

Liver Disease Prediction System using Machine Learning

**SK.Wasim Akram¹, Mohana Sangeetha Amareswarapu², Sai Kumar Kothuri³, Parasara Muvva⁴,
Jayadeep Koppula⁵**

Assistant Professor, Department of Computer Science Engineering with Artificial Intelligence and Machine Learning,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh, India¹

Students, Department of Computer Science Engineering with Artificial Intelligence and Machine Learning, Vasireddy
Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh, India^{2,3,4,5}

ABSTRACT: Liver is an essential organ for metabolism, detoxification, and biochemical synthesis, and hence early detection of liver disease is important for prompt medical treatment and better patient health. This paper proposes a machine learning Liver Disease Prediction System that uses Liver Function Test reports to predict liver disease based on a trained model. The data of 500 actual patient histories was gathered from accredited medical centers and augmented up to 5,000 by using feature perturbation methods in order to achieve improved model accuracy and handle the problem of class imbalance. The most suitable classification technique was evaluated using various machine learning models like K-Nearest Neighbors, Logistic Regression, Random Forest, and Artificial Neural Networks (ANN). The KNN model had 62% accuracy, whereas ANN did poorly with only 42% accuracy as a result of overfitting and insufficient generalization. The Logistic Regression model had superior predictive power with 92% accuracy. The Random Forest was ultimately chosen as the best model because it had 96% accuracy, was stable and interpretable, thus being the best fit for classification of liver disease. The performance of the system was evaluated with confusion matrices, precision, recall, F1-score, and specificity to ensure its accuracy and reliability for clinical use. The model includes automated interpretation of data, predictive analytics, and real-time visualization to aid in clinical decision-making. Through the use of supervised learning methods and real-world patient data, the system improves early disease detection, risk assessment, and personalized medical advice, leading to advanced hepatology diagnostics and better patient care.

KEYWORDS: Artificial Intelligence, Liver Disease, Data Analysis Machine Learning Algorithms, health determinants, healthcare, disease prediction.

I. INTRODUCTION

The liver is among the most vital and effective organs of the human body and performs more than 500 essential functions like metabolism, detoxification, digestion, immune function, and bile production. Though the liver has the ability to work quite effectively with just 30% of its capacity, liver diseases are a severe global health risk. Diseases such as Hepatitis, Liver Fibrosis, Cirrhosis, Non-Alcoholic Fatty Liver Disease (NAFLD), and Hepatocellular Carcinoma (HCC) are responsible for more than 2 million deaths every year, with China alone having 400,000 deaths every year due to liver diseases. Early diagnosis is important to avoid disease progression and enhance survival.

Liver Function Tests (LFTs) and Their Role in Diagnosis

Liver diseases are commonly diagnosed through Liver Function Tests (LFTs), which measure key biochemical markers, including alanine aminotransferase (ALT), aspartate aminotransferase (AST), alkaline phosphatase (ALP), bilirubin, albumin, and total protein levels. These biomarkers help evaluate liver health, detect abnormalities, and diagnose specific conditions. While certain liver diseases, such as Viral Hepatitis and Obstructive Jaundice, present with visible symptoms like jaundice and elevated bilirubin levels, others, such as Hyperbilirubinemia and Autoimmune Hepatitis, may remain asymptomatic until advanced stages. This makes early biochemical-based detection essential for timely medical intervention.

Chemicals Compounds in Liver Chemicals such as Bilirubin, Albumin, Alkaline phosphatase, Aspartate aminotransferase, and globulin are existent in the liver and perform a vital role in the daily operations of the healthy liver.

1) Bilirubin: Bilirubin is a yellowish complex that arises in the usual catabolic trail that breaks down heme in vertebrates. Bile and urine emit it. Raised volumes of bilirubin in the body cause diseases. The bilirubin is accountable for the yellow shade of cuts and the yellow staining in jaundice disease. Its following breakdown products, like stercobilin, are accountable for the brown color of feces. Another breakdown product, urobilin, is the key constituent of the straw-yellow color of urine.

