Journal of Advances in Developmental Research (IJAIDR)

E-ISSN: 0976-4844 ● Website: www.ijaidr.com ● Email: editor@ijaidr.com

Liver Disease Detection Using Machine Learning

K. Swapna[1], Ch. Spoorthi[2], K. Sirireddy[3], K. Sowmya[4], K. Harshitha[5]

[1]Assistant professor, [2,3,4,5]B. Tech 3rd year Student

[1,2,3,4,5]CSE (AI&ML), Vignan's Institute of Management and Technology for Women, Hyderabad, India

Abstract:

Liver disease is a major global health concern causing high mortality rates. Early detection plays a crucial role in improving treatment outcomes and reducing deaths. This research proposes a machine learning–based system for predicting liver disease using clinical data. The model utilizes the Random Forest algorithm for classification and applies the SMOTE + ENN technique to handle class imbalance in medical datasets. Performance is evaluated using accuracy, precision, recall, and F1-score. The proposed system significantly improves prediction accuracy and reduces false classifications compared to traditional approaches.

Keywords: Liver Disease, Machine Learning, Random Forest, SMOTE-ENN, Classification.

I. INTRODUCTION:

The liver is one of the most important organs in the human body, responsible for performing vital functions such as metabolism, detoxification, protein synthesis, and the production of biochemicals necessary for digestion. Liver diseases, including hepatitis, fatty liver, cirrhosis, and liver cancer, have become a major global health concern due to increasing lifestyle-related risk factors such as poor diet, alcohol consumption, obesity, and lack of physical activity. Early detection of liver disease is crucial, as delayed diagnosis can lead to severe complications and even death**.** The liver is one of the most important organs in the human body, responsible for performing vital functions such as metabolism, detoxification, protein synthesis, and the production of biochemicals necessary for digestion. Liver diseases, including hepatitis, fatty liver, cirrhosis, and liver cancer, have become a major global health concern due to increasing lifestyle-related risk factors such as poor diet, alcohol consumption, obesity, and lack of physical activity. Early detection of liver disease is crucial, as delayed diagnosis can lead to severe complications and even death. This research contributes to the development of an intelligent healthcare system that supports medical professionals in diagnosing liver disease more efficiently. In this study, a hybrid approach combining SMOTE and ENN is applied to preprocess the dataset effectively. Furthermore, the Random Forest algorithm, an ensemble learning method known for its robustness and high accuracy, is used for classification. The proposed system aims to enhance prediction performance, reduce false negatives, and provide a reliable tool for early detection of liver disease.

II. RELATED WORK:

Early research primarily focused on traditional machine learning algorithms such as Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Decision Trees. These models were applied to medical datasets to classify patients as either healthy or affected by liver disease. While these approaches provided a foundation for automated diagnosis, their performance was often limited due to issues such as overfitting, sensitivity to noisy data, and lack of robustness. In recent years, ensemble learning methods such as Random Forest and Gradient Boosting have gained popularity due to their ability to handle complex datasets and reduce overfitting. Random Forest, in particular, has been widely used in

healthcare applications because it combines multiple decision trees and provides stable and accurate predictions. A number of studies highlighted that medical datasets are usually imbalanced, where the number of healthy cases significantly exceeds the number of diseased cases. This imbalance leads to biased models that tend to favour the majority class, resulting in poor detection of liver disease cases. To address this challenge, researchers introduced data resampling techniques such as Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic samples for the minority class to balance the dataset. Further improvements were made by combining SMOTE with data cleaning techniques like Edited Nearest Neighbor (ENN). This hybrid approach not only balances the dataset but also removes noisy and misclassified samples, thereby improving the overall quality of the data. Studies have shown that such hybrid techniques significantly enhance model performance, especially in terms of recall and F1-score.

III. PROPOSED SYSTEM:

A. Overview of the proposed system:

The proposed system is designed to provide an efficient and accurate solution for the early detection of liver disease using machine learning techniques. It integrates data preprocessing, class imbalance handling, and ensemble learning to improve prediction performance and reliability. The system begins with the collection of patient clinical data, which may include parameters such as age, gender, bilirubin levels, enzyme levels, and other medical attributes. Since raw medical data often contains missing values, noise, and inconsistencies, a preprocessing step is applied to clean and prepare the dataset for analysis.
One of the major challenges in medical datasets is class imbalance, where the number of healthy cases significantly exceeds the number of liver disease cases. To address this issue, the proposed system employs a hybrid resampling technique combining SMOTE (Synthetic Minority Over-sampling Technique) and ENN (Edited Nearest Neighbour). SMOTE generates synthetic samples for the minority class, while ENN removes noisy and misclassified data points, resulting in a balanced and cleaner dataset.

