Liver Disease Detection

Paul Weiss, Haranadh Reddy Ravi, Bhavani Chalamalla, Dallas Grandy
CS5831 Final Project, Michigan Technological University
psweiss@mtu.edu, hravi1@mtu.edu, bchalama@mtu.edu, dmgrandy@mtu.edu

Abstract

Cirrhosis, a serious condition characterized by the scarring of the liver and impaired liver function, affects a significant portion of the U.S. population. As of recent data, it's estimated that approximately 20-30% of the population in developed countries is affected by chronic liver disease in varying levels (Pournik, 2014), underscoring its relevance in medical research and healthcare. This project aims to develop a reliable machine-learning framework for diagnosing liver cirrhosis, a critical health condition marked by liver scarring and impaired function. Leveraging a dataset published on Kaggle containing medical data of 583 patients (Indian, n.d.) to explore the effectiveness of K-Nearest Neighbor (KNN) and Logistic Regression models in classifying patients with liver cirrhosis. This dataset includes key indicators like age, gender, liver function tests (Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase, and more), and protein levels, crucial for understanding and predicting liver health. Initial preprocessing involved adjustments, primarily converting binary data fields to numeric formats. The approach for this study involved tuning hyperparameters for KNN and Logistic Regression models. The preliminary results indicated an average accuracy of approximately 65% for the KNN model and a slightly higher accuracy for Logistic Regression. However, both models struggled to balance false positives and negatives, necessitating further refinement. These initial outcomes underscore the complexities of medical data analysis and the need for a more sophisticated model tuning and balancing techniques. Overall, this project sought to explore how a machine learning model can assist with early detection of liver disease, where accurate and early detection can significantly impact patient outcomes.

Introduction[1]

The premise of this project centers on the early detection of liver disorders, a crucial area in healthcare that directly impacts patient outcomes and treatment strategies. Liver diseases, such as cirrhosis or hepatitis, are often diagnosed through invasive procedures like biopsies, which pose risks and discomfort to patients (Kumar, 2018). This highlights the importance of developing non-invasive diagnostic methods that can efficiently and accurately detect liver disorders at an early stage. Early detection is critical for timely treatment and plays a significant role in patient prognosis, allowing for interventions that could reverse damage or slow disease progression (Bhupathi, 2022).

This project addresses this challenge by applying machine learning techniques to clinical data for the predictive diagnosis of liver diseases. K-nearest neighbor (KNN) and Logistic Regression are the primary models this project explores. Model choice was driven by their proven effectiveness to handle complex and varied data sets. The KNN model's focus was on experimenting with different numbers of neighbors to optimize classification accuracy, while the Logistic Regression model concentrated on adjusting regularization strength to improve predictive performance. Initial outcomes indicate a moderate accuracy level, with the KNN model achieving about 65% and the Logistic Regression model performing slightly better. These results, while promising, also highlight the challenges in medical data analysis, particularly in balancing false positives and negatives. The findings lay a foundation for future research and development in this crucial area, aiming to ultimately improve patient outcomes and reduce the strain on healthcare systems.

Related Work

In addressing the complex challenge of diagnosing liver cirrhosis, various machine-learning techniques present distinct advantages and disadvantages (Volovici, 2022). While deep learning methods like Convolutional Neural

[1]Copyright © 2024, CS 5831 Final Project, Michigan Technological University. All rights reserved.

Networks (CNNs) and Recurrent Neural Networks (RNNs) are highly effective in pattern recognition within complex datasets, their need for extensive data and computational power, coupled with their "black box" nature, pose significant limitations, particularly in interpretability which is crucial in medical diagnostics.

Dutta et al. (2022) tested many models including Artificial Neural Networks, KNN, Decision Trees, and Logistic Regression to enable early detection of liver disease. Using accuracy, precision, recall, and F1 Score as the primary measures the study applied to these various models with and without the addition of Linear Discriminant Analysis (LDA). LDA was applied as a dimensionality reduction technique given the continuous variables that make up the features. This study utilized a very similar data set, differing only in size, with 30k samples. The features utilized are a replica of those available in the Indian Liver Patient Data Set. The larger data set required additional cleaning with null values and some duplication to complicate some pre-processing requirements. While implementing LDA made many of the models perform better, the best model was a Decision Tree without LDA. This model reached an accuracy of 99.9%.