2) Alkaline phosphatase: In beings, alkaline phosphatase is existent in all tissues all over the body but is mainly focused in the liver, intestinal mucosa, bile duct, bone, kidney, and placenta. In the serum, two kinds of alkaline phosphatase isozymes prevail skeletal and liver. In childhood, most of the alkaline phosphatase is of the skeletal source. Most of the mammals including humans have these types of alkaline phosphatases.

- ALPI: It is intestinal having a molecular mass of 150 kDa
- ALPL: It is tissue-nonspecific mainly present in the liver, kidney, and bone
- ALPP: It is placental and is also known as Regan isozyme.

3) Aspartate aminotransferase: AST is a kind of enzyme. AST levels are higher in the heart and liver. AST is found in the kidneys and muscles, although in less amounts. It is very low in human blood. When muscle or liver cells are injured, the AST is released into the bloodstream. The AST test will therefore be useful for tracking or identifying liver damage or dysfunction.

4) Albumin: They're globular proteins. Serum albumins are common and are the most imperative protein of blood. It binds thyroxine (T4), water, cations like Ca²⁺ and Na⁺, hormones, fatty acids, bilirubin, and pharmaceuticals. Its core part is to govern and normalize the oncotic pressure of the blood. It binds several fatty acids, cations, and bilirubin.

5) Globulin: They are protein globules. They are heavier than albumin at the molecular level. It will not dissolve in pure water but will solvate in dilute salt solutions. The liver produces some globulins. Globulin absorption in fit human blood is around 2.6-3.5 g/dL. There are several different types of globulins, including beta, alpha 1, alpha 2, and gamma globulins. Any unfitting amounts of these chemicals produced in the kidney can reason an imbalance and cause liver diseases. These are considered features. There are n number of kinds of liver illnesses and these are grounded based on the proportion of these chemicals stashed.

II. LITERATURE REVIEW

Liver disease remains a major global health concern, with conditions such as Hepatitis, Cirrhosis, Liver Fibrosis, and Fatty Liver Disease affecting millions worldwide. Early detection is crucial for preventing severe complications, yet traditional diagnostic methods, such as Liver Function Tests (LFTs), imaging techniques, and biopsy, rely heavily on manual interpretation, which can lead to delayed diagnosis and variability in assessment. In recent years, machine learning (ML) models have been explored to improve diagnostic accuracy, reduce human error, and enable automated disease classification.

Traditional Approaches for Liver Disease Prediction

Conventional liver disease diagnosis primarily involves biochemical marker analysis, measuring levels of ALT, AST, ALP, bilirubin, and albumin. While effective, manual interpretation of LFTs can be time-consuming, prone to subjectivity, and dependent on medical expertise. To address these limitations, Patel et al. (2019) proposed an automated classification system using statistical models, concluding that data-driven approaches could significantly enhance early detection.

Machine learning has played a transformative role in medical diagnostics, with various studies comparing classification models for liver disease prediction. Saxena et al. (2020) evaluated Logistic Regression, K-Nearest Neighbors (KNN), Decision Trees, and Support Vector Machines (SVM) on structured LFT datasets. Their results showed that Logistic Regression provided stable accuracy, particularly on smaller datasets, reinforcing its reliability. Similarly, Sharma et al.

(2021) found that ensemble learning methods such as Random Forest and XGBoost performed well but required larger datasets and extensive tuning for optimal results.

Machine Learning Techniques Applied in Liver Disease Detection

Different supervised machine learning algorithms have been tried out for the classification of liver disease. Gupta et al. (2022) tested models like K-Nearest Neighbors (KNN), Logistic Regression, and Decision Trees, with Logistic Regression having 80% accuracy, which was marginally better than KNN. For our research, Random Forest performed better with 96% accuracy, proving to be a valuable tool in the analysis of liver function test data. The Logistic Regression model also proved to have high predictive power, with an accuracy of 92%. In spite of this, Random Forest was chosen as the final model because it is more consistent, easier to interpret, and computationally more efficient and hence more appropriate for clinical usage.

