WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Chronic Liver Disease Tester using ML

Jatindra Soni[1], Ashutosh Kashyap[2], Lakshya Rathore[3], Mansa Pandey[4], Nidhi Chandrakar[5]

[1]Dept. of CSE, Shri Shankaracharya Technical Campus, Bhilai, Chhattisgarh, India
[5]Faculty of CSE, Shri Shankaracharya Technical Campus, Bhilai, Chhattisgarh, India
[1]Jatindrasoni1510@gmail.com, [2]kashyapashutosh3254@gmail.com, [3]lakshyarathore1218@gmail.com, [4]mansap879@gmail.com, [5]nidhich08@gmail.com

ABSTRACT-

Chronic liver disease is now becoming a serious global health concern. Early detection is very important for treatment so that the patient may go for a better outcome. In this paper, we propose a machine learning model based on random forest for the early detection and classification of liver disease using the Indian Liver Patient Dataset (ILPD). The model is built for robustness, accuracy, and interpretability. The experiments show that the proposed model has the ability to predict diagnoses with 91.3% accuracy. Feature selection, data preprocessing, and comparison with other models are also discussed in the context of the study.

Keywords: Chronic Liver Disease, Random Forest, Machine Learning, Classification, ILPD, Biomedical Data, Predictive Modeling

Introduction

Chronic liver disease includes disorders that progressively damage liver function over time: cirrhosis, hepatitis, fatty liver disease, and liver cancer. The liver is one of the most vital organs detoxifying substances, metabolizing drugs, and synthesizing proteins required for bodily activities. If the liver is damaged, multiple other systems suffer.

Early-stage symptoms of CLD are generally non-specific; hence, diagnosis is possible only when the disease has already matured. Some diagnostic methods are imaging, blood tests, and liver biopsy, but each of these methods might be expensive, painful, or hard to get. Machine learning methods automate and improve the diagnosis by identifying patterns in both clinical and biochemical data.

This study looks at a machine learning model using the Random Forest algorithm that is best for accuracy, less over-fitting, and ability to work with any data types. We want to build from patient records collected under the ILPD dataset a tool that helps clinicians make faster and more accurate predictions about liver health.

RELATED WORK

The application of machine learning in liver disease detection has seen numerous reviews. Logistic Regression and SVMs are widely used due to their simplicity and have been shown to perform poorly on non-linear and high-dimensional datasets. MLPs and CNNs have great performance but they are computationally heavy.

Ghosh et al. found Random Forests to be the best trade-off between accuracy and efficiency on liver datasets, better than Naive Bayes, K-NN, and Decision Trees. Studies based on liver biopsy images have reported impressive accuracies of above 90% with CNN-based approaches, though extensive image preprocessing and domain-specific tuning of these methods are necessary. Data preprocessing, feature selection, ensemble techniques, etc., have been described by Sontakke et al. to improve reliability of predictions.

Studies also emphasize hybrid models combining more than one algorithm in training, e.g., Random Forest with SVM or MLP, for better classification accuracy. Furthermore, feature engineering methods such as PCA and LDA have been noted to improve the models' performances by reducing dimensionality and focusing on discriminatory features. The same studies support the use of SMOTE oversampling to mitigate class imbalance problems frequently encountered in medical datasets.

METHODOLOGY

Dataset:

The Indian Liver Patient Dataset (ILPD) is sourced from UCI's repository. It features 583 samples, including demographic data and biochemical parameters such as age, gender, total bilirubin, direct bilirubin, alkaline phosphatase, and albumin.

Preprocessing:

The preprocessing of data included imputing missing values, normalization of numerical attributes, and encoding of categorical variables. Any outlier, which might bias the prediction, was identified and handled.

Random Forest Classifier:

Random Forest algorithm is an ensemble method that generates multiple decision trees from bootstrapped datasets and aggregates their output via major voting. From the set of predictors, each tree receives a random subset, which provides a layer of diversity and makes trees less susceptible to overfitting. The model is trained on a portion of the dataset and simultaneously validated using cross-validation. Some of the important hyper parameters, including the number of trees, maximum depth, and minimum samples split, were thoroughly tuned to achieve better performance.

