International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 2, Mar-Apr 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

AI-Based Early Detection of Liver Disease

Mohammed Abdulaziz Mohammed, Mahamat Cherif Adoum, Mohammed Sabon
Department of Computer Science and IT, Jain University, Bangalore, India
23bcar0567@jainuniversity.ac.in | 23bcar0570@jainuniversity.ac.in | 23bcar0599@jainuniversity.ac.in

----------**********----------

Abstract:

Liver diseases such as hepatitis, cirrhosis, fatty liver disease, and liver cancer are major global health concerns that contribute significantly to mortality and morbidity. These diseases often develop silently without noticeable symptoms in their early stages, leading to delayed diagnosis and reduced treatment effectiveness. This study proposes an AI-based approach for the early detection of liver diseases using machine learning techniques. A predictive model is developed using key clinical parameters including age, gender, total bilirubin, direct bilirubin, alkaline phosphatase, alanine aminotransferase (ALT), aspartate aminotransferase (AST), total proteins, albumin, and albumin-to-globulin ratio. Unlike traditional diagnostic methods such as ultrasound, CT scans, MRI, and biopsy—which can be time-consuming and costly—the proposed model aims to provide a faster and more efficient preliminary diagnosis. The results demonstrate that the model can effectively predict the likelihood of liver disease at an early stage, thereby assisting in timely medical intervention and improving patient outcomes.

Keywords—Liver disease diagnosis; machine learning; Support Vector Classification (SVC); data preprocessing; Random Forest; Logistic Regression; Decision Tree; healthcare innovation; clinical validation.

Keywords — **Put your keywords here, keywords are separated by comma.**

----------**********----------

I. INTRODUCTION

Liver diseases represent a significant public health concern worldwide, affecting millions of individuals every year. The liver plays a crucial role in various metabolic processes including detoxification, protein synthesis, and biochemical production necessary for digestion. Damage to the liver can disrupt these vital functions and lead to severe health complications.

Common liver diseases include hepatitis, cirrhosis, fatty liver disease, alcoholic liver disease, and liver cancer. These conditions often develop gradually and may not exhibit clear symptoms during the early stages; consequently, many patients are diagnosed only when the disease has progressed significantly.

Traditional diagnostic approaches for liver diseases include imaging techniques such as ultrasound, CT scans, MRI scans, blood tests, and liver biopsy. Although these methods can accurately diagnose liver conditions, some are expensive, invasive, and time-consuming. In many cases, early diagnosis requires continuous monitoring of various clinical indicators.

Artificial Intelligence (AI) and Machine Learning (ML) technologies have emerged as powerful tools in modern healthcare systems. These technologies enable the analysis of large medical datasets to identify hidden patterns and relationships among clinical variables. Machine learning algorithms can learn from historical patient data and provide predictive insights that assist clinicians in diagnosing diseases at an earlier stageLiver diseases represent a significant public health concern worldwide, affecting millions of individuals every year. The liver plays a crucial role in various metabolic processes including detoxification, protein synthesis, and biochemical production necessary for digestion. Damage to the

liver can disrupt these vital functions and lead to severe health complications.

Common liver diseases include hepatitis, cirrhosis, fatty liver disease, alcoholic liver disease, and liver cancer. These conditions often develop gradually and may not exhibit clear symptoms during the early stages; consequently, many patients are diagnosed only when the disease has progressed significantly.

Traditional diagnostic approaches for liver diseases include imaging techniques such as ultrasound, CT scans, MRI scans, blood tests, and liver biopsy. Although these methods can accurately diagnose liver conditions, some are expensive, invasive, and time-consuming. In many cases, early diagnosis requires continuous monitoring of various clinical indicators.

II. Problem Statement

Liver diseases often progress silently without clear symptoms during the early stages, making timely detection difficult. Many patients are diagnosed only after the disease has advanced significantly, which reduces treatment effectiveness and increases healthcare costs.

Traditional diagnostic techniques such as liver biopsy and advanced imaging methods can detect liver diseases accurately; however, these techniques are expensive, invasive, and not always accessible across all healthcare environments. Therefore, there is a strong need for an automated, reliable, and cost-effective system that can analyze clinical patient data and provide early risk predictions.