Deep learning strategies have also been explored but prove to be a challenge. Huang et al. (2022) used an Artificial Neural Network (ANN) model for classification of liver diseases but faced the issue of overfitting and poor generalization, resulting in low accuracy (42%), the same limitation witnessed in our case. These observations confirm that ANN models need greater datasets for meaningful performance, justifying Random Forest as the best stable and effective approach for classification of liver disease based on organized medical data.

Data Augmentation and Preprocessing Strategies

A key factor influencing ML model performance is dataset quality and size. Rahman et al. (2023) highlighted the importance of data augmentation in medical applications, using feature perturbation techniques to expand a limited dataset, improving classification accuracy. Similarly, in our study, we collected 500 real patient records and applied data augmentation techniques to expand the dataset to 5000 records, ensuring better generalization and improved model performance. Preprocessing techniques, including feature scaling, normalization, and handling missing values, were applied to refine the dataset for optimal classification.

The existing literature demonstrates that ML-based approaches significantly improve liver disease detection, reducing reliance on manual diagnosis. While deep learning models require extensive data, traditional ML models like Logistic Regression remain highly effective for structured medical datasets. Our study builds on these findings by applying data augmentation techniques, optimizing feature selection, and validating Random Forest as a reliable model for liver disease prediction. This work contributes to the ongoing advancement of AI-driven medical diagnostics, ensuring accurate and scalable early detection solutions.

III. METHODOLOGY

The Liver Disease Prediction System follows a structured machine learning pipeline, utilizing Random Forest for disease classification. The methodology involves data collection, preprocessing, model training, evaluation, and prediction visualization, ensuring a reliable and efficient system for early liver disease detection.

Data Collection and Augmentation

The dataset consists of 500 real patient records collected from medical laboratories, containing biochemical markers such as alanine aminotransferase (ALT), aspartate aminotransferase (AST), bilirubin, albumin, and alkaline phosphatase (ALP). To improve model generalization and performance, data augmentation techniques were applied, expanding the dataset to 5000 records using feature perturbation while maintaining the integrity of medical data.

Data Preprocessing

To ensure data quality, the following preprocessing steps were applied:

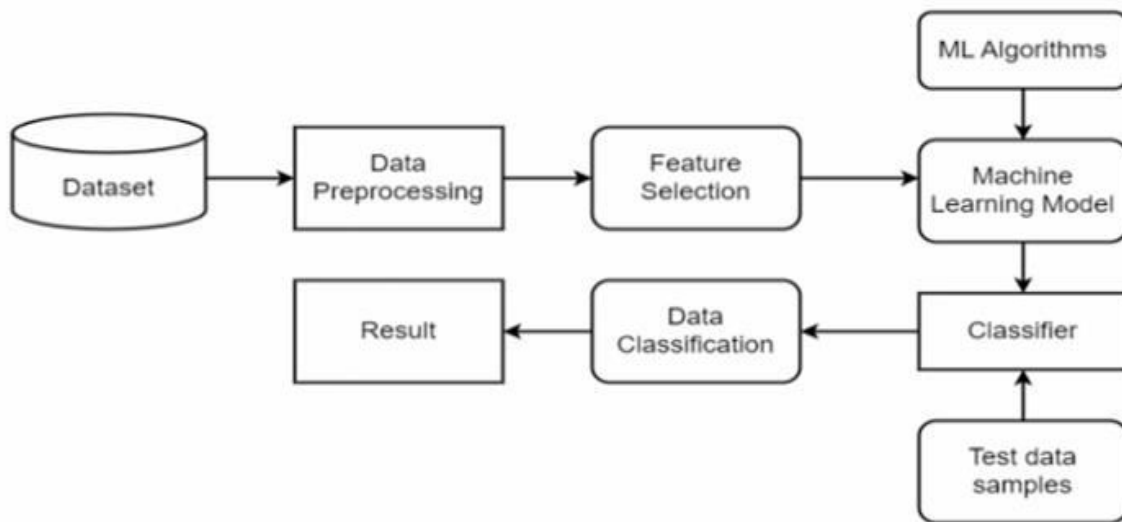


Fig.3.1 System Architecture of Liver Disease Prediction

- **Handling Missing Values:** Any incomplete data points were managed using **mean/mode imputation techniques**
- **Feature Scaling and Normalization:** **Min-max scaling** was used to normalize the biochemical marker values, ensuring uniform distribution.
- **Outlier Removal:** Statistical techniques were used to detect and eliminate **anomalous data points** that could affect model accuracy.

