LIVER DISEASE PREDICTION

US.Tharun[1], M. Usha Devi[2]
Department of Computer Science, Rathinam College of Arts and Science (Autonomous), Coimbatore, Tamil Nadu, India
Tharunus2005@gmail.com , Usha.devi@gmail.com

Abstract—*This electronic document is a "live" template and already defines the components of your paper [title, text, heads, etc.] in its style sheet. *CRITICAL: Do Not Use Symbols, Special Characters, or Math in Paper Title or Abstract. (Abstract)*

Keywords—component; formatting; style; styling; insert (key words)

I. INTRODUCTION

Liver disease is a serious medical condition that affects the normal functioning of the liver. The liver plays an important role in digestion, metabolism, and detoxification of harmful substances in the body. Diseases such as hepatitis, cirrhosis, and fatty liver are becoming increasingly common due to unhealthy lifestyle habits, alcohol consumption, and poor diet. Traditional methods of diagnosing liver disease involve laboratory tests and expert medical evaluation. These methods are often time-consuming and may delay early detection. In many cases, patients are diagnosed only at advanced stages, which makes treatment more difficult.

With the advancement of technology, machine learning has become a powerful tool in the healthcare sector. Machine learning algorithms can analyze large amounts of medical data, identify patterns, and make accurate predictions. By using patient data such as age, bilirubin levels, enzyme levels, and protein values, these models can help in early detection of liver disease. This study proposes a machine learning-based system that predicts liver disease using classification algorithms such as Logistic Regression and Random Forest, providing a fast and reliable solution.

II. LITERATURE REVIEW

Recent studies show that machine learning techniques are widely used in healthcare for disease prediction and diagnosis. Various models have been developed to analyze medical datasets and identify diseases at an early stage. Researchers have applied algorithms such as Decision Tree, Logistic Regression, Support Vector Machine, and Random Forest for liver disease prediction. These models analyze clinical attributes and provide predictions based on patterns in the data.

However, existing systems have limitations such as limited datasets, lack of proper feature selection, and reduced interpretability. Some models also fail to generalize well to real-world data. To overcome these issues, ensemble methods like Random Forest are used, which improve prediction accuracy and reduce overfitting. The proposed system enhances performance by applying proper preprocessing techniques and comparing multiple algorithms.

III. METHODOLOGY

A. System Architecture

The proposed system follows a structured machine learning pipeline that begins with data collection and continues through preprocessing, model training, prediction, and evaluation. Each stage is interconnected to ensure accurate and efficient processing. This architecture is designed to handle real-world medical data and provide reliable classification results while maintaining scalability for future improvements.

B. Data Collection

Data collection is an important step in building an effective machine learning model. In this project, the dataset is obtained from reliable sources such as Kaggle and contains patient medical records. Each row in the dataset represents an individual patient, while each column represents a specific medical attribute. The dataset includes features such as age, gender, total bilirubin, direct bilirubin, alkaline phosphotase, SGPT, SGOT, total proteins, albumin, and albumin-globulin ratio. The target variable indicates whether the patient has liver disease, where a value of 1 represents the presence of disease and 0 indicates absence.

C. Pre-processing

Data preprocessing is a critical step in the machine learning process as it ensures that the dataset is clean, consistent, and suitable for model training. Initially, missing values are handled by either removing or replacing them with appropriate statistical measures. Duplicate records are eliminated to maintain data integrity. Categorical data such as gender is converted into numerical form using encoding techniques. Feature selection is performed to retain only relevant attributes, and normalization techniques are applied to scale the data. Finally, the dataset is divided into training and testing sets to evaluate model performance effectively.

IV. MODEL TRAINING

Model training involves applying machine learning algorithms to the processed dataset to learn patterns and relationships between input features and the target variable. In this project, Logistic Regression and Random Forest algorithms are used. Logistic Regression serves as a baseline

model due to its simplicity and efficiency in binary classification tasks. Random Forest, on the other hand, is an ensemble learning method that combines multiple decision trees to improve accuracy and reduce overfitting. The models are trained using training data, and their performance is optimized through parameter tuning.