A study by Bhupathi et al. (2022) expanded beyond the utilization of the LDA technique to improve model performance. This newer study did not consider a Decision Tree and reviewed an SVM, Naive Bayes, Knn, and Classification and Regression Trees. The Knn model reached an accuracy of 91.7%. The Liver Disease Prediction (LDP) model developed was based on the Sample-Explore-Modify-Model-Assess(SEMMA) lifecycle to explore, build, and assess. Small tweaks were made to add Data Pre-Processing between Modify and Model as well as a Results section at the conclusion. This study uses the Indian Liver Patient Data and applies a method to create synthetic records to balance the class distribution.

Decision Trees and Random Forests, known for their robustness against overfitting and ability to provide feature importance scores, can become overly complex and less interpretable with a large number of trees. Furthermore, they may not handle imbalanced datasets efficiently, a common issue in medical data (Dwaraka et al., 2023).

Support Vector Machines (SVMs), while effective for binary classification and smaller datasets, face challenges with larger datasets and multi-class classification problems (Mills, 2018). The complexity of selecting appropriate kernels and parameters further complicates their use in diverse medical datasets.

Data

Liver disease data was collected from a previous liver cirrhosis study published in Kaggle (Indian, n.d.). The file contains 583 samples with 10 predictive features available from age, gender, and common blood panels. Preliminary exploration reveals that 71% of the patients in the study were diagnosed with liver cirrhosis and 29% were not.

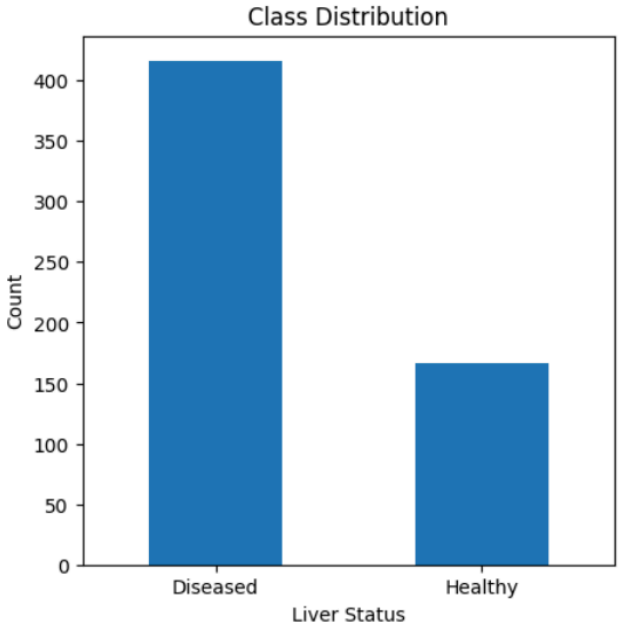

Figure 1: Bar plot representing the class distribution

Data Columns

Features available in the data are (Cleveland, 2024):

- Age - Patient's Age
- Gender - Patient's Gender (Binary)
- Total Bilirubin (TB) - Total amount of bilirubin in the blood, a yellow pigment produced by the breakdown of red blood cells.
- Direct Bilirubin (DB) - The direct fraction of bilirubin, specifically associated with liver function.
- Alkaline Phosphatase (Alkphos) - An enzyme associated with the biliary system; elevated levels indicate potential liver or bone issues.
- Serum Glutamic Pyruvic Transaminase (Sgpt) - An enzyme predictive of liver health; elevated levels suggest liver damage.
- Serum Glutamic Oxaloacetic Transaminase (Sgot) - An enzyme predictive of liver health; elevated values indicated potential liver disease
- Total Proteins (TP) - Total amount of proteins in the blood, including albumin and globulins.
- Albumin (ALB) - A protein generated by the liver; that impacts blood volume and pressure.

