Group 2 Final Project: Liver Disease Prediction

Selina Meng; Tianchuzi Qin; Tina Zhang; Yuyang Chen

MSDS 422: Practical Machine Learning

02/09/2025

1.0 Abstract Draft

Liver disease remains a major global health problem that requires innovative solutions for early detection and management. We developed an automated liver disease prediction process based on machine learning and deep learning principles using the Kaggle dataset “Liver Disease Patient Dataset 30K Training Data”, which has two subsets. Train consists of 20K training data in 10 columns, including key metrics such as total bilirubin, albumin, and demographic statistics. Test, on the other hand, contains only about 1k of data.

The data preparation phase involved addressing missing values, encoding categorical variables, scaling numerical data, and extracting key features. These steps were essential to ensure the stability, accuracy, and reliability of the models. Exploratory data analysis (EDA) was conducted to uncover important patterns and trends, including an in-depth understanding of the relationship between key biochemical markers and the presence of disease, to guide the model-building process.

After we followed data preparation and analysis, the study evaluated four models: logistic regression, random forest, gradient boosting, and a simple deep-learning model. We currently prefer gradient boosting because it is most likely to achieve the best performance, even if the data is unbalanced to provide powerful classification capabilities. In addition, our deployment strategy focuses on automated data extraction, model retraining, and prediction using cloud-based tools such as Amazon SageMaker and TensorFlow.

Our projects emphasize simplicity and practicality. Recommendations include integrating models into clinical workflows to assist with real-time predictions and exploring additional datasets to improve model performance. By leveraging machine learning, our model may show

how a data-driven approach can improve healthcare outcomes and provide a valuable tool for addressing liver disease diagnosis.

2.0 Problem Statement/Research Objectives

This study aims to develop ML-based predictive models for liver disease from liver disease datasets to improve predictive accuracy, model interpretability, and clinical applicability. Therefore, exploratory data analysis will first be performed to reveal key demographic, clinical, and biochemical factors associated with liver disease, analyze feature distributions, examine potential data imbalances, and assess correlations between variables. In addition, data preprocessing, such as dealing with missing values and outliers and handling categorical variables (e.g., feature selection, scaling, and transformation), will be performed to improve model performance. This study will further develop and compare at least four models, including models such as logistic regression and decision trees. We will use cross-validation, mean square error, and root mean square error. We will perform correlation analyses, such as feature importance analysis, to identify the variables that have the greatest impact on the prediction of liver disease and consider clinical characteristics associated with the predicted outcomes to improve model transparency. Finally, we will evaluate the generalizability and scalability of the model using different subsets of data and further improve the model by combining different factors. Finally, the model trained on top of the training dataset can be applied to our test dataset to predict the occurrence of liver disease. After all this work is done, an efficient, scalable, and interpretable model for early detection of liver disease will be created to provide a means for public health and healthcare.

3.0 Annotated Bibliography

1. Lu, Chien-Hung, Weu Wang, Yu-Chuan Jack Li, I-Wei Chang, Chi-Long Chen, Chien-Wei Su, Chun-Chao Chang, and Wei-Yu Kao. “Machine Learning Models for Predicting Significant Liver Fibrosis in Patients with Severe Obesity and Nonalcoholic Fatty Liver Disease.” *Obesity Surgery* 34, no. 12 (2024): 4393–4404. https://doi.org/10.1007/s11695-024-07548-z.

The objective of this review is to outline the performance of various ML models in predicting significant liver fibrosis among severely obese NAFLD patients. Among a total of 194 patients undergoing bariatric surgery, five machine learning algorithms were studied: support vector machine (SVM), random forest (RF), k-nearest neighbor (KNN), XGBoost, and logistic regression alone and combined into an ensemble model. The performance for the ensemble model was the highest: AUROC was 0.83, sensitivity 73.1%, and accuracy 83.1%. Key predictors included liver stiffness measurements (LSM), AST, ALT, and ultrasonographic fibrosis scores. This study highlights the challenges of liver fibrosis diagnosis through traditional methods such as biopsy, therefore requiring the use of non-invasive methods. Using ML models, this research illustrates a new way of identifying high-risk patients, especially bariatric surgical candidates. The multiple predictive variables included will help comprehensively assess the advancement of fibrosis and fill the lacuna of the conventional diagnostic framework.

The strengths are the prospective design and comprehensive inclusion of various ML algorithms that allowed a good assessment of predictive performance. Limitations include a small sample size and restriction to severely obese patients, which is a significant disadvantage for generalizability. That the study relies on invasive validation methodologies, such as liver biopsy, points to the still-pressing challenge presented by the development of completely non-invasive diagnostic tools.

This review article critically discusses the role of ML in enhancing the diagnostic accuracy of liver fibrosis diagnosis, especially in specific populations. Its ensemble modeling approach agrees with our interests in the investigation of advanced algorithms for healthcare applications. These findings emphasize the incorporation of clinical and imaging data into predictive models as a way to higher value in diagnostic precision and accessibility. This study will, therefore, inform our exploration of the role of ML in addressing metabolic health challenges, especially in high-risk groups.

2. Liu, Yuan-Xing, Xi Liu, Chao Cen, Xin Li, Ji-Min Liu, Zhao-Yan Ming, Song-Feng Yu, et al. "Comparison and Development of Advanced Machine Learning Tools to Predict Nonalcoholic Fatty Liver Disease: An Extended Study." *Hepatobiliary & Pancreatic Diseases International* 20, no. 5 (2021): 409–15. https://doi.org/10.1016/j.hbpd.2021.08.004.

This study investigates advanced machine learning methods for predicting Nonalcoholic Fatty Liver Disease, which affects a substantial part of the population in the world. With a sample dataset of over 15,000 Chinese adults, seven different predictive models were developed: XGBoost, support vector machines, and logistic regression, among others. The XGBoost model performed best with an AUROC of 0.873, AUPRC of 0.810, and an accuracy of 79.5%. Body mass index was ranked as the most important predictor, followed by fasting blood glucose and triglycerides. These results indicate that ML-based models could be effective, non-invasive tools for NAFLD population screening and, thus, for early intervention and disease management.

It gives an extensive comparison of the performance of each model regarding sensitivity, specificity, and F1 score for all nine metrics. XGBoost topped the models studied, demonstrating strong predictive performance and a high ability to classify individuals at high risk. This emphasis on ML then highlights its potential to address several limitations with traditional diagnostic methods that generally rely on invasive techniques like liver biopsy.

This work is methodologically exemplary, as it requires a big, variously biased data set to ensure that its outcomes are not the result of a type-one error. More value is brought into the current review with the inclusion of various ML models and details on different performance metrics. Unfortunately, it covers the study on one ethnic group alone; thus, the general relevance of the findings is less. Furthermore, reliance on ultrasonography as a diagnostic tool, being less precise than liver biopsy or advanced imaging, may introduce inaccuracies in the dataset.

This is fundamental research explaining the role of ML in improving healthcare delivery, specifically for non-invasive diagnostics. The results relate well with our research focus on the application of advanced algorithms in disease prediction and management. Also, with a focus on available predictors such as BMI, this study underlines the chances of ML democratizing healthcare solutions by making diagnostic tools available more broadly than ever. This paper presents a key reference for understanding how to apply ML in solving global health challenges.

