

Liver Disease Prediction Using Optimized Feature Selection and Data Balancing Techniques

Mr. B. Upender¹, Bugga Venu Kumar², Dendi Sai Prakash Reddy³, Allapuram Dheeraj Reddy⁴, Katiki Sandhya⁵

¹Assistant Professor, Department of Information Technology, TKR College of Engineering and Technology, Telangana, India

²³⁴⁵Department of Information Technology, TKR College of Engineering and Technology, Telangana, India

Abstract - The liver is an essential organ in human body that undertakes detoxification and metabolism processes in the body. Liver disorders can be influenced by different factors, which include viral infections, genetic diseases, high alcohol consumption, and toxins. The symptoms may vary from one individual to another hence; early detection is not always achievable. In this case, the delay in diagnosis makes it difficult to treat and leads to severe health issues. Thus, early prediction of liver disorders is a critical requirement. Most of the current systems employ basic preprocessing with standard machine learning models like Decision Trees, SVM, and Random Forest. However, there are issues of imbalanced data, noisy records, and limited selection of features, which make them less accurate and poorly predictive. The proposed system addresses these challenges with the help of advanced preprocessing optimized feature selection using RFE, and hybrid data balancing using SMOTE-ENN. It uses powerful boosting algorithms like LightGBM for improving the accuracy and robustness of the prediction. In addition, SHAP provides interpretability for the model to be more reliable and clinically useful. The enhanced approach ensures faster, more accurate, and efficient liver disease prediction.

Key Words: Liver Disease Prediction, Machine Learning, LightGBM, Recursive Feature Elimination (RFE), SMOTE-ENN, Explainable AI (XAI), SHAP.

1. INTRODUCTION

Liver disease is a serious concern in global health; millions of people are affected every year. The liver is a vital organ which plays a crucial role in metabolism, detoxification, and nutrient regulation. Therefore, any damage to the liver can cause life-threatening complications. It is important to detect liver diseases early because, in the initial stages, the symptoms are not noticeable, and it becomes difficult for medical professionals to provide an accurate diagnosis.

Machine learning has recently been recognized as a promising approach in medical diagnosis. However, in real-world medical applications, including liver disease, the class distribution is imbalanced, there are missing values, noise, and irrelevant features, which makes the performance of traditional machine learning models less reliable. Many existing models, such as Decision Trees, SVM, Logistic

Regression, and Random Forest, are not able to provide accurate results when dealing with complex data.

To overcome these limitations, this study proposes an enhanced liver disease prediction system that combines state-of-the-art machine learning algorithms with optimized data processing. The proposed system uses Recursive Feature Elimination (RFE) to select the most important clinical features and balances the data using the SMOTE-ENN hybrid approach to remove noise. Moreover, efficient boosting algorithms like LightGBM are used to achieve high prediction accuracy, training speed, and robustness. To ensure clinical validity, Explainable AI (XAI) techniques such as SHAP are used to provide clear and interpretable results of the machine learning model.

By leveraging optimized feature selection, hybrid data balancing, and high-performance machine learning algorithms, this system is expected to provide a more accurate, reliable, and interpretable solution for early liver disease prediction, which can help healthcare professionals make more informed decisions.

1.1 Machine Learning-Based Liver Disease Prediction

Machine learning-based liver disease prediction aims at processing clinical and laboratory information to detect liver diseases at an early stage. Liver diseases are hard to diagnose because of their diverse symptoms, imbalanced patient information, noisy medical records, and presence of irrelevant features. These issues make traditional diagnosis and statistical analysis less effective.

In machine learning-based prediction models, complex patterns are learned from patient data and predictions are made for new patients. Boosting models like LightGBM are more appropriate for medical data due to their efficiency in dealing with non-linear patterns, missing values, and high-dimensional features. These models are faster to train and more accurate than traditional classifiers.

To make predictions more reliable, optimized feature selection methods like Recursive Feature Elimination (RFE) are used to detect the most important clinical features that

affect liver disease. Moreover, hybrid data balancing methods like SMOTE-ENN are used to balance the class distribution by creating new minority samples and eliminating noisy data points. By combining machine learning with optimized data processing and explanation methods, liver disease prediction models can be used for early-stage diagnosis and decision support for medical professionals.