- A/G Ratio - The ratio of albumin to globulins, indicates liver and kidney function

Classifier:

- Selector - Identifies if a patient is diagnosed with Liver Cirrhosis through a biopsy

Data Preprocessing

SMOTE was not utilized to generate synthetic data to account for the slight imbalance. However, 4 records had NaN values in the A/G Ratio field. These values were imputed with median values. The median, instead of the mean, was additionally calculated based on the Selector to attempt to avoid influence from some of the extreme outliers found in this dataset. To better utilize binary classification in future model builds, the Selector was mapped to 0 and 1 values. The final pre-processing steps involved mapping the Gender to binary values of 0 and 1 for Male and Female respectively.

The number of features and related enzymes being measured implied that there would be varying amounts of correlation. A correlation matrix was used to view the extent of correlation between features. Total Bilirubin and Direct Bilirubin are factors of the same measure. Various model tests will be executed with one or the other of these fields dropped in pre-processing. A similar level of correlation is observed between Albumin and the A/G Ratio (Albumin/Globulin Ratio). As a result of needing to impute values for NaN, the A/G Ratio feature will be dropped from the dataset. The third area of strong correlation is between aspartate aminotransferase (Sgot) and alanine aminotransferase (Sgpt). Due to the significance of these features and because their values are not based on a ratio or calculation of the other, both features will be retained in the modeling (HexaHealth, 2024).

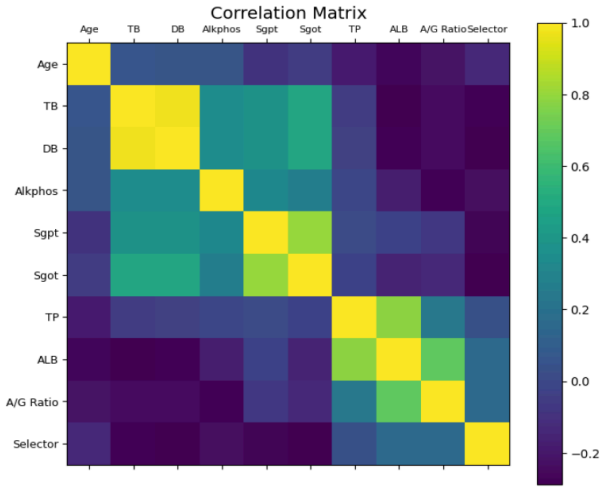

Figure 2: Plot showing the correlation between the variables

A significant concern identified in the data set was the large number of extreme outliers. The features measuring different enzymes, proteins, and pigments can also occur in other parts of the musculoskeletal portions of the body (Woreta, 2014). Therefore, some patients exhibited much higher results in the blood panel than others. While scaling and other techniques will be examined in the Methods, these box plots generated post-scaling highlight the extent of outliers within the data set.