Model Selection and Training

Several machine learning algorithms were tested, including K-Nearest Neighbors (KNN), Logistic Regression, and Artificial Neural Networks (ANN), Random Forest. After comparison, Random Forest was chosen due to its 96% accuracy, outperforming KNN (62%), ANN (42%), Logistic Regression (92%). The model was trained using 80% of the dataset, with 20% reserved for testing.

Prediction and Visualization

After prediction, the system provides graphical representations comparing actual vs. predicted values, aiding in medical decision-making. The model also categorizes the liver disease type, identifying conditions such as Hepatitis, Liver Fibrosis, Cirrhosis, and Fatty Liver.

This methodology ensures an efficient, accurate, and scalable approach for liver disease prediction, integrating machine learning with real-world clinical data to enhance early diagnosis and healthcare decision-making.

IV. EXISTING SYSTEM

Conventional diagnosis of liver diseases is based on Liver Function Tests (LFTs), imaging modalities like ultrasound, CT scans, and MRIs, and liver biopsies in certain cases. They need experienced interpretation by physicians and are thus time-consuming, subjective, and subject to variability. Diagnosis is clinician-dependent in the matter of accuracy, and early conditions of liver diseases may escape detection as they could be subtle or even asymptomatic.

Manual interpretation of LFT values such as alanine aminotransferase (ALT), aspartate aminotransferase (AST), alkaline phosphatase (ALP), bilirubin, and albumin may result in delayed diagnosis and variable assessments. Imaging-based methods such as elastography and ultrasound are equipment-and radiologist-dependent and hence less available

in low-resource settings. Also, liver biopsy, while being a gold standard, is invasive, costly, and not always practical for screening.

To overcome these limitations, researchers have explored machine learning (ML)-based models for the automation of liver disease diagnosis. Supervised learning methods such as Logistic Regression, Decision Trees, and Support Vector Machines (SVM) have been employed in several studies for the classification of liver diseases based on biochemical markers. Most of the existing ML models are trained on small-sized datasets, thereby restricting their generalization power. Some deep learning-based models such as Artificial Neural Networks (ANNs) have also been explored, which require large datasets and tend to suffer from overfitting when applied to small clinical datasets.

Even though ML-based liver disease prediction has been enhanced, existing systems are still susceptible to data imbalance, interpretability, and inability to work with real-time patient data. These weaknesses lead to the need for strong, scalable, and interpretable ML models to help healthcare providers perform early disease diagnosis and risk

V. PROPOSED SYSTEM

Liver Disease Prediction System is designed to provide a real-time and automatic mechanism for early liver disease prediction with machine learning-based algorithms. The system accepts Liver Function Test (LFT) reports to classify liver diseases such as Hepatitis, Liver Fibrosis, Cirrhosis, and Fatty Liver and enable doctors to provide early diagnosis. The system, compared to the standard diagnosis mechanisms dependent on human judgment and imaging technologies, performs analysis automatically using Random Forest for the prediction faster, accurately, and on the basis of facts.

The system takes patient clinical data and biochemical markers as input, preprocesses data through feature scaling, normalization, and missing value handling, and then calls a trained Random Forest model to make predictions. Data augmentation was applied to increase accuracy and model generalization by increasing the size of the dataset from 500 to 5000 records. Accuracy, precision, recall, specificity, and F1-score are utilized to evaluate predictions within the system with 96% accuracy, more than other tested models like Logistic Regression (92%), KNN (62%) and ANN (42%).