3. Kang, Baeki E, Aron Park, Hyekyung Yang, Yunju Jo, Tae Gyu Oh, Seung Min Jeong, Yosep Ji, et al. “Machine Learning-Derived Gut Microbiome Signature Predicts Fatty Liver Disease in the Presence of Insulin Resistance.” *Scientific Reports* 12, no. 1 (2022): 21842–21842. https://doi.org/10.1038/s41598-022-26102-4.

The following study will review the predictive power of gut microbiome data coupled with ML for the diagnosis of FLD in insulin-resistant individuals. Using 16S rDNA sequencing data on 777 subjects, the authors have developed and optimized a random forest classifier with an AUROC of 0.93 for insulin-resistant subjects. It identified a core set of ten microbial genera as key biomarkers, offering a non-invasive diagnostic alternative to traditional methods like liver biopsy and magnetic resonance imaging.

The research has pointed out the connection of gut microbiota with metabolic disorders and hence indicated that microbiome-based diagnostics may become a game-changer in disease prediction and management. The optimization of the classifier using genetic algorithms underlines the potential of advanced computational techniques to improve predictive accuracy further. The findings indicate significant differences in gut microbial diversity between individuals with and without FLD, especially in insulin-resistant populations.

In the paper, the integration of microbiome data and ML's new approach to FLD diagnostics is pursued. The use of a genetic algorithm in model optimization in this work deepens the analysis to illustrate the potential for personalized medicine. However, the moderate sample size and reliance on one dataset constrain the generalization of the results, while variability in the microbiome across populations remains a challenge for the wide application of such findings.

This article is instrumental in furthering our understanding of microbiome-based diagnostics and their application to metabolic health. The emphasis on insulin resistance as a key factor aligns with our research objectives, underlining the need for targeted and personalized diagnostic tools. This study paves the way for future research toward non-invasive diagnostics through integrating microbiome data with ML. This work will inform our exploration of innovative biomarkers and computational methods in healthcare, particularly for addressing complex metabolic disorders.

4.El–Serag, Hashem B, and K. Lenhard Rudolph. “Hepatocellular Carcinoma: Epidemiology and Molecular Carcinogenesis.” *Gastroenterology (New York, N.Y. 1943)* 132, no. 7 (2007): 2557–76. https://doi.org/10.1053/j.gastro.2007.04.061.

This article provides a comprehensive analysis of the global epidemiology of hepatocellular carcinoma (HCC), the most common and aggressive form of liver cancer, and an in-depth look at its carcinogenic molecular mechanisms. The authors carefully analyzed a range of risk factors for HCC, including chronic infection with hepatitis B and C viruses, excessive alcohol consumption, and non-alcoholic fatty liver disease (NAFLD). In addition, the study highlights the critical role of cirrhosis and chronic liver inflammation as precursors to HCC, thereby providing a nuanced understanding of the progression of the disease.

The molecular pathways involved in liver cancer progression, such as oxidative stress, DNA damage, and alterations in the Wnt/β-catenin signaling pathway, are also explored in depth. These mechanisms are believed to be key factors in tumor initiation and development, providing insights into the intricate interactions between genetic and environmental factors. This discussion is particularly important in advancing our understanding of the biology of liver cancer and guiding the search for potential therapeutic targets.

This study is highly relevant for machine learning applications in liver disease prediction as it highlights the importance of identifying risk factors and early biomarkers and integrating them into predictive models. By elucidating the pathophysiological and epidemiological basis of liver cancer, this paper provides a valuable framework for selecting meaningful features to improve model accuracy. For example, clinical markers such as alpha-fetoprotein (AFP) highlighted in the study represent key predictors that can be incorporated into algorithms to improve early detection and risk stratification.

In conclusion, this paper makes an important contribution to understanding the complexity of liver disease and its development in liver cancer. By emphasizing the importance of early diagnosis and providing a solid foundation for feature selection in predictive models, it provides an important resource for researchers seeking to develop machine learning algorithms that can integrate biochemical and epidemiological data. The insights provided in this paper highlight the potential for an interdisciplinary approach to improve liver disease management outcomes and advance the field of precision medicine.

5. Xiang, Kailai, Baihui Jiang, and Dong Shang. "The Overview of the Deep Learning Integrated into the Medical Imaging of Liver: A Review." Hepatology International 15, no. 4 (2021): 868–80. https://doi.org/10.1007/s12072-021-10229-z.

This review article explores in depth the integration of deep learning techniques in liver medical imaging. The authors discuss the transformative role of DL in imaging processes such as segmentation, lesion detection, and classification. They highlight the advantages of the DL algorithm over traditional methods, including higher accuracy, efficiency, and reduced observer bias in the diagnostic process. The application of key DL models such as convolutional neural networks (CNN) and their variants, including U-Net and VGG-16, in liver imaging tasks is highlighted.

The study systematically reviewed the application of deep learning to liver segmentation and lesion characterization in imaging modalities such as ultrasound (US), computed tomography (CT), and magnetic resonance imaging (MRI). For example, the authors show how transfer learning and fine-tuning can improve model performance on a limited set of healthcare data. They also address challenges to training deep learning models, such as the reliance on high-quality annotated data and the risk of overfitting.

This paper highlights the potential of deep learning in the non-invasive diagnosis of liver diseases, including hepatocellular carcinoma (HCC) and non-alcoholic fatty liver disease (NAFLD). The article focuses on advanced imaging methods such as contrast-enhanced ultrasound (CEUS) and elastography, illustrating their synergies with deep learning in improving diagnostic accuracy. The paper also identifies future research directions, including the integration of multimodal imaging and the development of real-time deep learning applications for endoscopy and intraoperative environments.

This comprehensive review provides valuable insight into the latest advances in deep learning for liver imaging, making it an important resource for researchers and clinicians. It Bridges the gap between computational advances and clinical applications, providing a roadmap for the use of AI in hepatology. This work is particularly important for medical imaging specialists who aim to apply deep learning techniques to liver diagnosis and give us new ideas for our projects.

6. Amin, Ruhul, Rubia Yasmin, Sabba Ruhi, Md Habibur Rahman, and Md Shamim Reza. “Prediction of Chronic Liver Disease Patients Using Integrated Projection Based Statistical Feature Extraction with Machine Learning Algorithms.” Informatics in Medicine Unlocked 36 (2023): 101155-. https://doi.org/10.1016/j.imu.2022.101155.

This study proposes a new approach to predicting liver disease with machine learning. Its classification has been successful on a patient dataset consisting of 583 records provided by UCI-I. The study mainly consists of preprocessing the data, which includes removing outliers and estimating missing values, along with integrated feature extractions through the integration of PCA, FA, and LDA.

