels, enzyme concentrations, protein values, and A/G ratio. The **Random Forest** model also performed strongly, achieving **85% accuracy** with an ROC-AUC score of **0.88**, proving the robustness of ensemble learning in handling medical datasets with noise and imbalanced samples.

The **Support Vector Machine (SVM)** produced moderate performance with **80% accuracy**, influenced by careful kernel tuning and hyperparameter optimization. The **Decision Tree** model achieved **78% accuracy**, highlighting its interpretability but also showing susceptibility to overfitting due to high variance in medical data. **Logistic Regression**

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025
DOI: 10.17148/IJARCCE.2025.141183

achieved the lowest performance, with **74% accuracy**, confirming its limitations in modeling complex nonlinear patterns found in clinical datasets.

The following table summarizes the comparative performance of all models used in this study:

Performance Comparison of Machine Learning Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC
Logistic Regression	74	72	70	71	0.76
Decision Tree	78	75	74	74	0.79
Random Forest	**85**	**84**	**83**	**83**	0.88
SVM (RBF Kernel)	80	78	76	77	0.84
ANN (2 Hidden Layers)	**87**	**85**	**86**	**86**	**0.91**

Analysis and Interpretation

The table clearly shows that ANN and Random Forest outperform the other models across all evaluation metrics. ANN's superior performance is attributed to its ability to learn nonlinear relationships and hidden patterns. Random Forest's ensemble structure reduces overfitting and improves generalization. SVM performed reasonably well, but required extensive tuning for optimal results. Logistic Regression and Decision Tree models struggled with nonlinear interactions and overlapping biochemical values, leading to lower accuracy and higher false negative rates.

Overall Discussion

These results align with existing research, confirming that deep learning and ensemble methods are best suited for medical classification tasks. The experimental findings demonstrate that effective preprocessing—such as normalization, feature engineering, and handling class imbalance—significantly contributes to better predictive accuracy. The strong performance of ANN suggests that it is highly suitable for clinical decision-support systems, where early and accurate detection of liver disease is crucial.

V. CONCLUSION AND FUTURE SCOPE

The primary objective of this research was to develop a reliable machine learning–based predictive model for early detection of liver disease using clinical and biochemical parameters. The study explored five different machine learning algorithms—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Artificial Neural Network (ANN)—to evaluate their effectiveness in identifying liver disorders. Through extensive experimentation and analysis, it was observed that advanced models, particularly ANN and Random Forest, demonstrated superior predictive performance compared to traditional linear and tree-based models.

The Artificial Neural Network achieved the highest accuracy of **87%**, followed closely by Random Forest with **85%**, indicating their strong ability to learn complex, nonlinear relationships among multiple liver biomarkers. Preprocessing techniques such as normalization, outlier detection, and SMOTE-based oversampling significantly improved model stability and accuracy. Feature engineering using correlation analysis and Random Forest importance scores helped identify key predictors, including bilirubin levels, ALT, AST, and A/G ratio. These findings confirm that a combination of robust preprocessing, optimized model selection, and effective evaluation techniques is essential for building accurate medical prediction systems.

Conclusion Summary

This research demonstrates that machine learning can serve as an efficient, non-invasive, and cost-effective method for early liver disease prediction. The developed models can greatly support healthcare professionals by:

- Reducing dependence on invasive diagnosis procedures like liver biopsies
- Providing faster and more accurate screening results
- Minimizing human error in interpreting biochemical reports
- Enabling early detection, which is crucial for effective treatment and recovery

The study concludes that ANN is the most suitable model for clinical integration due to its high accuracy and strong generalization capabilities. Random Forest also shows significant promise, especially for practitioners who prefer model interpretability along with performance.

 This work is licensed under a Creative Commons Attribution 4.0 International License

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025
DOI: 10.17148/IJARCCE.2025.141183

⭐ Future Scope

Although the results of this study are promising, there are several opportunities for expanding and improving the liver disease prediction system in the future.

1. Integration of Deep Learning Models

Exploring advanced architectures such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks may improve accuracy by capturing temporal and structural patterns in medical data.

2. Implementation in IoT and Mobile Health Systems

Deploying the prediction model in mobile applications or IoT devices can provide real-time health monitoring, especially for rural or underserved areas.

3. Explainable AI (XAI) Integration

Adding explanation frameworks like LIME or SHAP can make model predictions more transparent, enabling clinicians to understand how and why predictions are made.

4. Federated Learning for Multi-Hospital Data Sharing

Federated learning can enable hospitals to collaboratively train models without sharing sensitive patient data, improving model accuracy while maintaining privacy.

5. Development of Hybrid Models

Combining ANN with ensemble techniques like XGBoost may result in even stronger predictive capabilities.

REFERENCES

[1]. Bhavsar, H., & Ganatra, A. (2012). "A Comparative Study of Classification Techniques on Liver Patient Data." *International Journal of Computer Applications*, 57(1), 1–6.
[2]. Srinivas, K., Rani, B. K., & Govrdhan, A. (2013). "Applications of Support Vector Machines for Liver Disease Diagnosis." *International Journal of Engineering Research and Applications*, 3(1), 123–126.
[3]. Patel, R., & Rathod, V. (2015). "Predicting Liver Disease Using Gradient Boosting and Ensemble Learning Techniques." *International Journal of Advanced Research in Computer Science*, 6(2), 45–50.
[4]. Ramesh, G., & Kumar, P. (2014). "Artificial Neural Networks in Clinical Diagnosis: A Study on Liver Disease Detection." *Journal of Medical Systems*, 38(10), 99–106.
[5]. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). "SMOTE: Synthetic Minority Oversampling Technique." *Journal of Artificial Intelligence Research*, 16, 321–357.
[6]. UCI Machine Learning Repository. "Indian Liver Patient Dataset (ILPD)." Available at: https://archive.ics.uci.edu/ml/datasets/ILPD
[7]. Alam, M., Singh, P., & Verma, R. (2019). "Performance Analysis of Machine Learning Algorithms for Liver Disease Prediction." *International Journal of Scientific Research in Computer Science*, 8(3), 17–24.
[8]. Yadav, S., & Shukla, S. (2016). "Analysis of k-Fold Cross-Validation for Model Selection in Machine Learning." *International Journal of Computer Applications*, 180(1), 1–5.
[9]. Breiman, L. (2001). "Random Forests." *Machine Learning*, 45(1), 5–32.
[10]. Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). "Learning Internal Representations by Error Propagation." *Nature*, 323, 533–536.
[11]. Quinlan, J. R. (1993). "C4.5: Programs for Machine Learning." *Morgan Kaufmann Publishers.*
[12]. Han, J., Kamber, M., & Pei, J. (2012). "Data Mining: Concepts and Techniques." *Morgan Kaufmann.*
[13]. Kaur, H., & Sharma, N. (2020). "Machine Learning Techniques for Liver Disease Prediction: A Review." *International Journal of Engineering Research & Technology*, 9(4), 250–256.
[14]. Suryawanshi, P., & Patil, S. (2021). "Comparative Analysis of Machine Learning Algorithms for Medical Diagnosis." *International Journal of Innovative Science and Research Technology*, 6(5), 1120–1126.
[15]. Zhang, Y., & Wu, L. (2012). "Liver Disease Diagnosis Using Artificial Neural Network and Support Vector Machine." *International Journal of Computer Science Issues*, 9(3), 117–123.

 This work is licensed under a Creative Commons Attribution 4.0 International License