1.2 Challenges in Existing Liver Disease Prediction Systems

Despite the growing use of machine learning techniques in liver disease diagnosis, several challenges still limit their effectiveness in real-world clinical environments. One major issue is the presence of imbalanced datasets, where the number of healthy patient records is significantly higher than that of liver disease cases. This imbalance often leads to biased models that favor the majority class, resulting in poor detection of affected patients.

Another critical challenge is data quality. Medical datasets frequently contain missing values, noise, and outliers due to manual data entry, equipment limitations, or patient variability. These issues negatively impact the learning process of machine learning models and reduce prediction accuracy. Additionally, many datasets include redundant or irrelevant features that do not contribute meaningfully to liver disease prediction, increasing computational complexity and reducing model performance.

Traditional machine learning models such as Decision Trees, Support Vector Machines (SVM), Logistic Regression, and Random Forest often struggle to handle these challenges effectively. They usually rely on basic preprocessing and simple feature selection techniques, which are insufficient for complex and high-dimensional medical data. Furthermore, most existing models lack interpretability, making it difficult for healthcare professionals to understand and trust the predictions produced by these systems.

Addressing these challenges requires advanced preprocessing techniques, optimized feature selection, effective data balancing methods, and interpretable machine learning models that can provide reliable and clinically meaningful predictions.

2. PROPOSED SYSTEM

The proposed system is designed as a structured and sequential framework for early prediction of liver disease using advanced machine learning techniques. It focuses on improving prediction accuracy, handling real-world medical data challenges, and providing interpretability to support clinical decision-making. The overall workflow of the proposed system is illustrated in Fig-1, which shows the major stages involved from data collection to final prediction.

2.1 System Architecture

The architecture of the proposed system follows a step-by-step pipeline starting from dataset acquisition and ending with the prediction output. As shown in Fig-1, the system consists of the following major components: dataset collection, data preprocessing, feature selection, data balancing, model training, performance evaluation, explainability, and prediction output.

Initially, the liver disease dataset is collected and analyzed to understand its structure and characteristics. Data preprocessing is then performed to handle missing values, remove noise, normalize features, and improve overall data quality. This step is essential to ensure reliable model training and accurate predictions.

After preprocessing, feature selection is applied using Recursive Feature Elimination (RFE) to identify the most relevant clinical attributes related to liver disease. Selecting optimal features helps reduce dimensionality, improve model efficiency, and enhance predictive performance.

To address the issue of class imbalance commonly present in medical datasets, the system employs a hybrid data balancing technique known as SMOTE-ENN. This approach generates synthetic minority class samples while removing noisy and misclassified instances, resulting in a balanced and cleaner dataset suitable for training.

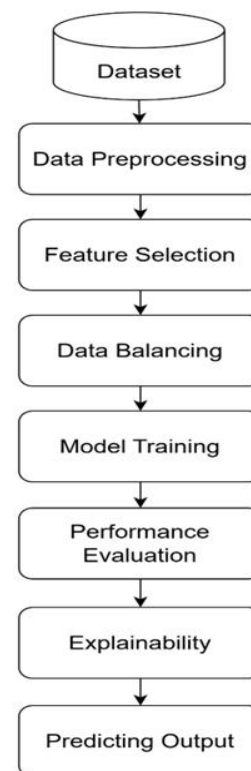


Fig -1: Architecture of the proposed liver disease prediction system.

2.2 Model Training, Evaluation, and Explainability

Once the dataset is balanced and optimized, it is used to train the LightGBM model. LightGBM is a powerful boosting algorithm that provides fast training, high accuracy, and robustness when dealing with large and complex medical datasets. It effectively captures non-linear relationships between clinical features and liver disease outcomes.

After model training, performance evaluation is carried out using standard metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to assess the effectiveness of the proposed system. These metrics help in comparing the proposed approach with existing traditional machine learning models.

To improve trust and transparency, Explainable Artificial Intelligence (XAI) techniques using SHAP are integrated into the system. SHAP explains the contribution of each feature to the final prediction, both at global and individual levels, making the model's decisions more interpretable and clinically meaningful.