Moreover, various classifiers such as Logistic Regression (LR), Random Forest (RF), K-nearest Neighbor (KNN), Support Vector Machine (SVM), Multilayer Perceptron (MLP), and Ensemble Voting Classifier were examined for their effect in this study. Their model states that by combining these features they could reach a classification accuracy of 88.10%, precision of 85.33%, recall of 92.30%, F1 score of 88.68%, and AUC of 88.20%. These results reflect an increase of 0.10%-18.5% over previous studies that anticipated a combination of different feature-extraction techniques for medical diagnosis through ML.

The combination of extractive features has increased the effectiveness of ML models in medical diagnosis, especially in predicting liver diseases. It ensures a balanced and well-distributed dataset using advanced feature extraction techniques such as PCA, FA, and LDA, along with optimal feature selection for better classification accuracy. This methodology heightens predictive models' ability to distinguish between healthy and diseased individuals, making it even more sensitive and specific to indicators of liver disease. Its early detection will provide great relief in terms of providing medical intervention and saving patients from falling prey to the brutality of disease progression, which is one of the main advantages of this approach. It also allows a good feature selection procedure that helps minimize false diagnoses, thus minimizing false positive and false negative rates, which is quite essential for reliable clinical decision-making. It, in turn, would allow healthcare workers to improve their screening, risk hopes, and disease monitoring selective strategies by strengthening predictive models and improving diagnostic accuracy, leading to higher survival rates and reduced healthcare burdens.

Future cure avenues could consider making use of deep-learning models with maximization of inputs through data on biomarkers, thus further bolstering the changeover detection by predictive performance improvement. The paper illustrates the tremendous potential that ML has in changing the entire paradigm in complex medical diagnosis of chronic liver disease while engaging in further refinements toward machine learning to increase diagnostic accuracy.

7. Singal, Amit G, Ashin Mukherjee, B Joseph Elmunzer, Peter D R Higgins, Anna S Lok, Ji Zhu, Jorge A Marrero, and Akbar K Waljee. “Machine Learning Algorithms Outperform Conventional Regression Models in Predicting Development of Hepatocellular Carcinoma.” *The American Journal of Gastroenterology* 108, no. 11 (2013): 1723–30. https://doi.org/10.1038/ajg.2013.332.

In this study, "Machine Learning Algorithms Outperform Conventional Regression Models in Predicting Development of Hepatocellular Carcinoma," the efficacy of various machine learning (ML) techniques in predicting HCC risk in cirrhotic patients is ascertained and compared with traditional regression models. As HCC cases rise primarily due to infections, such as those classified as hepatitis C or due to non-alcoholic fatty liver disease, early intervention could contribute to an increase in patient survival. Herein, the researchers seek to optimize risk stratification with ML algorithms aimed at accurately identifying higher-risk individuals.

Subjects of the study included 442 patients classified as either Child A or B, diagnosed with cirrhosis, recruited from UM between 2004 and 2006, and followed up until the development of HCC, liver transplantation, death, or end of the follow-up period. The cohort was used to develop both regression and ML models, as well as validation against an independent cohort derived from the HALT-C Trial. c-statistics, net reclassification improvement (NRI), and integrated discrimination improvement (IDI) were used to assess the model performance.

The authors state that at the median follow-up duration of 3.5 years, 41 individuals had developed HCC. The ML algorithm outperformed the UM regression model (c-statistic = 0.61) and the HALT-C model (c-statistic = 0.60), by achieving a c-statistic = 0.64 (95% confidence interval: 0.60-0.69). Also, they showed that the ML algorithms boost diagnostic accuracy significantly better as compared to regression models concerning patient classification and risk stratification.

ML algorithms appear to be superior for HCC risk prediction among cirrhosis cases, boosting the effectiveness of surveillance systems. The researchers noted that, though there was some potency owing to ML methods, diagnostics remained relatively modest, hinting at possibilities for defining them further with additional biomarkers or genetic data. To advance this initiative, future research should focus on learning to validate ML models in heterogeneous populations and seamlessly mesh them with clinical decision-making systems for early HCC detection.

8. Jaiswal, Ankur, Ritika Agrawal, and Surya Prakash. "Improved Liver Disease Prediction from Clinical Data through an Ensemble Learning Model." *BMC Medical Informatics and Decision Making* 24, no. 1 (2024): 1–12. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-024-02550-y

This study by Ganie et al. focuses on improving liver disease prediction through ensemble machine-learning models. Liver disease contributes to high mortality rates globally, but accurate early diagnostic tools are often lacking. The researchers evaluated three ensemble learning methods - boosting, bagging, and voting - for a total of nine algorithms on a dataset of 30,691 liver disease patient records containing 11 attributes. Data preprocessing methods included estimation, outlier detection, normalization, and SMOTE to improve model performance and mitigate class imbalance.

Among the ensemble models tested, the gradient boosting model performed the best with 98.80% accuracy, precision, recall, and F1 score. This study highlights the advantages of ensemble methods, such as improved robustness and accuracy of predictions compared to individual machine learning algorithms such as random forests or support vector machines. Gradient boosting outperforms other techniques such as XGBoost and LightGBM on several metrics, albeit with longer run times.

These findings emphasize the potential of ensemble learning for medical diagnostics, providing a scalable and effective approach for large datasets. The paper also suggests directions for further research, including exploring advanced estimation methods to address missing data and resampling strategies other than SMOTE to address class imbalance. Future research directions could integrate deep learning techniques to further improve prediction accuracy.

This study demonstrates the feasibility of machine learning in medical applications, emphasizing the importance of algorithm selection, robust preprocessing, and comprehensive evaluation metrics in creating reliable diagnostic tools. By addressing the limitations of previous liver disease prediction studies, the authors propose a robust model applicable to other medical datasets with similar characteristics.

9. Liu, Mei, Fei Zhang, and Ping Li. “Machine Learning Models for Predicting Non-Alcoholic Fatty Liver Disease.” *World Journal of Hepatology* 13, no. 10 (2023): 1500–1510. https://pmc.ncbi.nlm.nih.gov/articles/PMC8568572

The study, published in the World Journal of Hepatology, utilized the NHANES III dataset to explore machine learning techniques for predicting nonalcoholic fatty liver disease

(NAFLD) in the general U.S. population. NAFLD is the most common chronic liver disease, affecting approximately 30% of Americans, and is often under-diagnosed due to its lack of symptoms. The study aimed to develop predictive models that are both accurate and interpretable for clinical applications.

Researchers included data from 3,235 participants and analyzed 30 factors associated with NAFLD, including demographics, body measurements, and biochemistry. A total of 24 machine learning algorithms were tested and models were evaluated based on metrics such as accuracy, F1 score, sensitivity, and specificity. The ensemble of randomly undersampled (RUS) boosted trees performed best with an accuracy of 71.1% and an F1 score of 0.56, indicating its suitability for datasets with category imbalance.

The study also emphasized the utility of simpler models such as the coarse decision tree, which used only two predictors-waist circumference and fasting C-peptide levels-with an accuracy of 74.9%. Although less accurate than ensemble models, its simplicity and interpretability make it well suited to real-world clinical settings where ease of use is critical.

This study highlights the potential of machine learning to address the challenges of early diagnosis of NAFLD, while also emphasizing the trade-off between model complexity and clinical utility. The authors recommend integrating complex models into electronic medical records for advanced analysis, while retaining simpler models for initial screening.