B. Overall System Architecture:

The system architecture follows a pipeline-based design where data flows sequentially through multiple stages. Initially, medical data is collected and preprocessed to remove noise and handle missing values. The processed dataset is then balanced using the SMOTE + ENN technique to address class imbalance issues. The trained model is then used in the prediction phase, where new patient data is provided as input. The system processes this data and predicts whether the patient is likely to have liver disease. Finally, the system evaluates its performance using metrics such as accuracy, precision, recall, and F1-score to ensure reliability and effectiveness. provides a systematic and efficient approach for early liver disease detection, enabling accurate predictions and supporting healthcare professionals in decision-making.
Next, a Random Forest classifier is trained on the balanced dataset to build a robust predictive model. The trained model is used in the prediction module, where new patient data is analyzed to determine the presence of liver disease. Finally, the system evaluates performance using standard metrics and produces the prediction output.

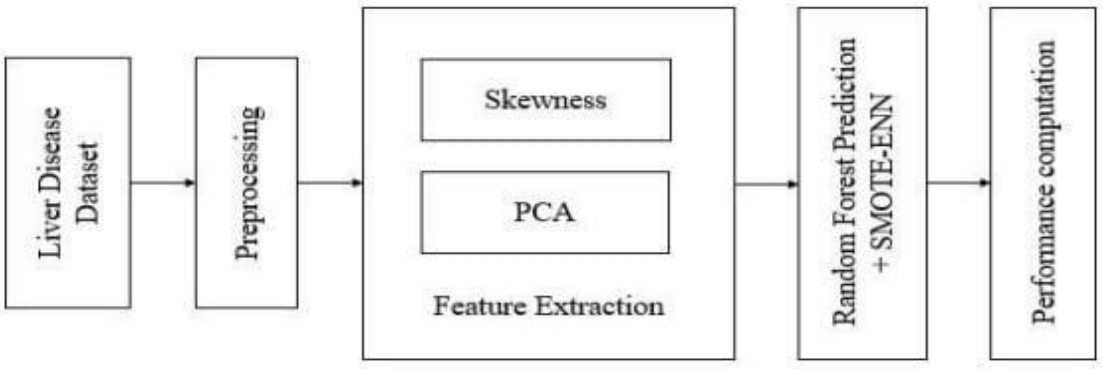


Figure: System Architecture of Live Disease Dataset

C. Data Collection Modules:

The Data Collection Module is the initial and one of the most critical components of the proposed liver disease detection system. It is responsible for gathering relevant and high-quality data required for training and evaluating the machine learning model.In this module, medical datasets containing patient information are collected from reliable sources such as hospitals, healthcare organizations, or publicly available repositories (e.g., UCI Machine Learning Repository). The dataset typically includes various clinical and biochemical parameters related to liver function, such as age, gender, total bilirubin, direct bilirubin, alkaline phosphatase, alanine aminotransferase (SGPT), aspartate aminotransferase (SGOT), total proteins, albumin, and albumin-globulin ratio.

The collected data may exist in different formats such as CSV files, databases, or spreadsheets. Therefore, the module ensures proper data acquisition and integration into a unified format suitable for further processing. It also verifies the authenticity and consistency of the data to avoid errors during model training. In addition, this module ensures that sufficient data samples are collected for both classes (liver disease and non-liver disease cases). This is important for building a robust and generalizable machine learning model. If the dataset is imbalanced, it is addressed in later stages using resampling techniques.

D. Data Preprocessing Module

The Data Preprocessing Module is a crucial stage in the liver disease detection system, as it prepares raw medical data for accurate analysis and model training. Since real-world medical datasets often contain missing values, noise, and inconsistencies, preprocessing ensures that the data is clean, structured, and suitable for machine learning algorithms.Initially, the dataset is examined to identify missing or null values. These missing values are either removed or replaced using appropriate techniques such as mean, median, or mode imputation. This step prevents errors during model training and improves reliability.

Next, data cleaning is performed to eliminate duplicate records and irrelevant information. This helps in reducing redundancy and ensures that only meaningful data is used for analysis. Noise and outliers present in the dataset are also identified and removed to improve the overall data quality. After cleaning, data transformation is carried out. This includes normalization or scaling of numerical features so that all attributes are brought to a similar range. This step is important because machine learning algorithms perform better when data is standardized.