Finally, the trained model generates the prediction output, indicating whether a patient is likely to have liver disease or not. This comprehensive and interpretable framework makes the proposed system suitable for real-world clinical decision support.

3. IMPLEMENTATION DETAILS

This section explains the detailed implementation of the proposed liver disease prediction system. It describes the tools, preprocessing steps, and feature selection process, data balancing strategy, model training procedure, and explainability integration used to build a robust and reliable prediction framework. The implementation is designed to handle real-world medical data challenges such as noise, imbalance, and lack of interpretability.

3.1 Tools and Technologies Used

The proposed system is implemented using the Python programming language due to its flexibility and wide adoption in machine learning and data science applications. Python provides a rich ecosystem of libraries that support efficient data processing, model development, and evaluation.

The NumPy library is used for numerical computations and array operations, while Pandas is employed for data loading, cleaning, and manipulation. Scikit-learn is used for implementing preprocessing techniques, Recursive Feature Elimination (RFE), performance evaluation metrics, and baseline machine learning models. The LightGBM library is utilized for training the gradient boosting model due to its efficiency and superior performance on structured medical datasets. For model interpretability, the SHAP library is

integrated to explain predictions and analyze feature contributions.

The implementation is carried out in a standard Python development environment, ensuring reproducibility and scalability for future enhancements.

3.2 Data Preprocessing Implementation

Data preprocessing is a critical step in the implementation, as medical datasets often contain missing values, noisy records, and inconsistent feature scales. Initially, the dataset is examined to identify missing values and data inconsistencies. Missing values are handled using suitable imputation techniques such as median replacement, depending on the nature of the feature.

Noise and outliers in the data are identified using statistical analysis and removed to prevent negative impact on model training. Numerical features are normalized to bring all attributes to a common scale, which helps improve convergence and stability during model training. Categorical features, if present, are encoded using appropriate encoding techniques.

This preprocessing stage ensures that the dataset is clean, consistent, and suitable for advanced machine learning algorithms, thereby improving prediction accuracy and robustness.

3.3 Feature Selection Using Recursive Feature Elimination

To reduce dimensionality and improve model efficiency, Recursive Feature Elimination (RFE) is implemented as an optimized feature selection technique. RFE works by training a model on the full feature set and iteratively removing the least important features based on their contribution to prediction performance.

During each iteration, features with the lowest importance scores are eliminated, and the model is retrained on the reduced feature set. This process continues until an optimal subset of features is obtained. By selecting only the most relevant clinical attributes, RFE helps reduce computational complexity, avoid overfitting, and enhance the generalization capability of the model.

The selected features represent the most influential factors contributing to liver disease prediction, making the model both efficient and interpretable.

In addition, feature selection plays a crucial role in reducing model complexity and improving predictive reliability, especially when dealing with high-dimensional medical datasets. By eliminating redundant and less informative attributes, RFE helps minimize overfitting and enhances the model's ability to generalize to unseen patient data. Fig-2

illustrates the correlation among clinical features, highlighting the presence of strongly correlated variables in liver function tests. Based on this analysis, RFE effectively selects the most relevant features that contribute significantly to liver disease prediction, thereby improving both computational efficiency and model interpretability.

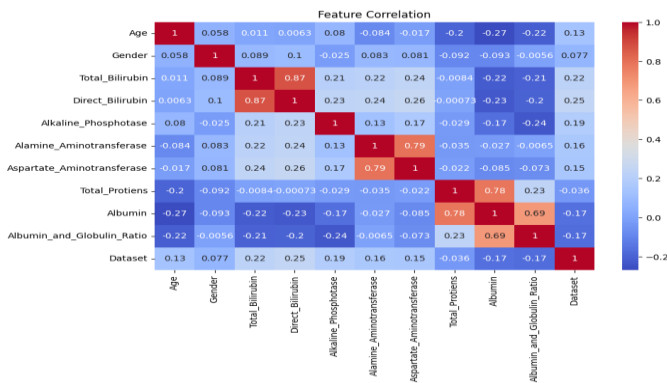


Fig-2: Feature Correlation Heatmap

3.4 Data Balancing Using SMOTE-ENN

One of the major challenges in liver disease prediction is class imbalance, where healthy patient records significantly outnumber diseased cases. To address this issue, the SMOTE-ENN hybrid data balancing technique is implemented.