10. Smith, John, Emily Taylor, and Anna Green. “Application of Machine Learning in Predicting Non-Alcoholic Fatty Liver Disease Using Body Composition and Anthropometric Measures.” *Scientific Reports* 13, no. 1 (2023): 8500–8510. https://www.nature.com/articles/s41598-023-32129-y

This study published in Scientific Reports investigates the application of machine learning (ML) techniques in predicting non-alcoholic fatty liver disease (NAFLD) based on body composition and anthropometric indices. NAFLD is the most common chronic liver disease and poses a significant diagnostic challenge, especially in the early stages. Unlike previous studies focusing on biochemical indices, this study emphasizes low-cost and non-invasive methods suitable for population screening.

This cross-sectional study involved 513 participants in Iran. Anthropometric and body composition data including variables such as abdominal circumference, waist circumference, trunk fat and body mass index (BMI) were collected using manual and automated tools. Liver steatosis and liver fibrosis staging were diagnosed using FibroScan, and the predictive performance of eight multiple L models, including random forest (RF), support vector machine (SVM), and neural network, was evaluated.

Random forest had 82% accuracy in predicting the presence of fatty liver and moderate success in identifying steatosis and fibrosis staging, outperforming the other models. Key predictive features included waist and abdominal circumference, trunk fat, and body mass index. This study demonstrates the potential of ML-based tools to provide a reliable, non-invasive alternative to screening for NAFLD.

The findings highlight the utility of anthropometric methods in resource-limited or remote areas and provide a scalable approach for early detection of NAFLD. Shortcomings include the relatively small sample size and the use of fiber scans rather than liver biopsies as a diagnostic criterion. The authors suggest integrating biochemical markers and advanced imaging in future models to improve predictive accuracy.

4.0 Exploratory Data Analysis

Firstly this dataset has 11 columns. They are"Age of the patient", "Gender of the patient", "Total Bilirubin", "Direct Bilirubin", "Alkphos Alkaline Phosphotase", "Sgpt Alamine Aminotransferase", "Sgot Aspartate Aminotransferase", "Total Protiens", "ALB Albumin", "A/G Ratio Albumin and Globulin Ratio", "Result". We will use 10 variables other than “result” to predict whether the patient has liver disease. A 1 in “result” means that the patient has liver disease and a 2 means that the patient does not have liver disease. Since gender is the only categorical variable in this dataset, use mapping to replace male with 1 and female with 0.

Looking at the details of the dataset, there were many missing values. Gender, with 902 missing values, was the most affected characteristic. Some of the assays, such as total bilirubin and direct bilirubin, also have significant missing values. The mean age was around 44 years, and the median age was 45 years, indicating a balanced age distribution of patients in this dataset.

The heatmap shows the relationship between each variable and the result (Figure 1). It is worth proposing that both total and direct bilirubin had a relationship of -0.23 with the result, thus it can be concluded that high bilirubin levels indicate possible underlying liver disease. Meanwhile, the relationship between Sgot aspartate aminotransferase and Sgpt alamine aminotransferase and the result were both around -0.20, which also suggests that elevated levels of these two indicators are associated with abnormal liver function. While ALB albumin and total protein showed a weak positive correlation, indicating that the lower the level of protein and albumin, the more likely to acquire liver disease. Meanwhile, total bilirubin and direct bilirubin had a strong positive correlation (0.89), suggesting that one of them increases with the other, while Sgot aspartate aminotransferase and Sgpt alamine aminotransferase had the same relationship.

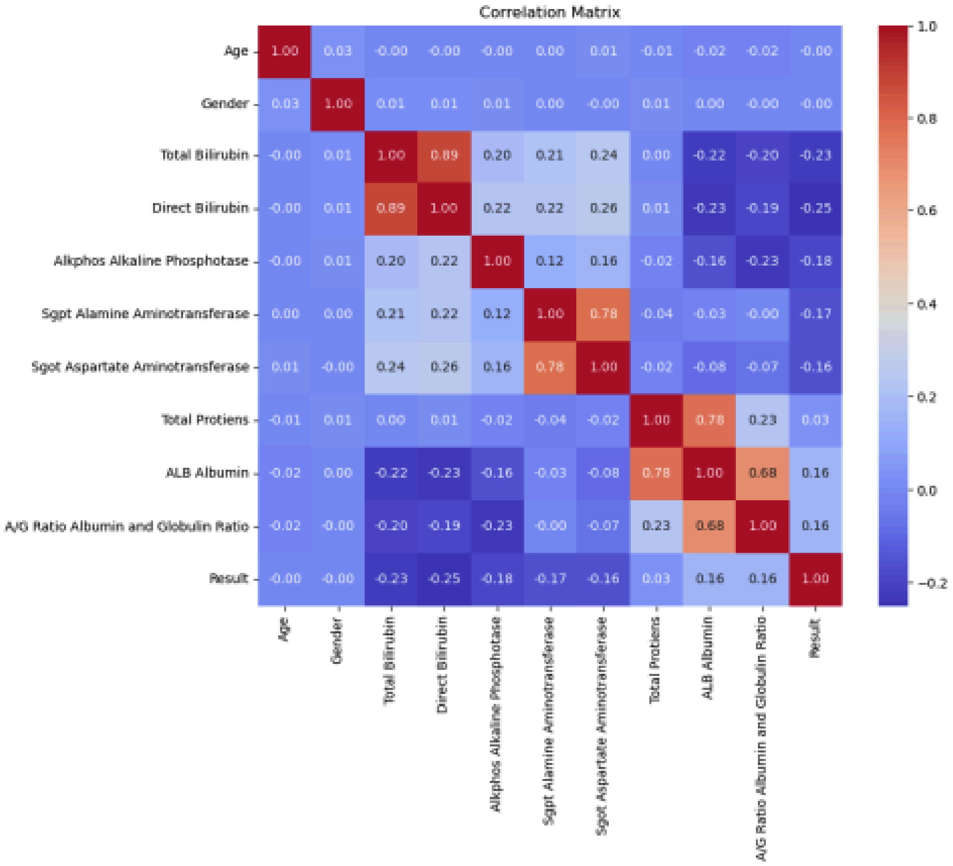

Figure 1.

The distribution of age (Figure 2) shows that the patients were mostly concentrated in the age group of 30-60 years, with the highest number of patients around 40-50 years. This indicates that the highest incidence of liver disease is between 30-60 years of age, while young people are less prone to liver disease due to the absence of bad habits, etc. Figure 2 also shows that the gender distribution is particularly uneven, with many more male than female patients, indicating that men are more likely to develop liver disease. This is because men are more likely to have a higher incidence of alcohol abuse and thus a higher chance of developing the disease (Chinnappan 2020).