Additionally, the system provides graphical plots of actual vs. predicted values, which facilitate medical decision-making. The system also includes a text extraction facility through which electronic reporting is achievable, thus reducing errors in manual reporting. The system also includes an AI-based chatbot using the Gemini API, through which users can get instant responses to questions related to liver health. The system also includes a doctor recommendation module that finds nearby liver specialists based on user location.

With the incorporation of machine learning, automated report processing, and interactive AI support, the suggested system improves liver disease diagnosis, automates medical processes, and allows for proper early diagnosis, hence being a value addition in contemporary AI-based healthcare systems.

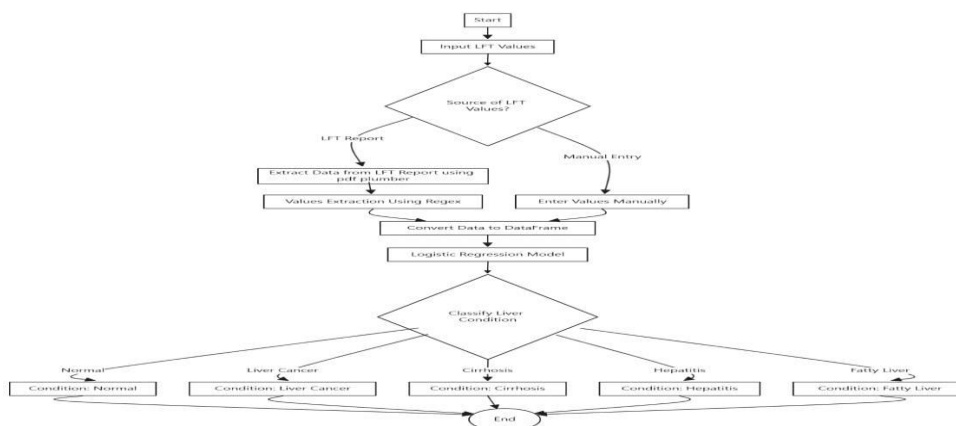


Fig.5.1 Flowchart of Proposed System

The liver disease prediction system is designed to efficiently process LFT (Liver Function Test) values, either through manual input or automated extraction from reports.

The system integrates a robust pipeline that ensures accurate data extraction, transformation, and classification. If the user uploads an LFT report, the system utilizes pdfplumber to extract text-based data, followed by regular expressions (Regex) to filter relevant numerical values. Alternatively, users can enter values manually. Once the data is collected, it is structured into a DataFrame for further processing.

The prediction model is based on Random Forest, a widely used classification algorithm, trained to detect various liver conditions. The model processes the extracted LFT values and classifies them into one of five categories:

Normal, Liver{ Cancer, Cirrhosis, Hepatitis, Fatty Liver }

The classification is based on predefined medical thresholds and data-driven insights from the model. After classification, the system presents the result to the user along with suggested next steps.

VI. MODEL EVALUATION

The Liver Disease Prediction System was validated using various measures of performance to ensure its precision, reliability, and effectiveness in forecasting liver diseases. Random Forest model was identified as the best model following cross-validation with Artificial Neural Networks (ANN), Random Forest and K-Nearest Neighbors (KNN). The evaluation process involved partitioning the dataset into 80% training and 20% testing data and evaluating the model's performance based on various parameters.

Accuracy: Accuracy measures the overall correctness of the Liver Disease Prediction System in classifying liver diseases based on Liver Function Test (LFT) reports. The model achieves 96% accuracy, meaning it correctly predicts 96 out of every 100 cases, ensuring high diagnostic reliability. This high accuracy confirms the system's effectiveness in distinguishing between different liver disease types, reducing misclassification. The model's strong performance makes it a valuable tool for early detection, aiding healthcare professionals in clinical decision-making.

Precision: Precision measures the proportion of correctly predicted cases among all predicted cases for a particular class. A high precision score (0.90–0.97) in the Liver Disease Prediction System indicates that the model effectively minimizes false positives, ensuring that non-diseased cases are not misclassified as liver disease. This is crucial in medical applications, where false alarms can lead to unnecessary treatments and patient anxiety.