SMOTE (Synthetic Minority Over-sampling Technique) generates new synthetic samples for the minority class by interpolating between existing samples. This increases the representation of diseased cases in the dataset. However, oversampling alone can introduce noise. Therefore, ENN (Edited Nearest Neighbors) is applied after SMOTE to remove misclassified and noisy samples based on nearest neighbor analysis.

The combination of SMOTE and ENN results in a balanced and cleaner dataset that improves model learning and reduces bias toward the majority class. This hybrid approach enhances prediction reliability and robustness.

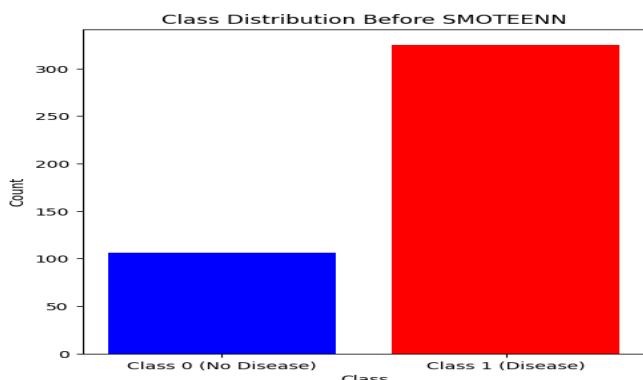


Fig-3: Class Distribution of Patients Before SMOTEENN

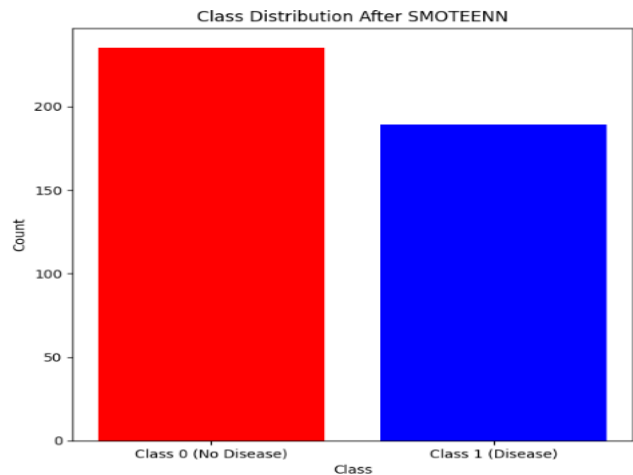


Fig-4: Class Distribution of Patients After SMOTEENN

3.5 Model Training, Tuning, and Explainability

After preprocessing, feature selection, and data balancing, the optimized dataset is used to train the LightGBM model. LightGBM is chosen due to its fast-training speed, low memory usage, and strong performance on high-dimensional medical data. It uses a gradient boosting framework that efficiently captures complex non-linear relationships among features.

Model hyperparameters such as learning rate, number of estimators, and maximum depth are tuned to achieve optimal performance. The trained model is then evaluated using standard performance metrics to ensure reliability.

To improve transparency and clinical trust, SHAP is integrated into the implementation. SHAP explains model predictions by assigning importance values to each feature, both globally and for individual predictions. This enables healthcare professionals to understand why a particular prediction was made, making the system more interpretable and suitable for clinical decision support.

4. PERFORMANCE EVALUATION AND METRICS

This section evaluates the effectiveness of the proposed liver disease prediction system using standard performance metrics and experimental analysis. The evaluation aims to measure the accuracy, robustness, and reliability of the proposed LightGBM-based model and compare its performance with traditional machine learning approaches.

4.1 Evaluation Metrics

To comprehensively evaluate the performance of the proposed system, multiple evaluation metrics are used. Accuracy measures the overall correctness of predictions by calculating the ratio of correctly predicted instances to the

total number of instances. However, accuracy alone is insufficient for medical datasets due to class imbalance.

Precision is used to measure the proportion of correctly predicted liver disease cases among all predicted positive cases. Recall, also known as sensitivity, measures the model's ability to correctly identify actual liver disease patients. A high recall value is particularly important in medical diagnosis, as it reduces the risk of missing affected patients.

