# Machine Learning Approach to Predict Survival State of Primary Biliary Cholangitis Patient

Konsta Laurila
konsta.laurila@student.oulu.fi

## Abstract

*Early disease detection and treatment are crucial for improving patient outcomes and reducing mortality rates. Advancements in machine learning (ML) and artificial intelligence (AI) have opened up new possibilities for more accurate and timely disease detection. While ML models have shown promise in predicting disease prognosis, their complexity and potential flaws require careful consideration and fine-tuning of predictors and parameters.*

*In this study, I utilized data from the D-penicillamine trial conducted by the Mayo Clinic to develop a supervised random forest classifier for categorizing Primary Biliary Cholangitis (PBC) patients into three groups based on their survival state. The dataset underwent preprocessing to ensure completeness, and model parameters were tuned using grid search.*

*To assess the performance of the model, I employed k-fold cross-validation and evaluated metrics such as accuracy, recall, and receiver operating characteristic (ROC) curves. Although the performance of the model was not optimal, my study demonstrates the feasibility of predicting the survival state of PBC patients using ML techniques, even with simple algorithms and tuning.*

*Moving forward, further research comparing different algorithms, feature combinations, and more comprehensive tuning approaches could lead to the development of more robust and accurate models for predicting disease outcomes in PBC patients.*

## 1. Introduction

Primary Biliary Cholangitis (PBC), previously recognized as Primary Biliary Cirrhosis, is a slowly progressive, cholestatic autoimmune liver disease characterized by cholestatic dysfunction [1]. Despite advancements in medical science, PBC's poor prognosis and few effective treatments often leads to more advanced levels of PBC or even mortality. PBC results from a combination of environmental triggers, individuals susceptibility, and epigenetic factors [5].

No known cure is available for PBC but the prognosis of the disease can be delayed. Studies have shown that the introduction of Ursodeoxycholic Acid (UDCA) has shown the effect of slowing down the PBC. Ultimately liver transplantation (LT) is the only option for patients with advanced or end-stage PBC, but the selection of patients needing LT has been a problem. The scarcity of available donor organs increases the importance of accurately identifying patients who would benefit from LT in the early stages of the disease [5].

The selection of patients needing a liver transplant is difficult. Machine learning has shown great results in prediction of diseases and prognosis in medical use cases before [3]. By using ML algorithms, especially supervised learning methods, it is feasible to construct predictive models capable of identifying patients at risk of adverse outcomes, including death from PBC. A working model could reduce the amount of futile LT procedures while ensuring that patients with dire need of transplantation receive appropriate treatment.

In this project a supervised random forest classifier is constructed to classify patients with PBC into three categories: Death (D), censored (C), or censored due to liver transplantation (CL). Imbalances in the dataset, lead to the implementation of second model with class CL data being removed to reduce biases. This approach prioritizes prediction features that can be collected with minimal invasiveness i.e. blood test and non-invasive measurements, while ensuring patient comfort and efficient data collection like suggested in the paper *Prognosis in Primary Biliary Cirrhosis: Model for Decision Making* by Dickson et. al. Cirrhosis Patient Survival Prediction-dataset [2] from 1984 with 424 PBC patients was used to train and validate this project's ML model.

In summary, PBC is a lethal liver disease when untreated with poor prognosis and few treatment options. LT is done to cure the disease but livers are scarce and can't be handed out to everyone. The decision of patients treatment type for the PBC is crucial for the patients survival. With the correct use of ML techniques and parametrization, the decision making process can be made faster and more efficient.

For ML to be used in PBC survival state predictions the features needs to be carefully selected and the model needs to be optimized taking the problems nature into consideration.

## 2. Related Work

As PBC is a lethal disease, the use of ML in many aspects of its treatment, identification and prognosis has been studied. This project closely refers to the paper *Prognosis in Primary Biliary Cirrhosis: Model for Decision Making* by Dickson et. al., published in 1989 which used the same dataset to fit the data to a novel pragmatic model to predicting survival of a PBC patients. The paper introduced the idea of using small number of features that are inexpensive, non-invasive to collect and are universally available. At the time liver biopsy was widely used method for collecting information about PBC patients. Dickson et. al. found that their model was comparable in quality to other survival predicting models which used liver biopsy for their prognosis. While the paper published by Dickson et. al. can be considered outdated in today's research it provides valuable information and methods about comparison of predictive models.

In 2022 Hanif & Khan studied the possibility of diagnosing liver cirrhosis in early stages of the disease with three different machine learning approaches. In their study, they utilized open-access Liver Cirrhosis dataset to train and validate Support Vector Machine, Decision Tree Classification, and Random Forest Classifier models to forecast the possibility of liver cirrhosis infection with similar features. In their paper, all three models achieved high evaluation scores in around 97%. Which can be an indication of well performing model or over fitted model. While using similarly sized dataset they might have encountered the problem with imbalanced dataset. Only having 12% of the cases labeled positive for liver disease can lead to biases in the models. While the results of Hanif & Khan's study could be re-evaluated, the study underscores the importance of critically analyzing the model's performance and the process of data collection and preprocessing.