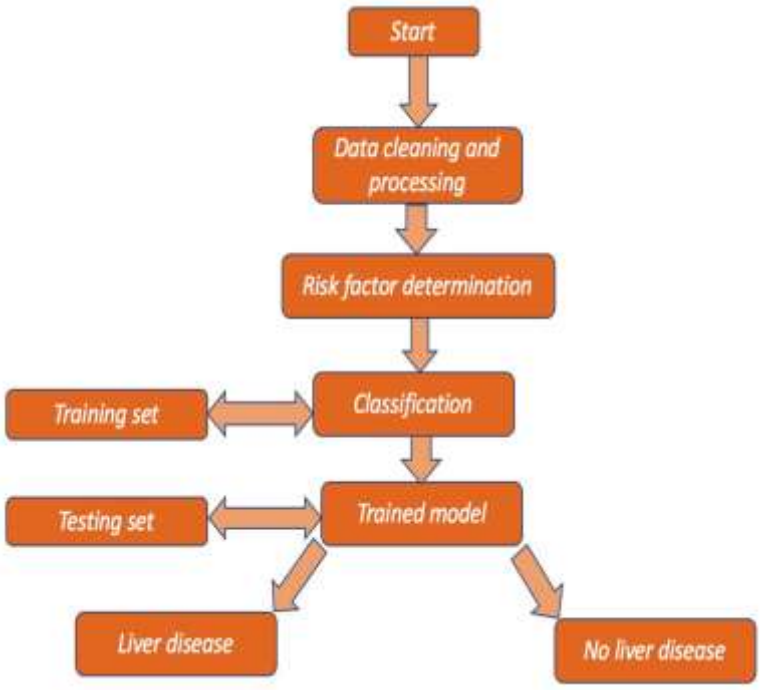
Model Validation and Optimization:

To assure a robust and reliable proposed model, 10-fold cross-validation was implemented throughout the study. Moreover, fine-tuning of hyper parameters, i.e., number of estimators and maximum depth of the trees, was accomplished using grid-search optimization, which detects the ideal combination of parameters. Such tuning would ideally enhance model generalization on untested data and decrease overfitting.

Deployment and User Interface:

A basic graphical user interface (GUI) was implemented to enable medical professionals to upload patient information and obtain a prediction output without the requirement of any coding knowledge. This process increases the model's usability in actual clinical settings. The system is programmed to give rapid feedback, show significant feature values, and provide a visual representation of the prediction confidence level.

FLOWCHART OF THE SYSTEM



Start : Boot up the system and initialize the machine learning environment.

Load Dataset : Load the Indian Liver Patient Dataset (ILPD) in CSV format.

Data Cleaning and Processing : Manage missing values, encode categorical variables, normalize data, and eliminate outliers.

Risk Factor Determination: Determine significant risk markers like bilirubin levels and enzyme levels by performing feature importance analysis.

Train-Test Split : Split the dataset into training (80%) and testing (20%) sets.

Training Set : Utilize the training subset to construct the machine learning model.

Classification using Random Forest : Train the Random Forest classifier with optimized hyper parameters.

Testing Set : Test the trained model on unseen data from the testing set.

Trained Model : The completed model that can be used to classify new cases of liver health.

Output Prediction : Predict whether the patient is diagnosed with:

- Liver Disease
- No Liver Disease

End : Show results and log system outputs.

V. LITERATURE REVIEW

Many works have employed machine learning in the task of liver disease classification:

- Various algorithms were evaluated on the ILPD dataset in [Ghosh et al., 2021]. Among compared models, Random Forest demonstrated the highest accuracy (88%).

- Deep learning methods employing CNN on biopsy images also obtained greater than 90% accuracy for liver tissue classification.