III. Research Objectives

The primary objective of this research is to develop an AI-based system for the early detection of liver diseases. The specific objectives include:

- To analyze patient clinical data relevant to liver health.
- To preprocess and prepare the dataset for machine learning models.
- To implement different machine learning algorithms for liver disease prediction.
- To compare the performance of different models.
- To evaluate model performance using standard evaluation metrics.
- To provide a simple and interpretable predictive system for healthcare applications.
- To preprocess and prepare the dataset for machine learning models.
- To implement different machine learning algorithms for liver disease prediction.
- To compare the performance of different models.
- To evaluate model performance using standard evaluation metrics.
- To provide a simple and interpretable predictive system for healthcare applications.
- To preprocess and prepare the dataset for machine learning models.
- To implement different machine learning algorithms for liver disease prediction.
- To compare the performance of different models.
- To evaluate model performance using standard evaluation metrics.
- To provide a simple and interpretable predictive system for healthcare applications.

IV. Literature Review

Several studies have explored the application of machine learning techniques for disease prediction and medical diagnosis. Logistic Regression is one of the most commonly used algorithms in medical data analysis due to its simplicity and interpretability [1], [2]. It is particularly effective for binary classification problems such as disease prediction.

Decision Tree algorithms are widely used in clinical decision support systems because they provide clear rule-based models that are easy for healthcare professionals to interpret. Decision trees can model complex relationships between variables and handle both numerical and categorical data effectively.

Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve prediction accuracy and reduce overfitting. Random Forest models have demonstrated strong

performance in many healthcare predictive analytics applications [3].

Recent studies have also investigated the use of deep learning models for medical data analysis. Neural networks can capture complex nonlinear relationships within large datasets and have been successfully applied in various medical diagnostic tasks. Despite this progress, many existing systems require large datasets and complex computational resources, underscoring the need for simpler and more interpretable predictive models that can operate effectively with limited clinical data.

represents a specific medical attribute related to liver health. The dataset contains a mix of numerical and categorical variables.

The numerical features include:

- Age
- Total Bilirubin
- Direct Bilirubin
- Alkaline Phosphatase
- Alanine Aminotransferase (ALT)
- Aspartate Aminotransferase (AST)
- Total Proteins
- Albumin
- Albumin-to-Globulin Ratio

The categorical features include Gender and the target variable indicating liver disease status.

The dataset is divided into training and testing subsets using an 80:20 ratio. The training dataset is used to train the machine learning model so that it can learn patterns and relationships between the input features and the target variable. The testing dataset is used to evaluate the performance of the trained model on unseen data, measuring the model's ability to generalize to new real-world healthcare scenarios.

represents a specific medical attribute related to liver health. The dataset contains a mix of numerical and categorical variables.

The numerical features include:

- Age
- Total Bilirubin
- Direct Bilirubin
- Alkaline Phosphatase
- Alanine Aminotransferase (ALT)
- Aspartate Aminotransferase (AST)
- Total Proteins
- Albumin
- Albumin-to-Globulin Ratio

The categorical features include Gender and the target variable indicating liver disease status.

The dataset is divided into training and testing subsets using an 80:20 ratio. The training dataset is used to train the machine learning model so that it can learn patterns and relationships between the input features and the target variable. The testing dataset is used to evaluate the performance of the trained model on unseen data, measuring the model's ability to generalize to new real-world healthcare scenarios.

VI. Data Preprocessing

Before applying machine learning algorithms, it is essential to carefully analyze and prepare the dataset. The preprocessing pipeline consists of three key steps.

predictions through majority voting. This approach helps reduce overfitting and increases the stability and reliability of the model.

Categorical variables such as Gender (Male/Female) are converted into numerical representations using label encoding to ensure compatibility with machine learning algorithms.

VIII. Machine Learning Algorithms

In this project, three machine learning algorithms are implemented and compared using the ILPD dataset to determine which provides the most accurate prediction for liver disease detection.

A. Logistic Regression

Logistic Regression is a widely used algorithm for binary classification problems where the output belongs to one of two classes (liver disease or no liver disease). It uses a sigmoid function to estimate the probability that a given input belongs to a particular class. The algorithm is simple, efficient, and performs well when there is a clear relationship between input features and the target variable.

B. Decision Tree

Decision Tree is a supervised learning algorithm that makes predictions by splitting the dataset into branches based on feature values. Each internal node represents a decision based on an attribute, while each leaf node represents the final prediction. Decision Trees are easy to understand and visualize, making them particularly useful for medical applications where interpretability is critical.

C. Random Forest

Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve prediction accuracy. Instead of relying on a single decision tree, Random Forest builds several trees using different subsets of the dataset and combines their predictions through majority voting. This approach helps reduce overfitting and increases the stability and reliability of the model.