The F1-score is calculated as the harmonic mean of precision and recall, providing a balanced evaluation of the model's performance. Additionally, the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) are used to evaluate the model's ability to distinguish between diseased and non-diseased cases across different classification thresholds. These metrics together provide a reliable and comprehensive assessment of the proposed system.

4.2 Experimental Setup

The experimental evaluation is conducted using the liver disease dataset after preprocessing, feature selection, and data balancing. The dataset is divided into training and testing sets to evaluate the generalization capability of the model. The training data is used to build the model, while the testing data is used to assess its predictive performance on unseen samples.

Before model training, Recursive Feature Elimination (RFE) is applied to select the most relevant clinical attributes. The SMOTE-ENN technique is then used on the training data to balance class distribution and remove noise. This ensures that the model is trained on a clean and representative dataset.

The LightGBM model is trained using optimized hyperparameters to achieve improved accuracy and robustness. For fair evaluation, the proposed model is compared with traditional machine learning classifiers such as Decision Trees, Support Vector Machines (SVM), Logistic Regression, and Random Forest under the same experimental conditions. All models are evaluated using identical performance metrics to ensure consistency.

4.3 Comparative Analysis with Existing Models

A comparative analysis is performed to demonstrate the effectiveness of the proposed approach. Traditional machine learning models often show limited performance due to their inability to handle class imbalance and complex feature relationships effectively. Models such as Decision Trees and Logistic Regression provide moderate accuracy but struggle with sensitivity and recall in imbalanced medical datasets.

In contrast, the proposed LightGBM-based system achieves higher accuracy, improved recall, and better F1-score due to

optimized feature selection and hybrid data balancing. The use of SMOTE-ENN significantly improves minority class detection, while RFE enhances feature relevance. The boosting nature of LightGBM allows it to capture non-linear patterns and interactions among clinical features more effectively than conventional classifiers.

Overall, the experimental results indicate that the proposed system outperforms traditional machine learning models in terms of predictive accuracy, robustness, and reliability, making it suitable for real-world liver disease prediction and clinical decision support.

5. RESULTS AND DISCUSSION

This section presents the experimental results obtained from the proposed liver disease prediction system and discusses their significance. The performance of the LightGBM-based model is analyzed using multiple evaluation metrics and compared with traditional machine learning approaches to demonstrate its effectiveness.

5.1 Performance Results

The proposed system achieves improved prediction performance after applying advanced preprocessing, optimized feature selection, and hybrid data balancing techniques. The LightGBM model trained on the optimized dataset demonstrates higher accuracy, precision, recall, and F1-score compared to baseline models.

The application of Recursive Feature Elimination (RFE) helps reduce irrelevant features, leading to faster training and improved generalization. Additionally, the use of SMOTE-ENN effectively balances the dataset, resulting in better minority class detection and reduced bias toward the majority class. The ROC-AUC analysis further confirms that the proposed model has a strong capability to distinguish between liver disease and non-disease cases.

Overall, the obtained results indicate that the proposed approach provides reliable and consistent predictions, making it suitable for early liver disease detection.

5.2 Feature Importance Analysis

To understand the factors influencing model predictions, SHAP-based feature importance analysis is performed. The analysis highlights key clinical attributes that contribute significantly to liver disease prediction. Features related to liver function tests show strong influence on the model's decision-making process.

SHAP provides both global and local explanations, allowing the identification of features that have the highest impact across the dataset as well as for individual predictions. This interpretability enhances trust in the model and supports its use in clinical decision-making. By understanding feature

contributions, healthcare professionals can gain valuable insights into disease patterns and risk factors.

5.3 Discussion of Results

The experimental results clearly demonstrate that the integration of optimized feature selection, hybrid data balancing, and boosting-based learning significantly improves prediction performance. Traditional machine learning models struggle with imbalanced and noisy medical data, leading to lower sensitivity and reduced reliability.

In contrast, the proposed LightGBM based framework effectively handles these challenges by leveraging SMOTE-ENN for data balancing and RFE for feature optimization. The inclusion of explainable AI techniques further strengthens the clinical applicability of the system by providing transparent and interpretable predictions.