Additionally, categorical data such as gender is converted into numerical form using encoding techniques. This allows the machine learning model to process non-numeric data effectively. Feature selection is also applied to identify the most relevant attributes that contribute to liver disease prediction. By removing unnecessary features, the system reduces complexity and improves computational efficiency.

E. Feature Extraction Module

The Feature Extraction Module plays a significant role in enhancing the performance of the liver disease detection system by identifying and transforming the most relevant features from the preprocessed dataset. In medical datasets, many attributes may be redundant, irrelevant, or less significant, which can negatively impact the efficiency and accuracy of machine learning models. Therefore, feature extraction is essential to reduce data dimensionality and improve model performance.

In this module, statistical and mathematical techniques are applied to analyze the distribution and importance of features. One of the key steps is handling skewness in the data. Medical data is often unevenly distributed, where some features have extreme values or are biased toward one side. To address this, appropriate transformations such as normalization or logarithmic scaling are applied to make the data more uniform and suitable for analysis.

Another important technique used in this module is Principal Component Analysis (PCA). PCA is a dimensionality reduction method that transforms the original set of features into a new set of uncorrelated variables called principal components. These components are arranged in such a way that the first few components retain most of the variation present in the original dataset. By selecting only the most significant components, the system reduces the number of features while preserving essential information. This process helps in eliminating redundant and less informative features, reducing computational complexity, and improving the overall efficiency of the model. Additionally, it minimizes the risk of overfitting by simplifying the dataset.

IV.IMPLEMENTATION DETAILS:

The implementation of the proposed liver disease detection system is carried out using a structured machine learning approach to ensure accurate and efficient prediction. The system is developed using the Python programming language, which provides extensive support for data analysis, preprocessing, and model development through various libraries.

Initially, the dataset containing patient clinical information is collected and loaded into the system. The data is then subjected to preprocessing to improve its quality and usability. This includes handling missing values, removing duplicate and inconsistent records, and converting categorical data into numerical format. Additionally, normalization techniques are applied to ensure that all features are on a similar scale, which enhances the performance of machine learning algorithms.Following preprocessing, feature extraction is performed to identify and retain the most relevant attributes. Techniques such as skewness handling are applied to correct uneven data distributions, while Principal Component Analysis (PCA) is used to reduce dimensionality and eliminate redundant features. This step helps in simplifying the dataset and improving computational efficiency.

To address the issue of class imbalance commonly found in medical datasets, the system employs a hybrid approach combining SMOTE (Synthetic Minority Over-sampling Technique) and ENN (Edited Nearest neighbour). SMOTE generates synthetic samples for the minority class, while ENN removes noisy and misclassified data points, resulting in a balanced and cleaner dataset.The processed data is then used to train a Random Forest classifier. This algorithm constructs multiple decision trees and combines their outputs using majority voting to produce the final prediction. The use of an ensemble method enhances accuracy, reduces overfitting, and improves the robustness of the model. After training, the model is evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score. These metrics provide a comprehensive assessment of the model's effectiveness in predicting liver disease.

V.ALGORITHM:

BEGIN
INPUT patient_data
LOAD dataset

PREPROCESS dataset
- handle missing values
- remove duplicates
- encode categorical data
- normalize features

