Enhancing Early Diabetes Detection: A Data-Driven Approach with the Indian Liver Patients Dataset

Alperen Dagli · Taha Ilgar

Department of Engineering, Istanbul Aydin University

COM511: Data Science

Dr. Hayder Ali Abdullah Mohammedqasim

December, 14, 2023

# Abstract

# 1 Introduction

The burden of liver disease in India is both profound and escalating, with recent figures painting a stark picture of a health crisis in motion. As the largest internal organ and a critical nexus for metabolic processes, the liver's health is indispensable to overall well-being. Yet, liver diseases remain a leading cause of disability and death globally, with India witnessing a rapid spread of these conditions—an epidemic that now touches one in every five adults. This is not merely a statistic but a clarion call for urgent action, as liver-related deaths in the nation have surged to 268,580 annually, which is 3.17% of all deaths and accounts for a staggering 18.3% of the global liver-related mortality rate.

Since 1980, India has seen a continual increase in liver disease deaths a trend in sharp contrast to countries like China, where the numbers have stabilized or even declined. This rise is inextricably linked to the impact of liver diseases on the Indian economy and healthcare resources, a situation further complicated by the diverse etiology of liver conditions. While common causes such as hepatitis viruses, alcohol consumption, and non-alcoholic fatty liver disease mirror those in the West, India also contends with tropical diseases that significantly affect liver health and contribute to the burden of disease, disability, and death.

Hepatitis B and C, pervasive and insidious in their spread, are a major concern due to their potential for chronic progression, cirrhosis, and liver cancer. With a hepatitis B surface antigen prevalence of 3-4.2% and hepatitis C antibody prevalence around 0.5%, India is home to over 40 million HBV carriers and between 4.7 to 10 million HCV carriers. Despite the grim scenario, it is heartening to note that interventions over the years -

such as screening blood products, safe injection practices, and the integration of the hepatitis B vaccine into the Universal Immunization Program have made significant inroads in combatting these infections. However, the road to the elimination of such infections, the 'zero-risk' goal, is still a distant reality.

The most enigmatic of pathogens, hepatitis E, is responsible for millions of infections annually in India, posing severe risks, especially to pregnant women. The country faces the challenge of HEV waterborne epidemics, which necessitate a robust public health response, focusing on clean drinking water and safe sewage disposal. With efficacious HEV vaccines on the horizon, there is a glimmer of hope for controlling this widespread infection.

Alcohol-related liver diseases and non-alcoholic fatty liver disease (NAFLD) further compound the liver health crisis. The former is directly implicated in a significant number of liver cirrhosis and cancer cases, while the latter has emerged as the most prevalent liver disease, affecting nearly two billion people globally. NAFLD stands at the intersection of liver health and other leading causes of morbidity, emphasizing the pressing need for lifestyle interventions and the exploration of therapeutic options.

Amidst this landscape of despair, liver transplantation presents a beacon of hope, with India performing around 1800 transplants annually. However, the journey is fraught with challenges, including the cost, donor availability, and a striking disparity between the need and the number of transplants performed. It underscores the imperative for innovative strategies tailored to India's diverse cultural and economic milieu [1].

In addressing this dire need, our study applies an array of advanced machine learning techniques to the Indian Liver Patient Dataset, aiming to create a robust predictive model. The dataset underwent rigorous preprocessing to ensure quality and relevance, including the removal of duplicate entries and the handling of missing values. Our preprocessing pipeline also tackled class imbalance, a common challenge in medical datasets, using Random Over-Sampling to create a balanced representation of outcomes, a crucial step to ensure the model's generalizability to real-world scenarios.

The cornerstone of our methodology is feature selection, which is essential to model accuracy and interpretability. Through correlation analysis, we identified and retained features with significant relationships to the target variable, thereby enhancing the predictive power of our model while maintaining computational efficiency. A feature importance analysis via a RandomForestClassifier provided further insight, guiding us towards the most relevant features for liver disease prediction.

Our model ensemble incorporated multiple algorithms, including neural networks, boosting classifiers, and decision trees. This ensemble approach allows us to leverage the strengths of diverse algorithms, mitigating the weaknesses inherent to any single model. Among these, we employed an innovative StackingClassifier that combined RandomForest, GradientBoosting, and ExtraTrees classifiers with a LogisticRegression meta-model. This stacked model was trained and validated through stratified k-fold cross-validation, ensuring that our results were robust against overfitting and reflective of the model's performance across various subsets of the data.

The metrics we used to evaluate our model accuracy, precision, recall, F1 score, and ROC-AUC paint a comprehensive picture of its diagnostic capabilities. With an average accuracy of 91.04%, our model demonstrates high reliability. Precision and recall rates of 89.32% and 93.46%, respectively, indicate a strong balance between the model's sensitivity and specificity, while the F1 score of 91.29% reflects the harmonic mean of precision and recall. The ROC-AUC score, also at 91.04%, assures us of the model's excellent discriminative ability between the classes representing the presence or absence of liver disease.