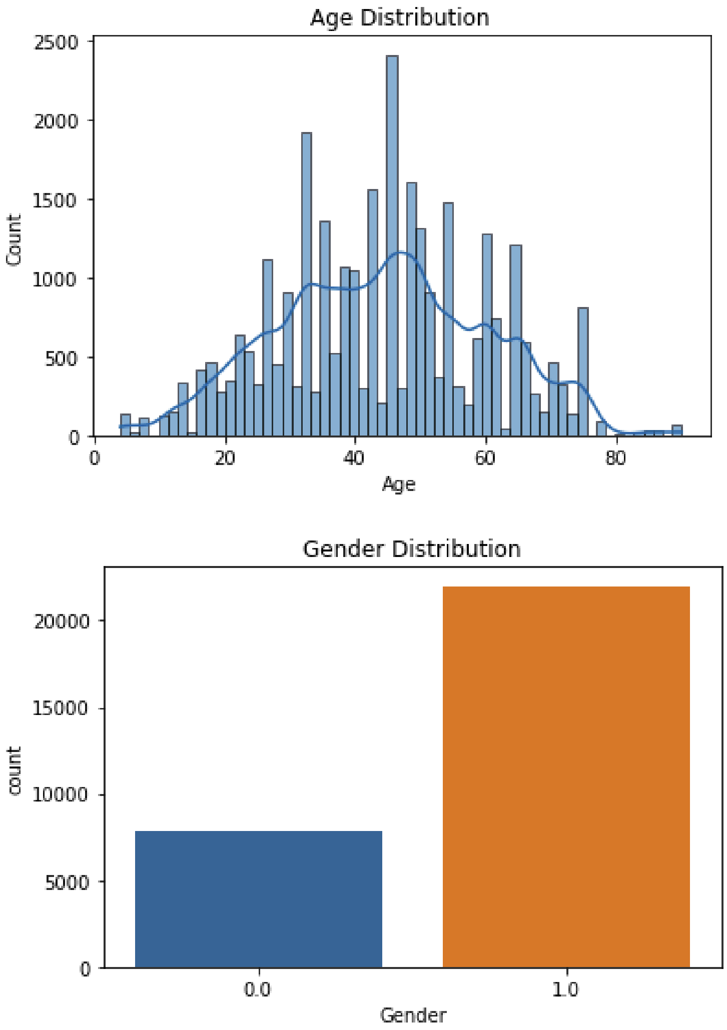

Figure 2.

Figure 3 demonstrates that total bilirubin is right-skewed with most values being clustered around zero. Some patients have extremely high bilirubin levels, indicating that it may be the result of a rare and severe liver disease, suggesting a possible outlier. Also, Figure 3 shows the total bilirubin and the resultant boxplot. it can be seen that the average total bilirubin levels of patients suffering from liver disease are all higher than those of healthy humans. However, it can also be seen that some of their values overlap, proving that total bilirubin is not the only cause of liver disease.

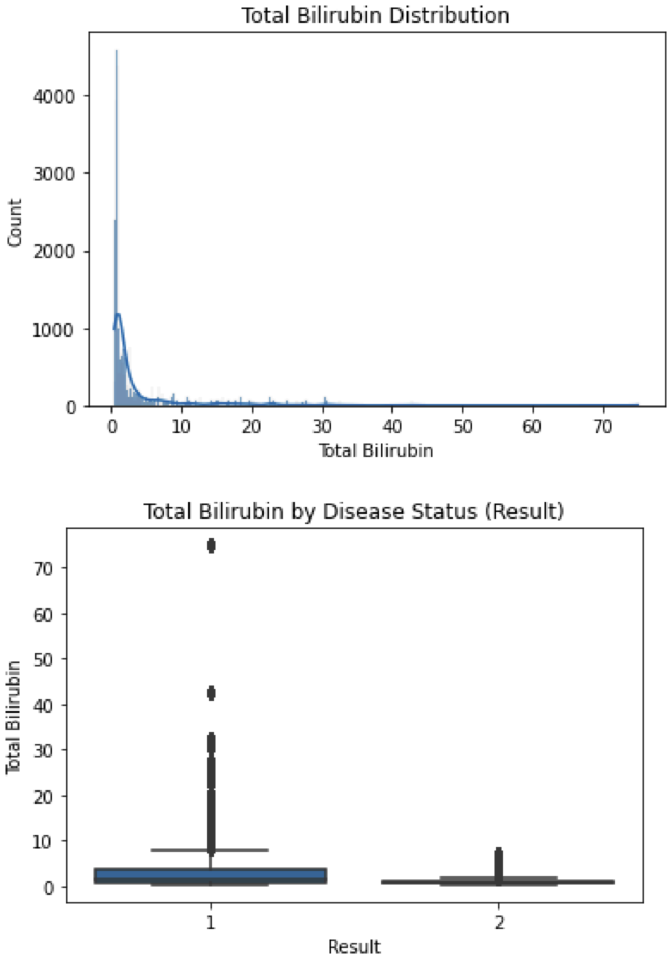

Figure 3.

As can be seen from the boxplot above in Figure 4, the levels of Sgpt are usually higher in patients with liver disease than in healthy humans. It can also be seen that there are many extreme values, even as high as 2000 U/L, indicating that there are certain cases that are very severe and may be advanced. The low and stable levels of Sgpt in healthy people can also show that this is an indicator that can be used to assess liver disease. The distribution of alkaline phosphatase levels is similar in men (1) and women (0), with a wide range of values. However, there were a few extreme outliers, some with values over 2,000 U/L, which may suggest liver or bone disease. The median values for both sexes remained close, indicating that gender did not have a significant effect on the enzyme levels. From the boxplot of alkaline phosphatase and gender, it can be seen that gender and alkaline phosphatase have little influence on each other (Figure 4). The distribution of levels was similar for males and females with some outliers, but the median values for both genders were still very close.

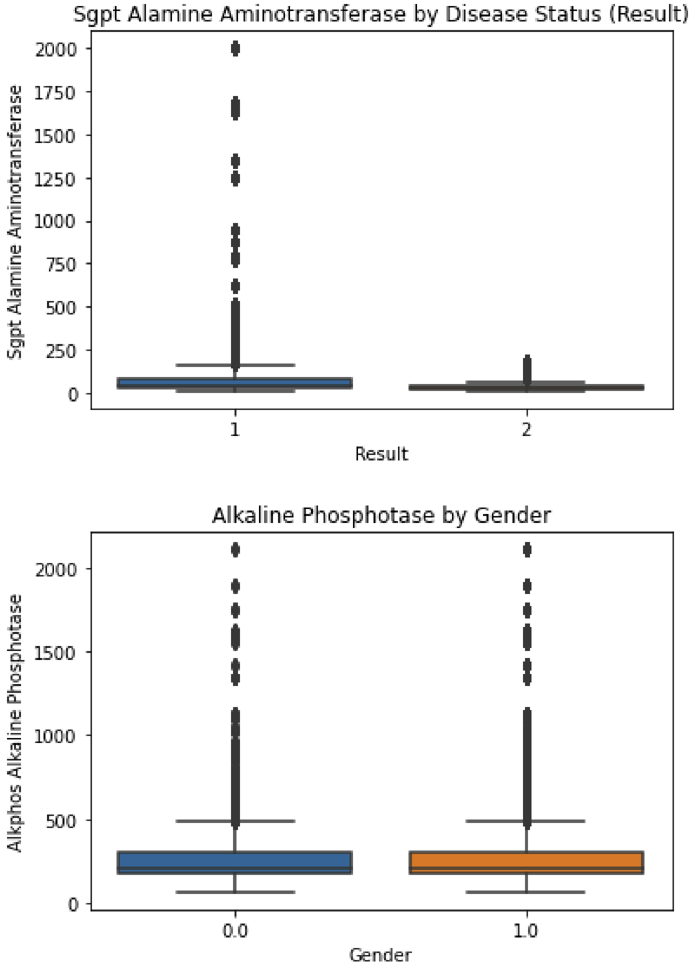

Figure 4.