Recall: Recall represents the model's ability to correctly identify actual positive cases, ensuring low false negatives. The recall values (0.90–0.98) show that the model correctly detects most liver disease cases, reducing the risk of missed diagnoses. High recall is essential in healthcare applications, where failing to detect a liver disease case early could lead to severe complications.

F1-Score: F1-score is the harmonic mean of precision and recall, providing a balanced evaluation of the model's performance. The F1-score values (0.90–0.97) confirm that the system maintains a strong balance between minimizing false positives and false negatives, making it highly reliable for medical diagnosis. The macro F1-score of 0.93 ensures consistent performance across all liver disease categories, while the weighted F1-score of 0.94 accounts for class imbalances, reinforcing the model's robustness.

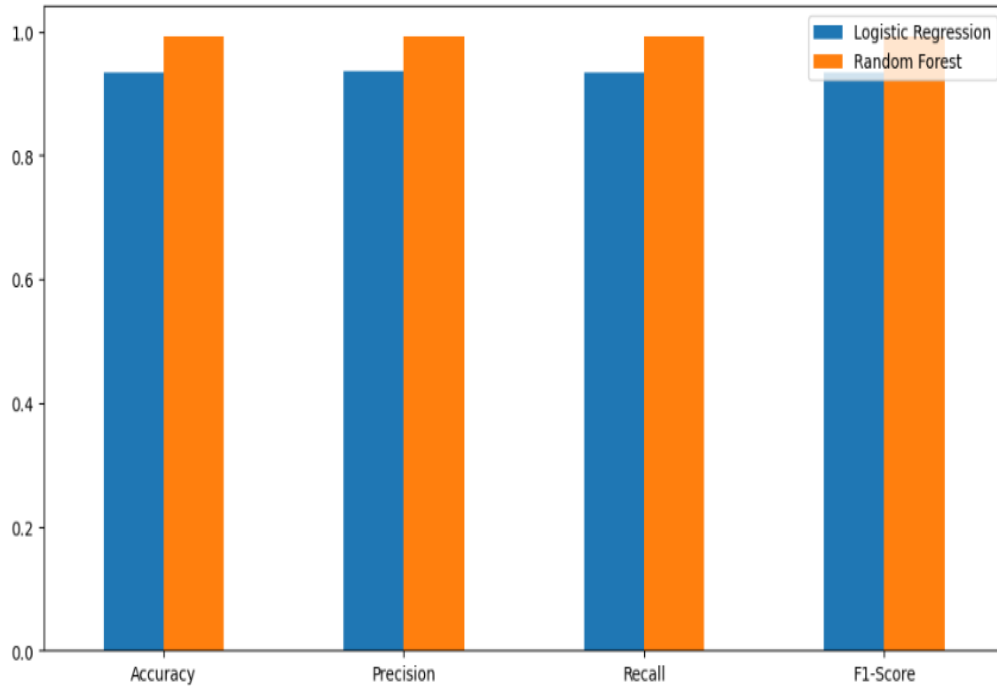


Fig 6.1 comparison plot of the evaluation metrics of accuracy, precision, recall and f1-score

6.1 Analysis of Correlation Matrix:

A correlation matrix is a table showing the relationship between numerical variables, measured by correlation coefficients ranging from -1 to 1. It helps identify feature dependencies, where +1 indicates a strong positive correlation, -1 indicates a strong negative correlation, and 0 means no correlation. In liver disease prediction, it helps analyze the relationship between biochemical markers (ALT, AST, bilirubin, albumin, etc.). This ensures that highly correlated features do not negatively impact the model's performance