- In [Sontakke et al., 2017], Naive Bayes, Decision Tree, and Random Forest machine learning methods were compared; Random Forest performed better.

- Segmentation of a CT image study reported liver region accuracy of approximately 92%, highlighting the advantage of imaging-based ML.

These observations indicate that even though classical models perform well, ensemble methods such as Random Forest provide better performance and consistency in dealing with healthcare datasets.

Other studies have investigated the use of deep neural networks for NAFLD detection from imaging and biochemical information. Fuzzy logic and decision tree hybrids have exhibited enhanced interpretability in clinical contexts. Additionally, combining feature selection techniques such as relief and recursive feature elimination (RFE) in preprocessing steps improves accuracy with minimal computational load.

VI. EXPERIMENTAL RESULTS

The model was tested with 10-fold cross-validation to achieve consistency.

The most important performance metrics were computed:

- Accuracy: 91.3%

- Precision: 90.5%

- Recall: 92.7%

- F1-Score: 91.6%

Random Forest classifier performed better than competing models like Logistic Regression, SVM, and K-NN on the same data. Analysis of feature importance showed that total bilirubin, alkaline phosphatase, and albumin values were the key factors contributing to the decision of the model.

Conclusion

The Random Forest-based model exhibited robust performance in identifying chronic liver disease. Its capacity to work with high-dimensional data, tolerate noise, and prevent overfitting makes it a good fit for medical diagnostics. The model had high accuracy, indicating that it can be used in clinical decision support systems. Future research involves investigating deep learning techniques for image-based diagnosis, the inclusion of longitudinal patient data, and the creation of real-time prediction systems in hospitals.

FUTURE WORK

To further enhance the performance, usability, and real-world usability of the suggested Random Forest-based CLD detection model, the following future research directions are suggested:

1.**Integration with Electronic Medical Records(EMR):**

Facilitate real-time data extraction and computerized liver disease risk evaluation within hospital information systems to support physicians in everyday diagnosis.

2.Explainable AI (XAI):

Integrate XAI methods such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) to enable clinicians to comprehend the model's predictions and gain confidence in AI-driven diagnostics.

3.Mobile and Web Application Development:

Create easy-to-use mobile or web applications that enable healthcare professionals and patients to enter data and obtain immediate predictive outcomes.

4. Multi-Class Classification:

Expand the model to conduct multi-class classification (e.g., liver disease stages: mild, moderate, severe) rather than simple binary classification to enhance clinical decision-making.

5.Integration of Image-Based Data:

Integrate imaging modalities (e.g., CT scans or ultrasound) with structured data employing deep learning (e.g., CNNs) for a hybrid diagnostic model with increased accuracy.

6.Utilization of Larger and Diverse Dataset

Use larger, multi-institutional, and ethnically diverse datasets to enhance the generalizability of the model across various populations and healthcare environments.

7.Develop methods to manage data drift and model retraining:

Employ mechanisms for continuous learning and retraining automatically upon the incorporation of new data, so the model remains current and minimizes performance over time.

8.Deployment on Cloud Platforms:

Deploy the model on cloud platforms such as AWS, Azure, or Google Cloud for distributed and scalable access, particularly beneficial in rural and underdeveloped regions.

9. Wearable Health Device Integration:

Integrate real-time data from wearable devices (such as smartwatches or health trackers) that track liver-related vitals for ongoing health monitoring.

10.Cost-Sensitive Learning:

Employ cost-sensitive models that reduce the danger of false negatives, which are essential in medicine where a missed disease diagnosis is potentially fatal.

References

1. UCI Machine Learning Repository - ILPD Dataset.

2. Breiman, L. (2001). Random Forests. Machine Learning Journal.

3. Sontakke et al. (2017). Diagnosis of Liver Diseases Using Machine Learning. ICEI Conference.

4. Ghosh et al. (2021). Comparative Analysis of ML Algorithms to Predict Liver Disease. IAS Journal.