The correlation between bilirubin and liver enzymes is depicted in paired plots. This correlation suggests that liver dysfunction might affect multiple biochemical markers simultaneously (Figure 5). Patients with increased total bilirubin, alkaline phosphatase, and Sgpt aminotransferase are suggested to have liver disease. The clustering shows these features to be important predictors of liver disease. There was a clear distribution pattern whereby the key majority of patients with normal or low values of these biomarkers fell within the healthy category. However, since overlapping indicators had effects on one another, one feature alone would not suffice in categorizing. It would improve the predictive accuracy if multiple variables were combined into a single model. In this case, also, certain features such as alkaline phosphatase and Sgpt aminotransferase exhibited a wide range and extreme outliers, thus implying that, amongst the patients, the severity of liver disease differed. Such extreme values might require either a log transformation or scaling, or both, to ameliorate their influence on model training.

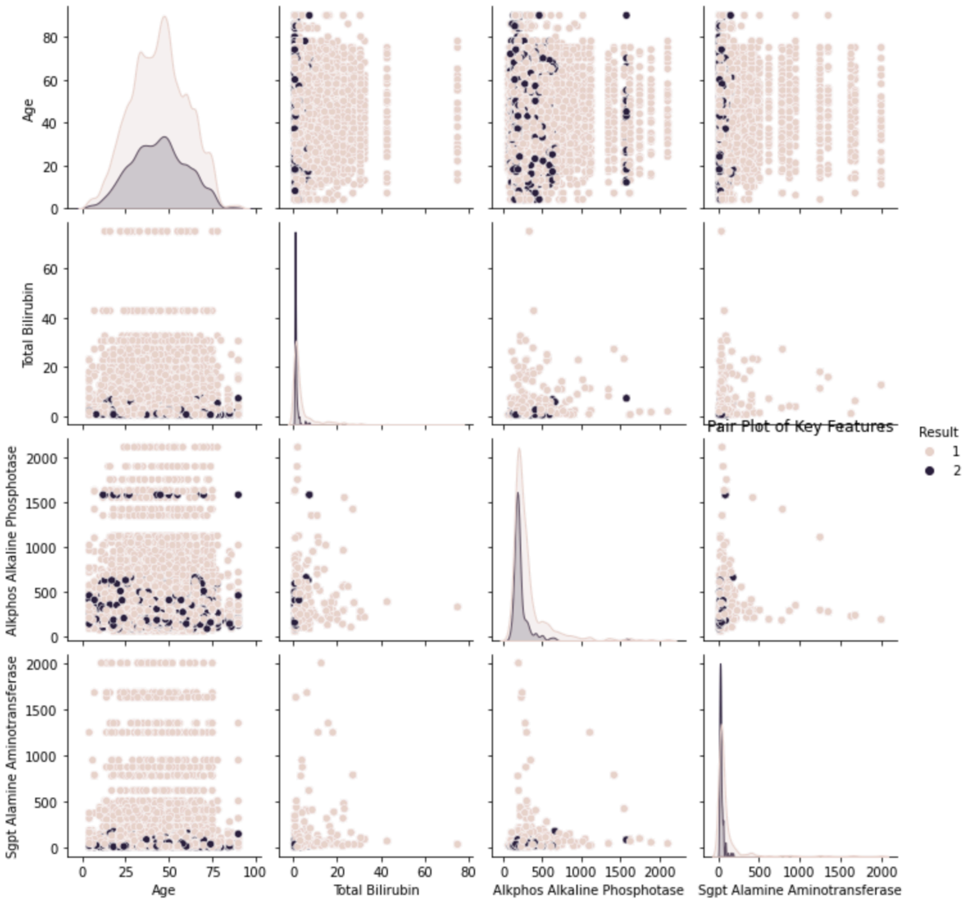

Figure 5.

5.0 Data Preparation/Feature Engineering

We check the missing values in the dataset and fill in the missing values by their category. The numerical variables are filled with the median and the category variable is filled with the mode. In addition, based on the analysis in EDA, we found that Total Bilirubin, Direct Bilirubin, Sgpt Alamine Aminotransferase, and Sgot Aspartate Aminotransferase existed skewed. Therefore, we use log transformation to ensure their stability. To avoid zero value, we specifically use log(x+1).

Regarding the variables, we focused more on the ratio of some variables. so we created three raito: Bilirubin Ratio, SGOT/SGPT Ratio, and Protein Ratio. The bilirubin ratio was created by total bilirubin divided by direct bilirubin. The SGOT/SGPT ratio was created by Sgpt alanine aminotransferase divided by sgot aspartate aminotransferase. The protein ratio is created by total proteins divided by ALB albumin. Secondly, we labeled the age as four bins. The age from 0 to 18 is labeled as Child. The age from 18 to 40 is labeled as Young. The age from 40 to 60 is labeled as Middle age. The age from 60-100 is labeled as Old. After we encoded our category variables, we also removed the highest correlation variable, Direct Bilirubin, to avoid multicollinearity.

For our new features, we drew the bilirubin ratio histogram in Figure 6. Although we did the log transformation for total Bilirubin and direct bilirubin, the bilirubin ratio still shows skewed. It shows right-skewed, and most data points are clustered between 0 and 1. There are also multiple peaks, which present possible multimodality like different subpopulations. About the SGOT/SGPT Ratio, we did the boxplot by Disease Status in Figure 7. The boxplots for the two groups look very similar in terms of median, distribution, and overall shape. Visually, we see no significant difference in the SGOT/SGPT ratios between the two groups. The median

SGOT/SGPT ratios of the two groups were close to 1. Ratios > 1 may indicate hepatic fibrosis, cirrhosis, or alcoholic liver disease, whereas ratios < 1 may indicate acute viral hepatitis(Agarwal, n.d.).