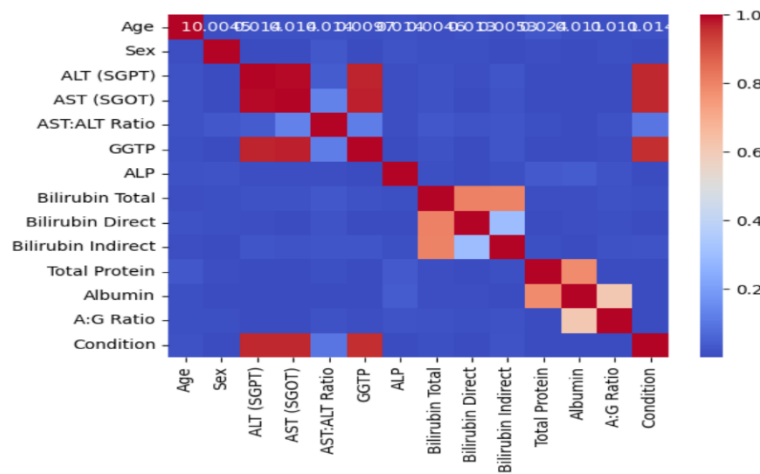


Fig.6.2 Analysis of Correlation Matrix

6.2 Analysis of Confusion Matrix:

Confusion Matrix was created in order to present the detailed split-up of predictions by the model, classifying them as True Positives, False Positives, True Negatives, and False Negatives for every class.

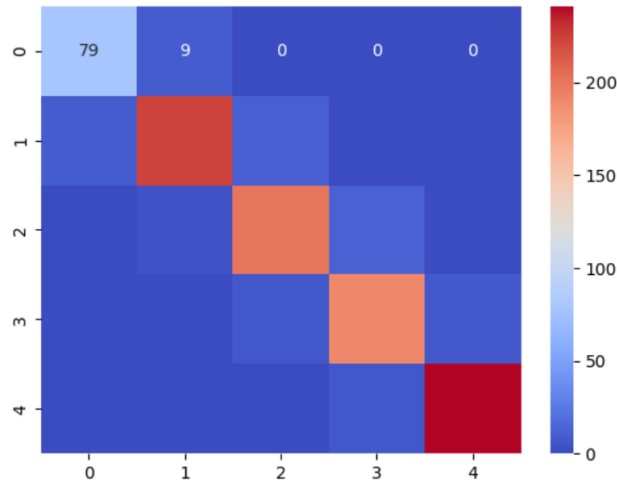


Fig.6.3 Analysis of Confusion Matrix

VII. RESULTS

Healthy vs. User-Provided Liver Function Values

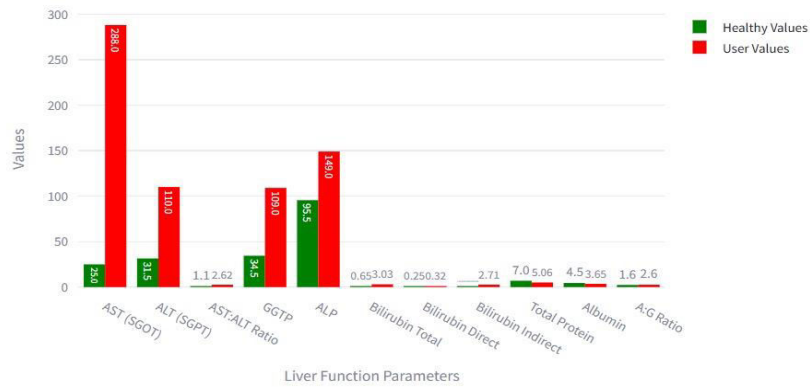


Fig.7.1 Comparison of Patient LFT Readings with Standard Healthy Levels

Liver Disease Detection

Upload Liver Function Test Report

Drag and drop file here
Limit 200MB per file • PDF

1.pdf 68.4KB

Liver Function Test Report

| | Age | Sex | AST (SGOT) | ALT (SGPT) | AST:ALT Ratio | GGTP | ALP | Bilirubin Total | Bilirubin Dire |
|--|-----|------|------------|------------|---------------|------|-----|-----------------|----------------|
| | 80 | Male | 288 | 110 | 2.62 | 109 | 149 | 3.03 | 0.32 |

You have Fatty Liver Disease!!