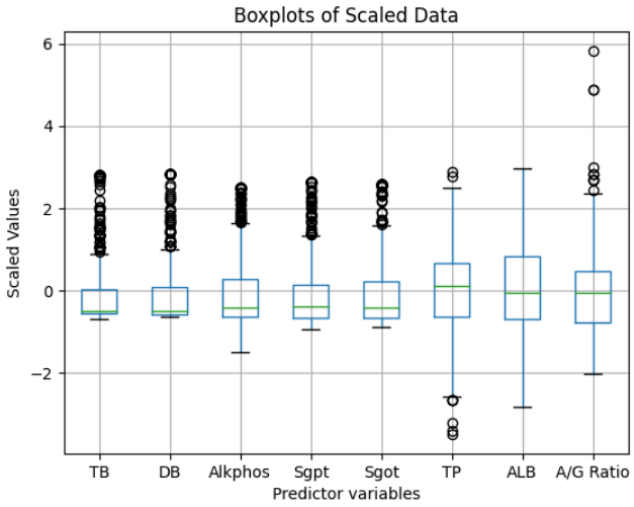

Figure 3: Distribution Across Features

Methods

Two primary approaches utilized are K-Nearest Neighbor (KNN) and Logistic Regression. Instead of dimensionality reduction techniques, this project utilizes winsoring to attempt to reduce the impact of massive outliers identified in various enzyme measures. KNN's simplicity and lack of data distribution assumptions make it suitable for smaller, medical datasets, offering transparency often lacking in more complex models. Logistic Regression's efficiency in binary classification and probability estimation provides a straightforward and interpretable method, essential for medical applications where understanding the model's reasoning is as important as its predictive accuracy. This balance is vital in the medical field, ensuring that the diagnostic tools are not only effective but also transparent and easily understandable, key factors in medical decision-making and patient care. Both models will be used to classify if the patient has cirrhosis of the liver or not.

For the K-Nearest Neighbor model the method in which it predicts this classifier is by using previous data closest to the new data point given. It does this by finding the distance between each feature point and the k closest points are used to classify the prediction. This project experimented with multiple neighbor lengths and

concluded that more hyperparameters are needed to tune the accuracy of the model efficiently.

Logistic regression is a command machine learning technique that performs exceptionally well with a binary target and independent features. A correlation matrix will be compiled to show feature independence or the lack thereof.

In many situations, medical data is noisy and outlier-prone. Winsorization is a technique to mitigate the effects of these outliers (Sharma, 2021). Winsorizing stems a feature's range at the lower and upper quartile(s). Both a Lower and Upper Interquartile Range is chosen and by applying the winsoring technique values outside the defined range are replaced with the value at the appropriate range's extent.

As an additional factor to adjust for overweighting features with large outliers we also applied the StandardScaler() function. As seen in Figure 3 above, standardization helped bring all features into a similar range while maintaining the within feature data distribution.

In addition to tuning model hyperparameters and applying preprocessing techniques, the project also considered bias. Specifically testing against potential Gender Bias. For both the KNN and Logistic Regression models data was filtered by gender and processed. No additional tuning was conducted to enable comparisons. Specifically for this analysis, only the Accuracy will be observed.

Like the nearest neighbor classifier, additional fine-tuning on the hyperparameters is necessary. After some exploratory graphing of different regularization weights, it seems to have the same effect as neighbor length, but within a slightly smaller range of accuracy than the neighbor model. Future testing and tuning of other hyperparameters is recommended as well as implementing pipelines and ensemble methods. Finally adding a grid-search method to find the appropriate parameter settings. The logistic regression model will take each feature and assign its weight based on the effect it has on model training. These weights are then used to classify the percentage chance that the observed data point would be considered 'positive', if this percentage is above 50%, then it classifies the data as positive and vice versa.

Experiment and Results

Following the pre-processing steps to prepare the data sets, the data was split into training and testing sets with 80/20 stratified splitting. This study concentrated on the effects of winsoring and checking models for gender bias with minimal hyperparameter tuning. For the KNN model, only the number of neighbors was adjusted and for logistic regression, only the regularization strength was tuned.

K-Nearest Neighbor Model

The first method used was a simple K-Nearest Neighbor model with various neighbor lengths as a preliminary test. Below in Figure 4, observe the results of the KNN model's accuracy as the number of neighbors changes.

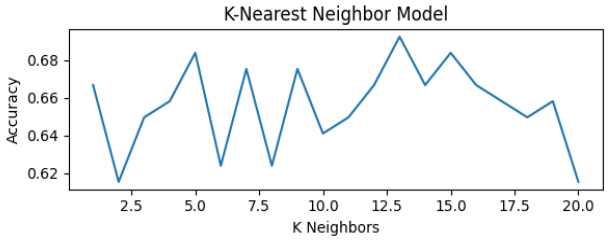

Figure 4: KNN Accuracy by K Values

An average Accuracy of around ~65% is not a very accurate model and will need further exploration and tuning of hyperparameters to achieve a better score. This may also be because NaN values were imputed with median values of each column as some variables even when standardized have very generous outliers within the data. Testing still needs to be done on the effects of different data sanitation/preprocessing.