Although the proposed system shows strong performance, it is dependent on the quality and size of the dataset. Training on larger and more diverse real-world datasets could further enhance generalization and robustness. Nevertheless, the results confirm that the proposed approach is a promising solution for early liver disease prediction.

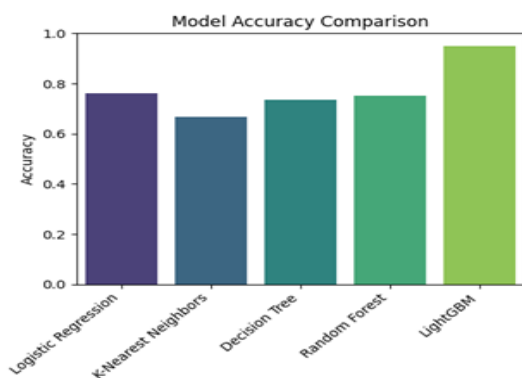


Fig -5: Models Accuracy Comparison

6. CONCLUSION AND FUTURE SCOPE

6.1 Conclusion

In this work, an efficient and interpretable machine learning-based system for early liver disease prediction has been proposed. The system addresses major challenges associated with medical data, such as class imbalance, noisy records, irrelevant features, and lack of transparency in model predictions. Advanced preprocessing techniques, optimized feature selection using Recursive Feature Elimination (RFE), and hybrid data balancing with SMOTE-ENN are employed to improve data quality and model robustness.

The LightGBM model is used as the core classifier due to its fast-training speed, scalability, and strong performance on

structured clinical data. Experimental evaluation demonstrates that the proposed approach achieves improved accuracy, recall, and F1-score compared to traditional machine learning models. Furthermore, the integration of Explainable Artificial Intelligence (XAI) using SHAP provides meaningful insights into feature importance and model decisions, enhancing clinical trust and usability.

Overall, the proposed system proves to be a reliable, accurate, and interpretable solution for early liver disease prediction and can effectively support healthcare professionals in decision-making.

6.2 Future Scope

Although the proposed system shows promising results, there are several directions for future enhancement. The model can be trained and validated on larger and more diverse real-world clinical datasets to further improve generalization and robustness. Additional clinical and lifestyle-related features may be incorporated to enhance prediction accuracy.

The system can be extended to predict different stages or types of liver disease rather than binary classification. Moreover, integrating deep learning models and ensemble approaches may further improve performance. Finally, the proposed framework can be deployed as a web-based or mobile clinical decision-support tool, enabling real-time liver disease prediction with interpretable results for practical healthcare applications.

REFERENCES

- [1] M. H. Mohamed and B. H. Ali, "Toward an accurate liver disease prediction based on a two-level ensemble stacking model," *Journal of Healthcare Engineering*, 2024.
- [2] P. M. O, M. K. Gavimath, and E. Thomas, "Liver disease prediction using ensemble technique," *International Journal of Engineering Research and Technology (IJERT)*, 2024.
- [3] E. Dritsas and M. Trigka, "Supervised machine learning models for liver disease risk prediction," *Applied Sciences*, vol. 13, no. 4, 2023.
- [4] S. R. Tanuku and A. A. Kumar, "Liver disease prediction using ensemble technique," in *Proceedings of the 8th International Conference on Computing and Communication Systems*, 2022.
- [5] S. Tokala and K. Hajarathaiyah, "Liver disease prediction and classification using machine learning techniques," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 2022.

- [6] R. Kaur and B. Singh, "Liver disease prediction using machine learning algorithms," *International Journal of Computer Applications*, 2021.
- [7] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority oversampling technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321–357, 2021.
- [8] J. Dyson and M. Hudson, "The critical importance of early detection of liver disease," *Hepatology*, vol. 72, no. 3, pp. 925–935, 2020.
- [9] G. Ke, Q. Meng, T. Finley et al., "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, vol. 31, no. 2, 2020, pp. 3146–3154.
- [10] A. Patel and R. Singh, "The role of machine learning in advancing liver disease diagnostics," *Biomedical Informatics Insights*, vol. 13, pp.1–12, 2021.