Another study, conducted by Kanwal et al. in 2020 [7], focused on predicting mortality in PBC patients. They used three different machine learning approaches to pick predictors to use with Cirrhosis Mortality Model (CiMM). The predictors were then used to predict the mortality of cirrhosis patients. In the study, Gradient descent boosting, logistic regression with least absolute shrinkage and selection operator (LASSO) regularization, and logistic regression with LASSO constrained were used for the selection of the predictors. Among thesem, the best performer were used to refit the predictors for the CiMM model. Results from CiMM were compared to Model for End Stage Liver Disease with sodium (MELD-Na) score to evaluate the predictive performance of their model. Kanwal et. al. study had significantly larger dataset with over 100 000 instances. Larger datasets will more likely lead to more accurate models on a specific problem. Their study shows that the right approach and careful consideration of the predictors and models ML can yield proficient results.

There have been various studies about the subject which aiming to aid patients suffering from PBC. Although many studies have been conducted to develop ML models predicting mortality rate, cirrhosis stage etc. further studies are required to ensure the robustness of the predictors and models.

## 3. Proposed Method

In this project, various preprocessing methods and algorithms are used to process the dataset before a random forest classifier is trained and evaluated. After preprocessing the data, the model is trained and grid search is used to optimize the models parameters from given initial values. Parameters from the grid search is used to train and validate the model using 5-fold cross-validation method to compensate for the small size of the dataset. Figure 1 shows the schematic diagram of this projects ML process.

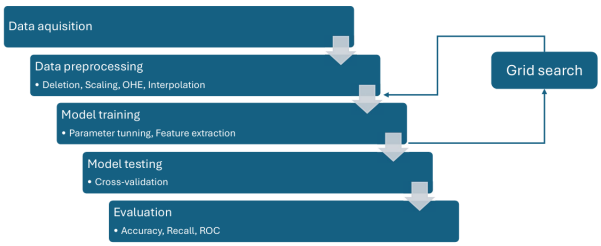


Figure 1: Model Diagram

### 3.1. Preprocessing

Inspecting the dataset revealed some properties that needed attention before the data could be used for ML application. For this dataset missing values, categorical data and uneven value distribution were present. For this specific dataset the required preprocessing steps to tackle these problems were deletion, interpolation, One-hot-encoding (OHE) and scaling.

#### 3.1.1 Deletion and interpolation

Some incidents needed to be deleted from the dataset due to incomplete measurements. Column with cholesterol values included some missing values which were handled by interpolating the mean of the cholesterol values to the missing incidents.

### 3.1.2 One-Hot encoding

The dataset included categorical variables which need to be converted to integers for the model to work. For this conversion I used OHE. OHE creates new variables to substitute one string to one integer. For example, this dataset included a binary variable for presence of ascites with "Y" meaning that the patient had ascites and "N" meaning the absence of ascites. In this case OHE created *ascites_Y* and *ascites_N* variables and if the patient had ascites the variable *ascites_Y* and *ascites_N* would be set to "1" and "0" respectively.

## 3.2. Scaling

Scaling is used to reduce the algorithms convention to bias greater values. It can happen because the algorithm doesn't know the difference in units, it only sees values. Scaling is done to handle highly varying magnitude of values which is present in this dataset. Scaling converts the numerical values to a fixed scale. In this case zero mean scaler was used to reduce the effect of highly varying values in the features.

## 3.3. Model and tuning

Various ML algorithms exits, all with their own benefits. In this project, I'm proposing the use of random forest classifier to predict the state of the PBC patient to three classes discussed before. Random forest classifier was chosen for this project for its ability to not overfit, high accuracy and robustness.

Decision tree consists of root node, branches, internal nodes and leaf nodes. Decision tree starts with root node which then branches to chosen path to an internal node with another decision for it to make. For example in this dataset the decision tree will come up with a internal node with question about the presence of ascites and it will tilt to specific class depending on these values. This decision process continues until it reaches leaf node which will be the output class. Figure 2 shows an example of a simple decision tree with only 2 internal nodes.

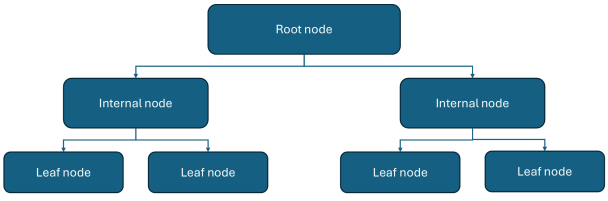


Figure 2: Decision tree schematic

Random forest classifier (Figure 3 uses multiple decision trees together to predict the class for the given data. Given data is send to predetermined number of trees which all will give its output class. The output class acts as a vote. These votes are then compared to the other trees output classes. The class with majority of votes will determine the final output class for the given data.