Below in Table 1 is a confusion matrix of the KNN model with a neighbor length of 10. As observed the KNN model performs with an accuracy of 64.1% but a precision score of 73.2%. It also had a false positive rate of 18.8% and a false negative rate of 17.1%. In terms of using this model in a real-world application, various tuning is recommended for it to be ready for use.

Table 1: K-Nearest Neighbor Confusion Matrix

		True	**False**
Label	**Diseased**	60 (TP)	20 (FN)
	Healthy	22 (FP)	15 (TN)
		Predicted Label	

Logistic Regression Model

The logistic regression model built is in a similar process of testing as the nearest neighbor model. Initial model testing involved tuning the regularization strength parameter to get a feel for how the Logistic Regression model operates. As seen in Figure 5, the accuracy of the model is ~68%, which is a slight improvement from the nearest neighbor model.

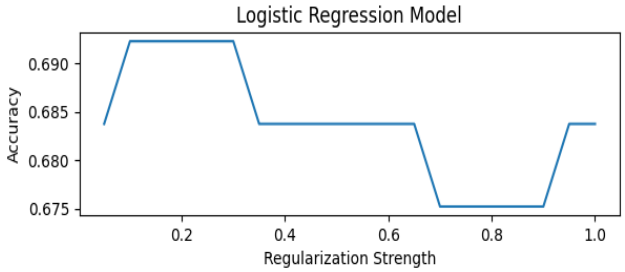

Figure 5: Logistic Regression Accuracy by C Values

Below in Table 2, the confusion matrix reveals the results from the logistic regression model. The model was made with a regularization strength of 0.5. The model had an accuracy of 79% and a precision score of 68.7%, lower than that of the KNN model. Its false positive rate was 30.8% and its false negative rate was 0.8%, so this model was better at predicting liver disease with great precision, but would also falsely classify the majority of false cases as diseased when the patient was healthy. This may be due to an activation parameter being too low in which most of the data is being falsely flagged as positive. Another reason for this may be due to the higher amount of disease data being supplied than healthy data. These problems will be resolved by recommendations for the next wave of testing.

Table 2: Logistic Regression Confusion Matrix

		True	**False**
Label	**Diseased**	79 (TP)	1 (FN)
	Healthy	36 (FP)	1 (TN)
		Predicted Label	

Gender Bias in the Model

Testing the model against only male records revealed greater accuracy than that of the whole sample population. The KNN model produced 78% accuracy and the Logistic Regression model yielded 73% accuracy.

Table 3: Male KNN Regression Confusion Matrix

		True	**False**
Label	**Diseased**	63 (TP)	2 (FN)
	Healthy	17 (FP)	7 (TN)
		Predicted Label	

Table 4: Male Logistic Regression Confusion Matrix

		True	**False**
Label	**Diseased**	65 (TP)	0 (FN)
	Healthy	24 (FP)	0 (TN)
		Predicted Label	

For female samples, results were significantly lower. The KNN model produced 55% accuracy and the Logistic Regression model yielded 65% accuracy.

Table 5: Female KNN Regression Confusion Matrix

		True	**False**
Label	**Diseased**	13 (TP)	6 (FN)
	Healthy	7 (FP)	3 (TN)
		Predicted Label	

Table 6: Female Logistic Regression Confusion Matrix

		True	**False**
Label	**Diseased**	19 (TP)	0 (FN)
	Healthy	10 (FP)	0 (TN)
		Predicted Label	

The Linear Regression model continued to favor a positive diagnosis across both genders which resulted in many false positive findings. However, given the KNN model results, there clearly a significant amount of gender bias within the machine learning process for current models. These differences could be explained by varying test levels and sample sizes. It may also be indicative of why decision trees in previous studies performed with superior accuracy over KNN and Linear Regression models (Dutta et al., 2022). Potentially gender is a key feature to split in the tree building process.