V. Prediction and Evaluation

After training, the models are tested using unseen data to evaluate their performance. The prediction is carried out using the .predict() function, where the output indicates whether a patient has liver disease or not. A value of 1 represents the presence of liver disease, while 0 indicates absence. The performance of the models is evaluated using metrics such as accuracy, precision, recall, and F1-score. Logistic Regression provides moderate accuracy, while Random Forest achieves higher accuracy and better reliability, making it more suitable for this application.

VI. System Design

The system design explains how input data is processed and how predictions are generated. The system is designed to be simple, efficient, and user-friendly. It accepts patient medical data as input, processes it using machine learning algorithms, and produces prediction results indicating the presence of liver disease.

A. Data Flow Design

The data flow begins with user input, where patient medical details are provided to the system. The system processes this data through preprocessing and applies trained machine learning models to generate predictions. Finally, the output is displayed in the form of disease prediction along with performance metrics. The system consists of modules such as input processing, prediction generation, and result visualization, each contributing to the overall functionality.

VII. Result and Discussion

The results of the machine learning models show that Logistic Regression achieves an accuracy of approximately 75%, while Random Forest achieves around 82%. This indicates that Random Forest performs better due to its ability to handle complex data and reduce overfitting. Graphical representations such as bar graphs, heatmaps, and distribution charts help in understanding model performance and dataset characteristics. The results demonstrate that the proposed system is effective in predicting liver disease.
Compared to traditional methods, it is faster, more accurate, and reduces human error. Random Forest proves to be the most reliable model for this application, making it suitable for real-world healthcare systems.

VIII. Conclusion

The liver disease prediction system uses machine learning techniques to analyze patient medical data and predict disease presence. Different models were implemented and evaluated, and the results show that Random Forest provides the highest accuracy and reliability. The system is efficient, easy to use, and helps in early detection of liver disease. It can assist healthcare professionals in making better decisions and improving patient outcomes. Overall, this project highlights the importance of machine learning in modern healthcare and its potential to improve diagnostic processes.

IX. References

1. Singla, B., Taneja, S., Garg, R., & Nagrath, P. (2022). *Liver disease prediction using machine learning and deep learning: A comparative study*. Intelligent Decision Technologies.
2. Rakshith, D. B., Srivastava, M., Kumar, A., & Gururaj, S. P. (2021). *Liver Disease Prediction System using Machine Learning Techniques*. International Journal of Engineering Research & Technology (IJERT).
3. Mostafa, F., Hasan, E., Williamson, M., & Khan, H. (2021). *Statistical Machine Learning Approaches to Liver Disease Prediction*. MDPI Livers Journal.
4. Lu, J. (2023). *Research on Prediction of Liver Disease Based on Machine Learning Models*. Highlights in Science, Engineering and Technology.
5. El Atifi, W., El Rhazouani, O., Khan, F. M., & Sekkat, H. (2025). *Optimizing ensemble machine learning models for accurate liver disease prediction in healthcare*. PLOS ONE.
6. Alenizi, A. S. F., & Al-Karawi, K. A. (2025). *Machine Learning Approach for Liver Disease Prediction*. Springer (ICT Conference Proceedings).
7. Chavda, T., & Swarndeep, S. (2022). *A Comparative Study for Liver Disease Prediction Using Machine Learning*. International Journal of Novel Research and Development.
8. An, M. E., Griffin, P., Stine, J. G., Balakrishnan, R., Sriram, R., & Kumara, S. (2025). *Predicting Metabolic Dysfunction-Associated Steatotic Liver Disease using Machine Learning Methods*. arXiv.
9. Miao, Z., Ravi, S., & Ahmed, A. (2026). *Early Prediction of Liver Cirrhosis Using Machine Learning Models*. arXiv.
10. Deivendran, P., Selvakanmani, S., Jegadeesan, S., & Kumar, V. (2023). *Liver Infection Prediction Analysis using Machine Learning*. arXiv.