The deployment of such a model in clinical settings could revolutionize the early detection and treatment of liver diseases, providing healthcare professionals with a powerful tool to assess risk and make informed decisions. The subsequent sections will delve deeper into the dataset description, preprocessing steps, and an in-depth discussion of the machine learning models.

Finally, we will discuss the results in the context of existing diagnostic methods and explore the implications of our findings for the future of liver disease diagnosis and treatment.

In the following sections, we will outline the dataset, detail the preprocessing methodologies employed, and elucidate the machine learning framework adopted. Our results will be presented and discussed in the context of their potential impact on diagnostic practices. The paper will conclude with a contemplation of our findings and their implications for future research avenues.

# 2 Dataset Description

Liver disease, with its multifaceted etiology encompassing factors such as excessive alcohol use, exposure to hepatotoxic substances, and various infections, presents a formidable global health burden. Within this spectrum, the early detection of liver pathology stands as a pivotal determinant of patient outcomes. Notably, the disparity in disease manifestation between genders underscores the exigency for diagnostic models that offer equitable sensitivity across populations.

Our analysis leverages the ILPD (Indian Liver Patient Dataset), a rich compilation of patient records aimed at facilitating the development of machine learning models for the early detection of liver disease. Sourced from the UCI Machine Learning Repository, this dataset is instrumental in addressing not only the disease's clinical aspects but also the health equity challenges it poses. It has been pivotal in studies examining disparities in liver disease prediction between male and female patients, highlighting the need for sex-stratified analysis in healthcare algorithms.

The dataset is characterized by its multivariate nature, comprising 583 instances and 10 features designed for classification tasks. These features encompass a range of both integer and real data types, reflecting various demographic and biochemical markers. Each feature has been meticulously curated to represent a significant aspect of liver health, providing a window into the metabolic state of the patient.

The dataset includes the following variables:

* Age (Age of the patient)
* Gender (Gender of the patient)
* Total Bilirubin (TB)
* Direct Bilirubin (DB)
* Alkaline Phosphotase (Alkphos)
* Alamine Aminotransferase (Sgpt)
* Aspartate Aminotransferase (Sgot)
* Total Proteins (TP)
* Albumin (ALB)
* Albumin and Globulin Ratio (A/G Ratio)
* Class Label (Selector)

Each patient record is a vector of these variables, annotated with a class label indicating the presence (disease) or absence (no disease) of liver pathology.

Prior to analysis, the dataset was meticulously preprocessed. Patients above the age of 89 were recorded as '90' to maintain anonymity, ensuring compliance with privacy considerations. A thorough examination revealed the dataset to be complete with no missing values, though 13 duplicate records were identified and subsequently addressed to prevent any skew in the analysis. The original dataset displayed a class imbalance, favoring liver disease instances; this was rectified using SMOTE to ensure a balanced dataset for our machine learning tasks.

This dataset has been a cornerstone in comparative studies between patients from different geographical regions, namely the USA and India, and has served as a foundational element in investigating gender-based biases in healthcare algorithms. Such analyses are crucial, as they inform the development of algorithms that aspire to equitable healthcare outcomes.

The dataset from the UCI repository, along with the insights garnered from it, has been instrumental in shaping our approach to this study. It provides a concrete foundation upon which our predictive models are built and evaluated, in the pursuit of advancing non-invasive, accurate, and equitable diagnostics for liver disease [2].

# 3 Methodology

## 3.1 Data Preprocessing and Feature Importance

As part of the data processing, the problem of class imbalance in the data set, a widespread and significant phenomenon in medical data analysis, was first addressed.

To ensure a balanced representation of different classes and to eliminate potential biases of the models in favor of the majority class, the Synthetic Minority Oversampling Technique (SMOTE) was used. This technique generates synthetic data points for the underrepresented class by identifying the characteristics of the minority class and creating new, similar data points. This results in a more balanced and representative data set, which forms the basis for more robust and fairer modeling.

After correcting for class imbalance, a feature importance analysis was performed to determine the importance of individual features for the prediction of liver disease. Random Forest classifier algorithms were used to rank the features according to their importance. This ranking method is based on the investigation of the influence of changes in individual features on the result of the classification. The findings on feature importance provide insightful information about the data structure and clarify which variables have the greatest influence on the prognosis of liver diseases. This enables the development of models that not only make precise predictions but can also emphasize the relevant features and neglect less important ones.

## 3.2 Model Selection

We selected a suite of machine learning models known for their effectiveness in classification tasks. These models included:

1. **Random Forest Classifier**: Recognized for its performance and ease of use, it was chosen for its ensemble learning capabilities, which help in reducing variance and improving generalizability.
2. **Gradient Boosting Classifier**: Valued for its predictive power, it builds an additive model in a forward stage-wise fashion, allowing for the optimization of arbitrary differentiable loss functions.
3. **Extra Trees Classifier**: An ensemble learning method similar to the random forest, known for fitting a number of randomized decision trees on various sub-samples of the dataset and using averaging to improve predictive accuracy and control overfitting.