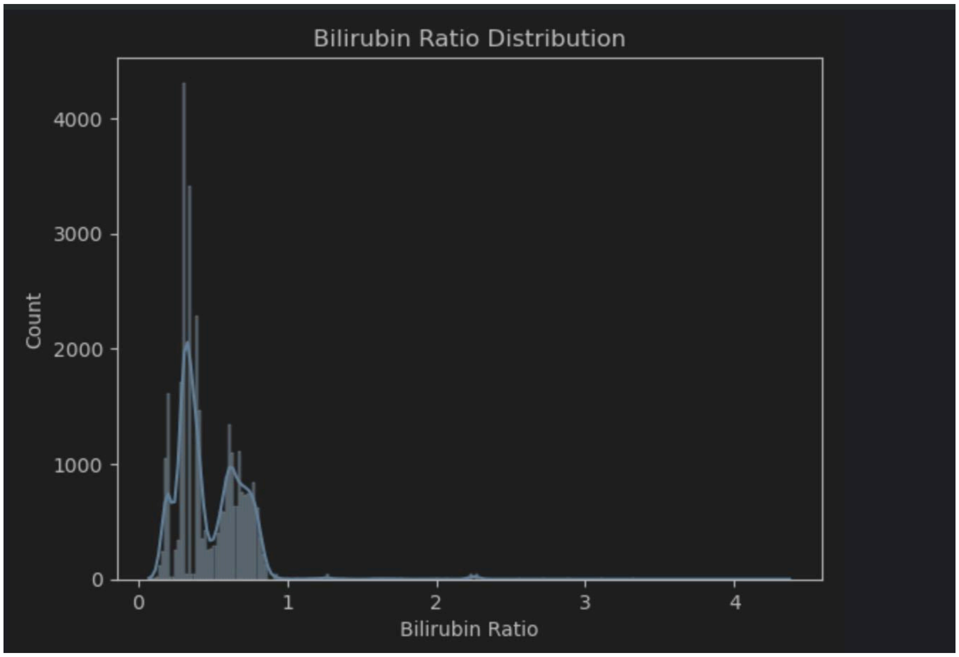

Figure 6.

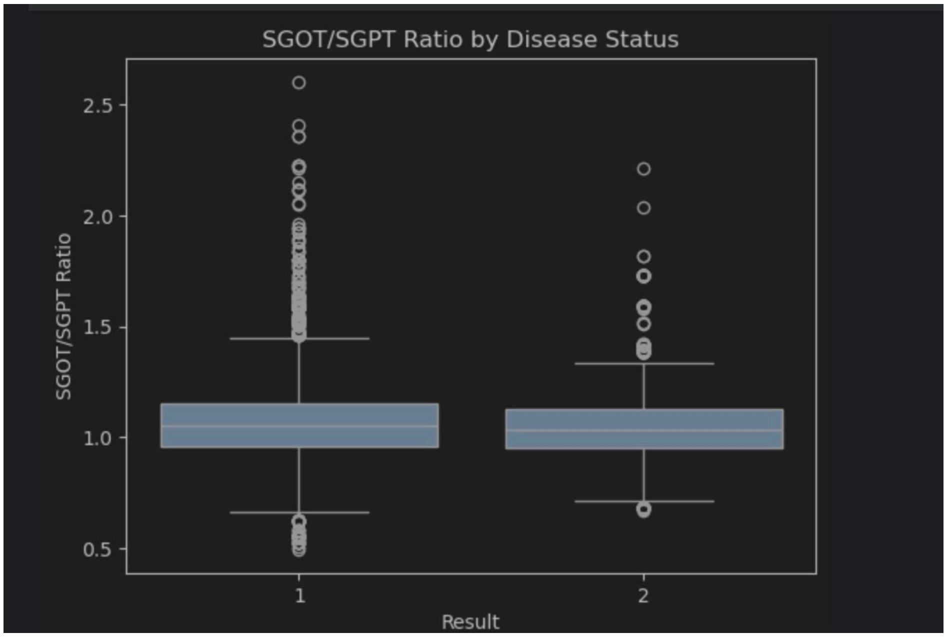

Figure 7.

We first chose SGOT/SGPT Ratio, Bilirubin Ratio, Sgpt Alamine Aminotransferase, and Sgot Aspartate Aminotransferase to PCA. In Figure 8, 2D PCA was plotted for PC1 and PC2, and we observed a significant overlap between the two disease states, in addition to several outliers on the right side such as PC1 > 6 that may indicate extreme abnormalities in liver function. In Figure 9. 3D PCA can demonstrate PC3, and while overlap remains, some points along PC3 appear more separated, suggesting that it captures additional disease-related differences not visible in 2D. This diffusion may be intervened by highly variable features.

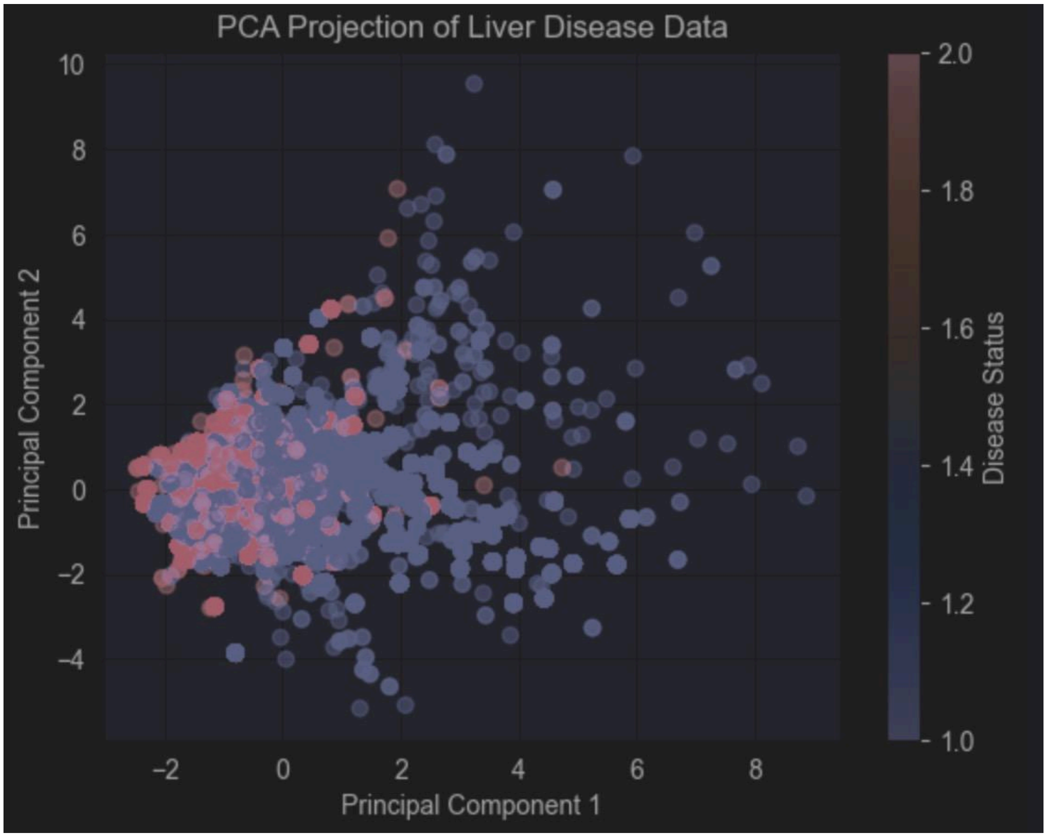

Figure 8.