Conclusions

No significant findings were identified after winsorizing fields with significant outliers. Preliminary results indicated an average accuracy of approximately 65% for the KNN model and slightly higher accuracy for Logistic Regression. However, both models exhibited challenges with balancing false positives and negatives, necessitating further refinement. These initial outcomes underscore the complexities of medical data analysis and the need for a more sophisticated model tuning and balancing techniques. The project, thus, sets the stage for advanced machine learning applications in healthcare, particularly in the domain of liver disease diagnosis, where accurate and early detection can significantly impact patient outcomes.

In addition to accuracy, Specificity is an important observation. The primary purpose of this is to account for possible measures that indicate potential False Negative results. This Type-2 error is the worst-case scenario from any chosen model as it may prevent a patient from receiving much-needed care or additional testing to enable early treatment. While the Logistics Regression model over-predicted False Positive results, the specificity was much lower (0.8% vs. 17.1%).

To address the complexities presented in healthcare problems such as this, additional models would be required to achieve improved results. The next steps for future projects would include potentially more complex models or using ensemble methods to capitalize on these simpler predictive tools. Ultimately, the medical field strives to ensure patients receive the treatment they need as soon as possible while limiting the overuse of resources, spending, and high-risk or invasive procedures.

References

1. Bhupathi, Deepika & Tan, Christine Nya-Ling & Sremath Tirumala, Sreenivas & Ray, Sayan. (2022). Liver disease detection using machine learning techniques.

2. Cleveland Clinic. “Liver Function Tests.” Cleveland Clinic, https://my.clevelandclinic.org/health/diagnostics/17662-liver-function-tests. Accessed 05 APR 2024.

3. Dutta, K., Chandra, S., & Gourisaria, M. K. (2022). Early-stage detection of liver disease through machine learning algorithms. Lecture Notes in Networks and Systems, 318, 155–166. https://doi.org/10.1007/978-981-16-5689-7_14

4. Dwaraka Srihith, I., Vijaya Lakshmi, P., David Donald, A., Aditya Sai Srinivas, T., & Thippanna, G. (2023). A forest of possibilities: Decision trees and beyond. Journal of Advancement in Parallel Computing, 6(3), 29–37. https://doi.org/10.5281/zenodo.8372196

5. HexaHealth HealthCare Team, Reviewed by: Dr. Aman Priya Khanna. https://www.hexahealth.com/blog/difference-between-sgpt-and-sgot, Difference between SGOT and SGPT, Published: Jan. 2024, Accessed: 14 APR 2024

6. Indian Liver Patient Records. (n.d.). Retrieved Feb. 14, 2024, from https://kaggle.com/uciml/indian-liver-patient-records

7. Kumar, S. and Katyal, S., "Effective Analysis and Diagnosis of Liver Disorder by Data Mining," 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2018, pp. 1047-1051, doi: 10.1109/ICIRCA.2018.8596817.

8. Mills, P. (2018). Solving for multi-class: A survey and synthesis. Journal of Advancement in Parallel Computing, 6(3), 29–37. https://doi.org/10.5281/zenodo.8372196

9. Pournik O, Dorri S, Zabolinezhad H, Alavian SM, Eslami S. A diagnostic model for cirrhosis in patients with non-alcoholic fatty liver disease: an artificial neural network approach. Med J Islam Repub Iran. 2014 Oct 21;28:116. PMID: 25678995; PMCID: PMC4313459.

10. Sharma, S.; Chatterjee, S. Winsorization for Robust Bayesian Neural Networks. Entropy 2021, 23, 1546. https://doi.org/10.3390/e23111546

11. Volovici, V., Syn, N.L., Ercole, A. et al. Steps to avoid overuse and misuse of machine learning in clinical research. Nat Med 28, 1996–1999 (2022). https://doi.org/10.1038/s41591-022-01961-6

12. Woreta TA, Alqahtani SA. Evaluation of abnormal liver tests. Med Clin North Am. 2014 Jan;98(1):1-16. doi: 10.1016/j.mcna.2013.09.005. Epub 2013 Oct 28. PMID: 24266911.