EXTRACT features using PCA

BALANCE dataset using SMOTE + ENN

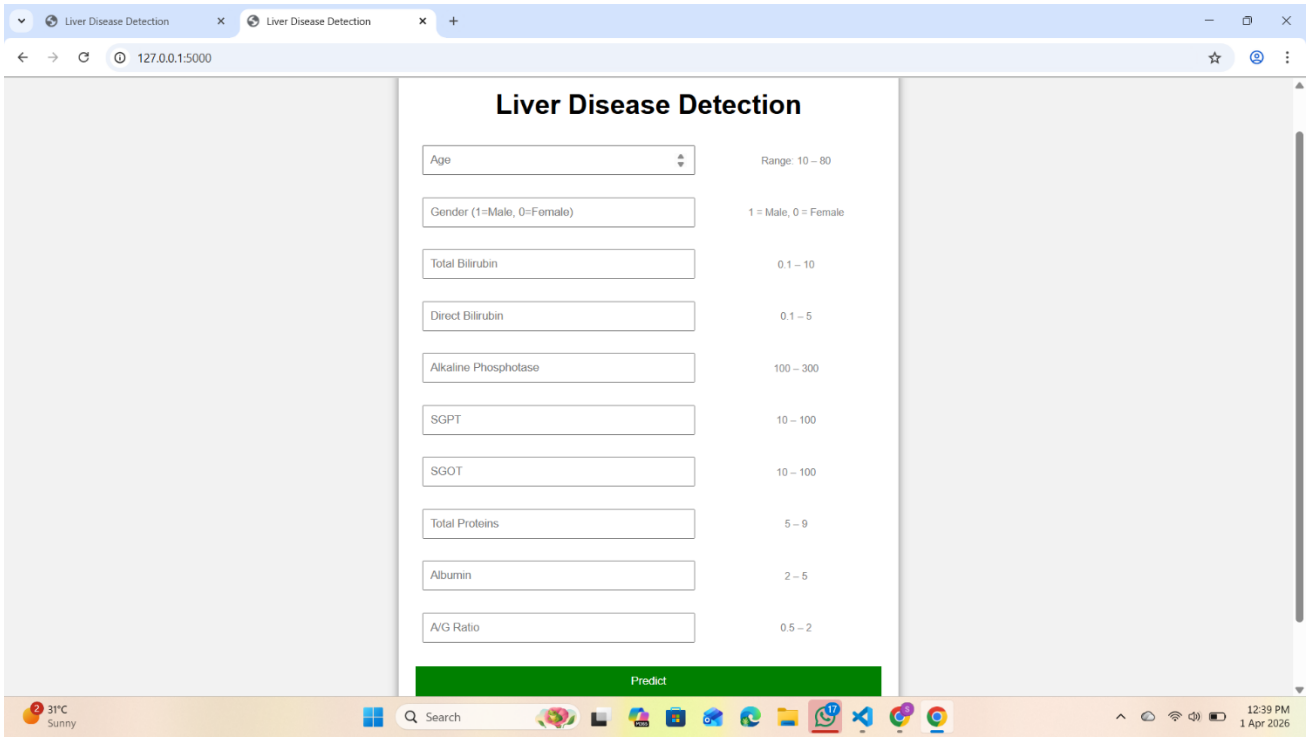
SPLIT dataset into training and testing

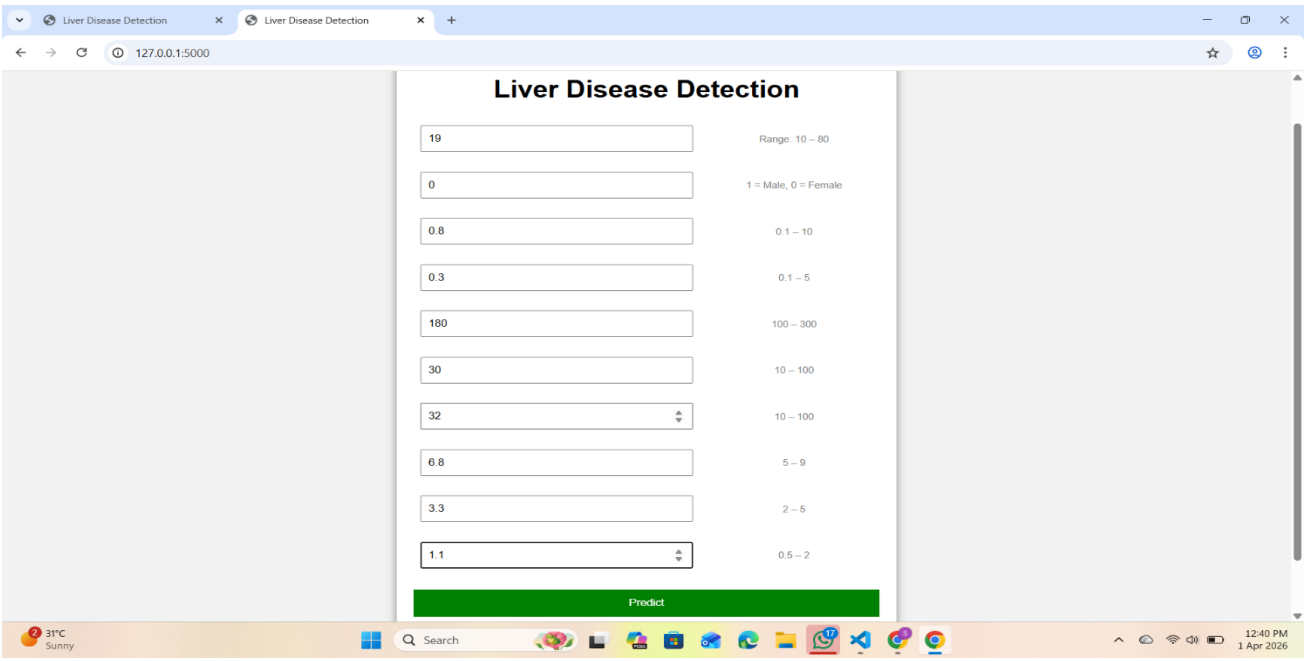
TRAIN Random Forest model

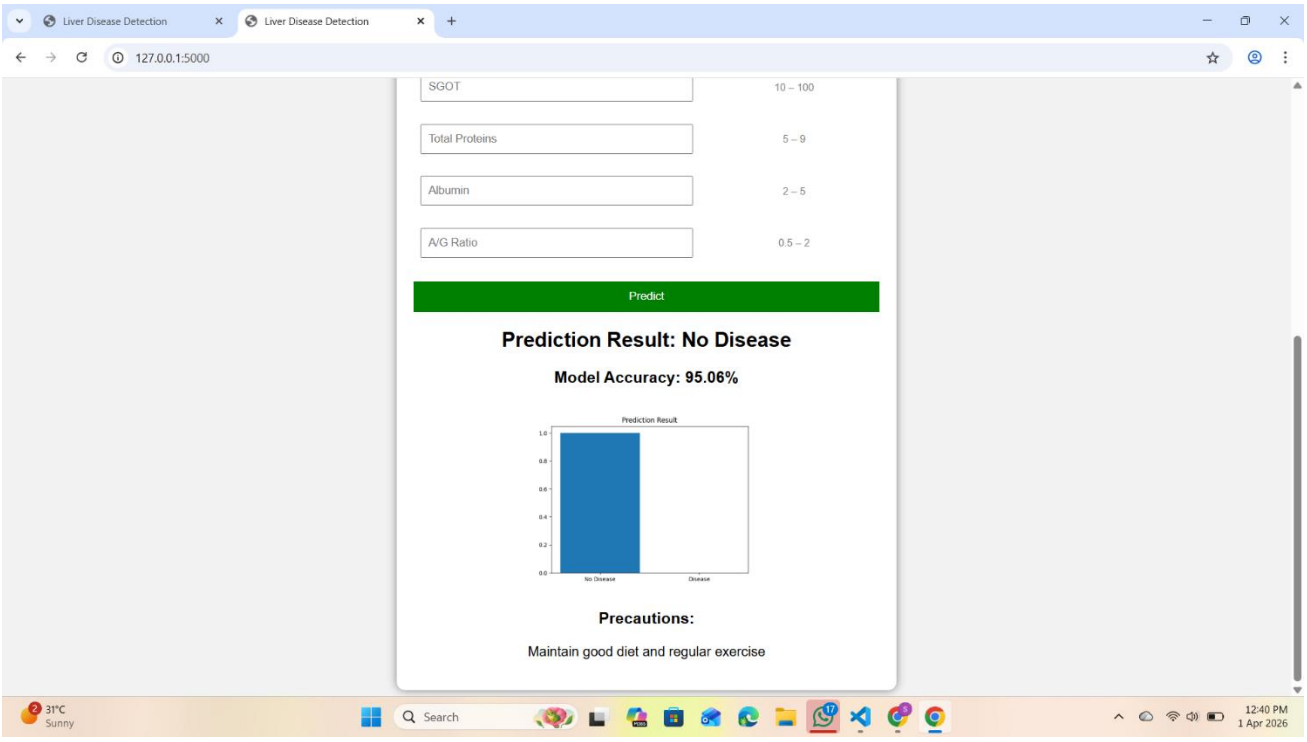
PREDICT using patient_data

IF prediction == 1 THEN
OUTPUT "Liver Disease"
ELSE
OUTPUT "No Liver Disease"
END IF
END

VII. RESULT







VIII. CONCLUSION

This research presents an effective and reliable machine learning-based approach for the early detection of liver disease using clinical data. The proposed system integrates multiple stages, including data preprocessing, feature extraction, class imbalance handling, and classification, to ensure accurate prediction. A key contribution of this work is the use of a hybrid resampling technique, SMOTE combined with ENN, which successfully addresses the problem of class imbalance commonly found in medical datasets. By generating synthetic samples for the minority class and removing noisy data points, the system improves data quality and enhances the learning capability of the model. The Random Forest classifier is employed due to its ensemble nature, which allows it to build multiple decision trees and combine their

outputs for better prediction. This approach reduces overfitting, increases stability, and provides higher accuracy compared to traditional single-model classifiers.