These models were used as base estimators in a stacking ensemble.

## 3. 3 Stacking Ensemble Model

In our endeavor to develop a robust prediction model, we decided to construct a stacking classifier. This advanced approach combines the predictions of multiple base estimators to overcome the weaknesses of individual models and increase the overall strength of the system. The base estimators, selected for their individual performance and diversity in data modeling, were trained extensively on the complete data set. This included a variety of algorithms, from tree-based methods to instance-based models, which were designed to fully capture the heterogeneity of the dataset.

Our final estimator, a carefully calibrated logistic regression model, was trained not only to learn the best combination of predictions from the base models, but also to provide a fine-tuned assessment of the probabilities of class membership. Logistic regression was chosen as the meta-model due to its ability to calibrate the probabilities, which is beneficial when assessing the certainty of a prediction.

Using the stacking classifier allowed us to increase the predictive power by combining the strengths of the base estimators, while logistic regression helped to synthesize this information and produce a unified prediction. Through this multi-level approach, a more complex representation of the solution space was achieved, leading to an improvement in generalizability and robustness to unpredictability in the data.

## 3.4 Cross-Validation and Model Evaluation

To ensure the reliability and stability of our predictive model, we opted for a rigorous validation method: StratifiedKFold cross-validation with 10 splits. This method ensures the preservation of class proportions in all subsets and thus provides a reliable estimate of model performance that is free from bias due to unequal class distribution.

Each convolution in the cross-validation process involved careful training of the model on 90% of the dataset and testing on the remaining 10% to evaluate the performance of the model. This procedure was repeated until each subset of the dataset served as a test set. We emphasized providing a holistic evaluation of the model by using a variety of metrics that reflect different aspects of the model's predictive capabilities:

* **Accuracy**: A general metric that indicates the overall accuracy of the model.
* **Precision**: This metric measures how accurate the model is in predicting liver disease.
* **Recall (Sensitivity)**: It evaluates the model's ability to identify all relevant cases of liver disease.
* **F1 Score**: A combination of precision and recall that provides a balanced measure of the model's accuracy.
* **ROC-AUC Score**: A metric that indicates how well the model can distinguish the classes.

By calculating these metrics for each convolution and aggregating the results, we obtained an average performance score that allows for a comprehensive evaluation of the model. This comprehensive evaluation structure ensures that we can assess the generalizability of our models beyond the training data and ensure that the final results are both reliable and reproducible.

# 4 Results

The performance of the stacking ensemble model was rigorously evaluated using StratifiedKFold cross-validation with 10 splits to ensure the consistency and reliability of the predictive model across the dataset. The model demonstrated robust performance across various metrics, which are critical indicators of its effectiveness in classifying liver disease.

The accuracy score, a measure of the model's overall correctness, yielded an average of 91.04%. This suggests that the stacking model correctly identified liver disease presence or absence in approximately 91% of cases across the validation folds.

Precision, which quantifies the model's exactness, was on average 89.32%. This high precision rate indicates that when the model predicted liver disease, it was correct approximately 89% of the time.

Recall, also known as sensitivity, measures the model's ability to identify all relevant cases of liver disease. The model achieved an average recall of 93.46%, suggesting it was able to identify 93% of liver disease cases within the dataset.

The F1 score, which balances precision and recall in a single metric, was 91.29% on average. This high F1 score indicates that the model has a harmonious balance of precision and recall.

The ROC-AUC score, which assesses the model's ability to discriminate between classes, was equivalent to the accuracy at 91.04%. An ROC-AUC score close to 1 indicates a high level of model discrimination capability.

These metrics collectively suggest that the stacking ensemble model, composed of RandomForestClassifier, GradientBoostingClassifier, ExtraTreesClassifier as base estimators, and LogisticRegression as the final estimator, is highly effective in predicting liver disease from the given dataset.

# 5 Discussion

# 6 Conclusion and Future Directions

# Acknowledgements

This data was sourced from the UCI Machine Learning Repository, courtesy of Lichman, M. (2013), University of California, Irvine, School of Information and Computer Science.

# References

1. Khuroo MS (2023) "Liver Diseases in India: Hope and Despair," Greater Kashmir. [Online]. Available: https://www.greaterkashmir.com/todays-paper/op-ed/liver-diseases-in-india-hope-and-despair.

[Accessed: 15-Dec-2023].

1. Ramana, B., & Venkateswarlu, N. (2012). ILPD (Indian Liver Patient Dataset). UCI Machine Learning Repository. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset). [Accessed: 15-Dec-2023].