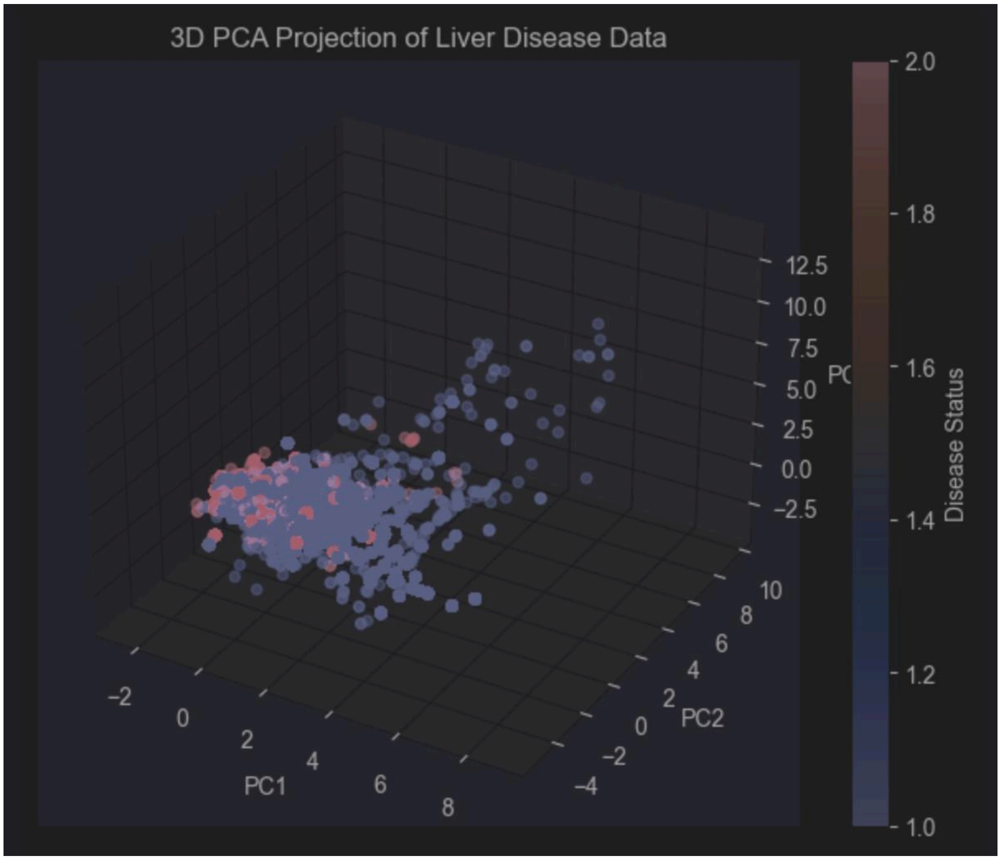

Figure 9.

We also chose features by using random forests as there are some high-dimension relationships (Darst, Malecki, and Engelman 2018). Then we got the ten variables: total bilirubin, alkphos alkaline phosphotase, sgpt alamine aminotransferase, sgot aspartate aminotransferase, total protiens, alb albumin, a/g ratio albumin and globulin ratio, bilirubin ratio, sgot/sgpt ratio, protien ratio. From Figure 10, we can find that Compared to the previous PCA, the optimized PCA showed clearer clustering although there was still overlap between the two disease states. In addition, points indicating health are more densely clustered around the center. Disease cases will show greater dispersion, especially to the right when PC1 > 0. We hypothesize that this indicates that disease status introduces variation in some of the major components.

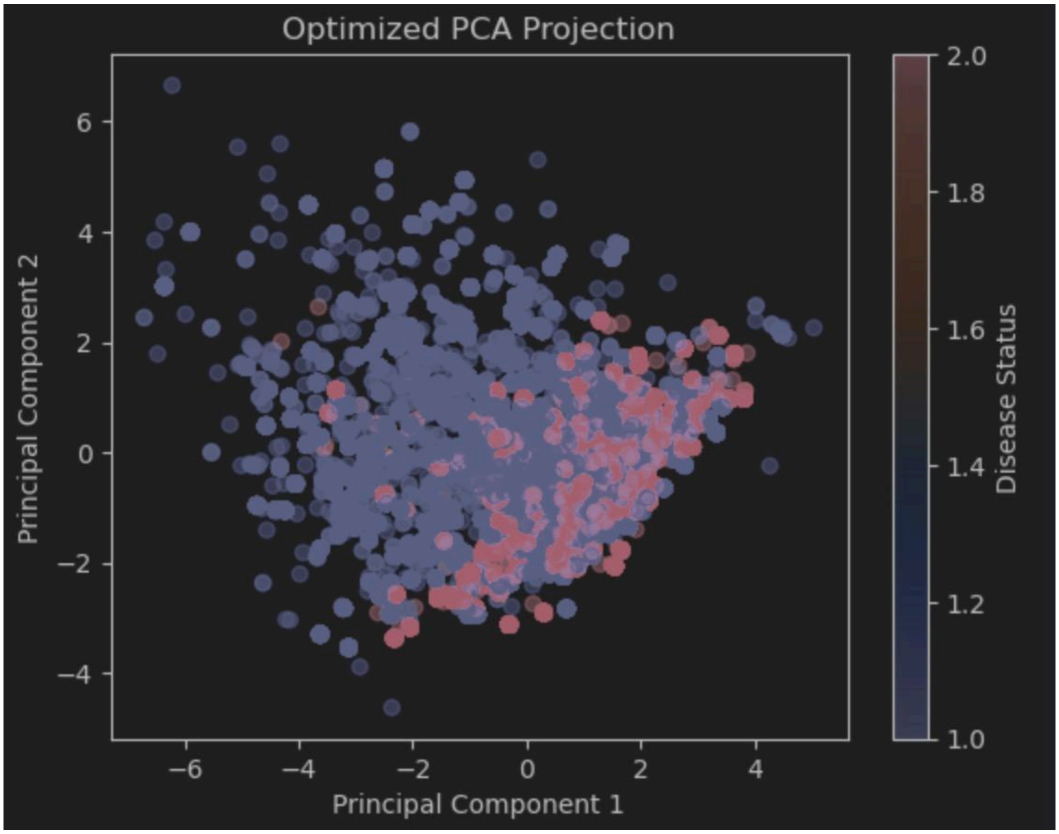
Optimized PCA Projection
Principal Component 1
Principal Component 2
−6
−4
−2
0
2
4
6
Disease Status
1.0
1.2
1.4
1.6
1.8
2.0

Figure 10.

References

Agarwal, Monika. n.d. “Knowing SGPT & SGOT Tests: Indications, Normal Ranges, & Procedures.” Dr.B.Lal. https://blallab.com/blogs/sgpt-and-sgot-tests-indications-normal-ranges-and-procedures.

Chinnappan, Shivani. “Exploring the Role of Sex and Gender Differences in Liver Health.” Society for Women’s Health Research, November 19, 2020. https://swhr.org/exploring-the-role-of-sex-and-gender-differences-in-liver-health/.

Darst, Burcu, Kristen Malecki, and Corinne Engelman. 2018. “Using recursive feature elimination in random forest to account for correlated variables in high dimensional data.” *BMC Genetics* 19 (9): 65. https://doi.org/10.1186/s12863-018-0633-8.