Experimental results demonstrate that the proposed system achieves strong performance across all evaluation metrics, including accuracy, precision, recall, and F1-score. In particular, the model shows improved recall, which is crucial in medical applications, as it ensures that most of the actual liver disease cases are correctly identified.

REFERENCES:

[1] J. Javad Hassannataj, S. Hamid, D. Abdollah and S. Shahaboddin, ―Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection‖, Informatics in Medicine Unlocked, vol. 17, 2019, 100255, no. 2352- 9148, https://doi.org/10.1016/j.imu.2019.100255.

[2] S. Ambesange, V. A, R. Uppin, S. Patil and V. Patil, ―Optimizing Liver disease prediction with Random Forest by various Data balancing Techniques‖, 2020 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM), 2020, pp. 98-102, doi: 10.1109/CCEM50674.2020.00030.

[3] M. A. Kuzhippallil, C. Joseph and A. Kannan, ―Comparative Analysis of Machine Learning Techniques for Indian Liver Disease Patients‖, 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020, pp. 778-782, doi: 10.1109/ICACCS48705.2020.9074368.

[4] S. Deshmukh, A. Lokhande, R. Wasnik and N. Singhal. ―Vacuole Segmentation and Quantification in Liver Images of Wistar Rat‖, Annu Int Conf IEEE Eng Med Biol Soc. 2020 Jul;2020:1396-1399. doi: 10.1109/EMBC44109.2020.9176500. PMID: 33018250.

[5] H. Hartatik, M. B. Tamam and A. Setyanto, ―Prediction for Diagnosing Liver Disease in Patients using KNN and Naïve Bayes Algorithms‖, 2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS), 2020, pp. 1-5, doi: 10.1109/ICORIS50180.2020.9320797.

[6] R. Kalaiselvi, K. Meena and V. Vanitha, ―Liver Disease Prediction Using Machine Learning Algorithms‖, 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), 2021, pp. 1-6, doi: 10.1109/ICAECA52838.2021.9675756.

[7] G. Shobana and K. Umamaheswari, ―Prediction of Liver Disease using Gradient Boost Machine Learning Techniques with Feature Scaling‖, 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 1223- 1229, doi: 10.1109/ICCMC51019.2021.9418333.

[8] Fahad Mostafa, Easin Hasan, Morgan Williamson and Hafiz Khan. Stastical Machine Learning Approaches to Liver Disease Prediction 2021,1,294-312.

[9] Rong-Ho Lin. An intelligent model for liver disease diagnosis. Artificial Intelligence in Medicine 2009;47:53—62.

[10] Schiff's Diseases of the Liver, 10th Edition Copyright ©2007 Lippincott Williams & Wilkins by Schiff, Eugene R.; Sorrell, Michael F.; Maddrey, Willis C.

[11] Michael J. Sorich,† John O. Miners,*,‡ Ross A. McKinnon,† David A. Winkler,§ FrankR. Burden,| and Paul A. Smith‡ Comparison of linear and nonlinear classification algorithms for the prediction of drug and chemical metabolism by human UDP- Glucuronosyltransferase Isoforms.

[12] N. A. Shackel, D. Seth, P. S. Haber, M. D. Gorrell and G. W. McCaughan. ―The hepatic transcriptome in human liver disease. Comp Hepatol‖, 2006 Nov 7;5:6. doi: 10.1186/1476-5926-5-6. PMID: 17090326; PMCID: PMC1665460.

[13] R. H. Lin. ―An intelligent model for liver disease diagnosis‖, Artif Intell Med. 2009 Sep;47(1):53-62. doi: 10.1016/j.artmed.2009.05.005. Epub 2009 Jun 21. PMID: 19540738.

[14] Y. Kumar and G. Sahoo. ―Prediction of different types of liver diseases using rule based classification model‖. Technol Health Care. 2013;21(5):417-32. doi: 10.3233/THC- 130742. PMID: 23963359.
[15] S. Sontakke, J. Lohokare and R. Dani, ―Diagnosis of liver diseases using machine learning,‖, 2017 International Conference on Emerging Trends & Innovation in ICT (ICEI), 2017, pp. 129-133, doi: 10.1109/ETIICT.2017.7977023.