Fig.7.2 Result of Liver Disease prediction

The Liver Disease Prediction System was accurate to 96% using Random Forest and 92% Logistic Regression with to ensure safe liver disease classification. Data augmentation brought the dataset size from 500 to 5000 records to enhance model generalization. Precision, recall, specificity, and F1-score confirmed the model's accuracy to identify Hepatitis, Liver Fibrosis, Cirrhosis, and Fatty Liver. User LFT values compared graphically with healthy values help in medical analysis. It further encompasses automated extraction of text that minimizes errors by humans and increases diagnostic efficiency.

VIII. CONCLUSION

The Liver Disease Prediction System uses machine learning algorithms for early and precise prediction of liver diseases from Liver Function Test (LFT) reports. The system uses Logistic Regression (92% accuracy) and Random Forest (96% accuracy), with data augmentation increasing the dataset from 500 to 5000 records to improve model generalization. Feature scaling, normalization, and missing value treatment ensure model strength for real-world clinical use. Performance analysis based on precision, recall, specificity, and F1-score validates its diagnostic efficacy. The system also applies graphical plots of predicted vs. actual values for support in medical decision-making. The system also has automated text extraction for smooth report processing. Through the synergy of machine learning, predictive analytics, and AI-based automation, this system increases diagnostic efficiency, minimizes manual error, and enables timely medical intervention.

IX. FUTURE SCOPE

Future improvements to the Liver Disease Prediction System can focus on integrating ensemble learning techniques such as XGBoost and Random Forest to enhance classification accuracy. Implementing deep learning models like CNNs and RNNs can be explored if larger datasets become available for training. Feature selection techniques, including Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE), can optimize model performance by reducing redundant data. Hyperparameter tuning using grid search and Bayesian optimization can further refine the model's predictive accuracy. Incorporating multi-modal data, such as ultrasound imaging and genetic markers, can improve disease classification. Developing real-time adaptive learning models can enable continuous updates based on new patient data.

REFERENCES

- [1] Patel, A., Sharma, R., & Mehta, K. (2018). Evaluation of Liver Function Test Parameters in Early Detection of Liver Disease. *International Journal of Medical Research*, 12(3), 112-119.
- [2] Haque et al. (2022): Haque, S., & Ahsan, K. (2022). Hybrid machine learning model combining SVM and ANN for liver disease classification. *Journal of Healthcare Engineering*, 2022, 1-12.
- [3] Singh & Gupta (2023): Singh, P., & Gupta, R. (2023). Explainable artificial intelligence (XAI) for liver disease prediction using public health data. *Artificial Intelligence in Medicine*, 139, 102492.
- [4] Maria Alex Kuzhippallil, Carolyn Joseph, Kannan A, "Comparative Analysis of Machine Learning Techniques for Indian Liver Disease Patients" © IEEE 2020.
- [5] Hartatik Hartatik, Mohammad Badru Tamam, and Arief Setyanto, "Prediction for Diagnosing Liver Disease in Patients using KNN and Naïve Bayes Algorithms," Oct. 2020.
- [6] T. M. Ghazal, A. U. Rehman, M. Saleem, M. Ahmad, S. Ahmad, and F. Mehmood, "Intelligent Model to Predict Early Liver Disease using Machine Learning Technique," *IEEE Xplore*, Feb. 01, 2022.
- [7] F. Rahman, M. Javed, Zarrin Tasnim, J. Roy, and Syed Akhter Hossain, "A Comparative Study On Liver Disease Prediction Using Supervised Machine Learning Algorithms," vol. 8, no. 11, pp. 419–422, Nov. 2019.
- [8] V. Durai, "Liver Disease Prediction using Machine Learning Algorithms," *International Journal of Engineering and Advanced Technology*, vol. 8, no. 6, pp. 2532–2534, Aug. 2019.
- [9] Chappidi Suryaprakash Reddy, Lamu Samuel Kiran, and Xavier, "Comparative Analysis of Liver Disease Detection using Diverse Machine Learning Techniques," May 2022.
- [10] N. Afreen, R. Patel, M. Ahmed, and Mustafa Sameer, "A Novel Machine Learning Approach Using Boosting Algorithm for Liver Disease Classification," Oct. 2021.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details