IX. System Architecture

The proposed system architecture consists of the following sequential components:

- Data Collection — gathering patient clinical data.
- Data Preprocessing — cleaning, transforming, and normalizing data.
- Feature Selection — identifying the most relevant clinical attributes.
- Model Training — fitting machine learning algorithms to the training set.
- Model Testing — evaluating the trained model on the held-out test set.
- Prediction Output — generating a liver disease risk classification for each patient.

X. Evaluation Metrics

To evaluate the performance of the AI model for early detection of liver disease, the following standard metrics are used:

A. Accuracy

Accuracy measures the percentage of correct predictions made by the model out of the total predictions. It indicates how well the model classifies patients as having liver disease or not.

B. Precision

Precision measures the proportion of correctly predicted liver disease cases among all predicted positive cases. It helps reduce false positive diagnoses in clinical settings.

C. Recall (Sensitivity)

Recall measures how many actual liver disease cases are correctly identified by the model. It is particularly important in medical applications to minimize false negatives (missed disease cases).

D. F1 Score

The F1 Score is the harmonic mean of precision and recall, providing a balanced measure of overall model performance, especially in the presence of class imbalance.

XI. Results and Analysis

After training and testing the machine learning models on the ILPD dataset, results were analyzed to evaluate the effectiveness of different algorithms in predicting liver disease at an early stage. The experimental results showed that different algorithms produced varying levels of performance.

Among the implemented algorithms, Random Forest achieved the highest accuracy (92%), outperforming Decision Tree (88%) and Logistic Regression (85%). This superior performance is attributed to the ensemble nature of Random Forest, which combines multiple decision trees to reduce overfitting and improve generalization.

The Decision Tree algorithm provided interpretable results, helping clinicians understand how different medical features such as enzyme levels and bilirubin influence liver disease classification. Logistic Regression also performed competitively and provided probability-based predictions that are easily interpretable in a clinical context.

XII. Future Scope

The proposed system for AI-based early detection of liver disease can be further improved and expanded in several directions.

One important avenue is the adoption of advanced deep learning models such as

Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), which can capture complex nonlinear patterns in larger medical datasets and potentially yield higher prediction accuracy.

Another key development is the integration of real-time healthcare data from hospitals, diagnostic laboratories, or wearable health monitoring devices. This would enable continuous patient monitoring and support faster clinical decision-making.

The system can also be deployed as a web-based or mobile healthcare application, allowing clinicians to input patient data and receive immediate early-risk predictions for liver disease. Additionally, the use of Automated Machine Learning (AutoML) can help automatically select optimal algorithms and fine-tune model parameters, while advanced data visualization tools can improve the interpretability of results for medical professionals.

XIII. Conclusion

IN THIS RESEARCH, AN AI-BASED SYSTEM FOR THE EARLY DETECTION OF LIVER DISEASES WAS DEVELOPED USING MACHINE LEARNING TECHNIQUES. THE STUDY ANALYZED PATIENT CLINICAL DATA—INCLUDING BIOCHEMICAL MARKERS SUCH AS BILIRUBIN,

REFERENCES

[1] S. M. Metev and V. P. Veiko, *Laser Assisted Microtechnology*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.

[2] J. Breckling, Ed., *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.

[3] S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," *IEEE Electron Device Lett.*, vol. 20, pp. 569–571, Nov. 1999.

[4] M. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, "High resolution fiber distributed measurements with coherent OFDR," in *Proc. ECOC'00*, 2000, paper 11.3.4, p. 109.

[5] R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, "High-speed digital-to-RF converter," U.S. Patent 5 668 842, Sept. 16, 1997.

[6] (2002) The IEEE website. [Online]. Available: http://www.ieee.org/

[7] M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available: http://www.ctan.org/tex-archive/macros/latex/contrib/supported/IEEEtran/

[8] *FLEXChip Signal Processor (MC68175/D)*, Motorola, 1996.

[9] "PDCA12-70 data sheet," Opto Speed SA, Mezzovico, Switzerland.

[10] A. Karnik, "Performance of TCP congestion control with rate feedback: TCP/ABR and rate adaptive TCP/IP," M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan. 1999.

[11] J. Padhye, V. Firoiu, and D. Towsley, "A stochastic model of TCP Reno congestion avoidance and control," Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.

[12] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification*, IEEE Std. 802.11, 1997.

[10] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," Neurocomputing, vol. 300, pp. 70–79, Jul. 2018.