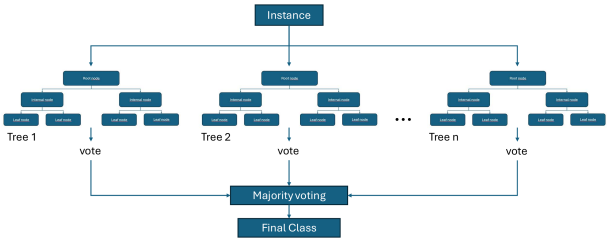


Figure 3: Random forest schematic

Random forest classifiers can be parameterised in various ways. For this project max depth of the tree, minimum number of samples required to be at a leaf node, minimum number of samples required to split an internal node and number of estimators were adjusted with grid search.

Grid search trains the model with multiple different parameter options and evaluates the performance of each model. Then it outputs the parameters which gave the best results. These results from grid search were used to train the final model in this project.

## 3.4. Evaluation methods

To get most of a small dataset, k-fold cross-validation was used, in this case 5-fold. In k-fold validation the data set is split into k number of folds. 1 of the folds is used as testing set and the rest are used to train the model. Then another fold is used to test the model while others are used to train the model. This is done until all folds have been used to test the model. The mean result can be then calculated to evaluate the model. With cross-validation one can get more training data without the need of gathering the data. This can be beneficial when the data collection is difficult and slow and the datasets remain small.

Three evaluation metrics are used o evaluate the performance of the model constructed for the project: accuracy, recall and ROC AUC. These three metrics give good overview of the performance of the model together with confusion matrices. In addition another model is trained without the CL class to see the effect of imbalanced dataset while using random forest classifier.

# 4. Experiments

## 4.1. Dataset

Cirrhosis Patient Survival Prediction dataset [2] was used in this project work. Data includes 418 instances with 17 histological and other features of PBC patients. The dataset is based on Mayo Clinics study of randomized placebo-controlled trial testing the drug D-penicillamine. D-penicillamine is nowadays used mainly to treat patients

with Wilson's disease, which is a disease where excess copper builds up in the body. The study was conducted from 1974 to 1984.

Dataset is in tabular form which can be efficiently manipulated with Python's pandas library. Dataset used in this project work includes integer, continuous and categorical variables, missing values and imbalances. Methods used to tackle these problems were discussed earlier in chapter 3.

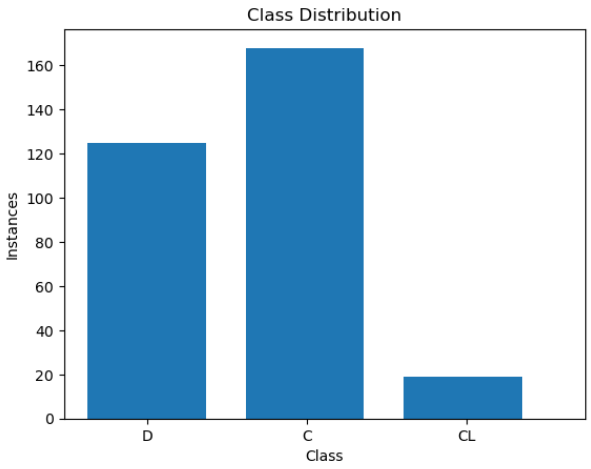


Figure 4: Target variable class distribution

Target variable in this dataset set is the survival state of the patient, which was classified into three categories Censored (C), Censored due to liver transplant (CL) and Death (D). Target variable showed major imbalances. Only 19 out of the 312 instances being classified as CL as shown in the figure 4. This lead to the decision of training two models with one using the low incident variable (CL) and other with the CL incidents removed.

112 of the patients didn't join the clinical trial but agreed to record basic metrics and undergo survival tracking in the original dataset. These incidents were excluded from the dataset due to missing measurements.

The 17 features and descriptions of the dataset is shown in the table below. From the 17 features (Figure 3), 12 of them can be measured and collected with minimal invasive methods like Dickson et. al. suggested in their paper. The 12 features included in the model; age, ascites, hepatomegaly, spiders, edema, bilirubin, cholesterol, albumin, platelets, prothrombin, aspartate transaminase (SGOT), and alkaline phosphatase.

Missing values were present in Platelets and Cholesterol columns showed 4 and 28 instances respectively. Missing values were handled by interpolating the mean value to the missing instances.

Categorical data included in the final dataset was handled with OHE to get binary variables like discussed before, which lead to 4 features (Ascites, Hepatomegaly, Spiders, and Edema) to be encoded as 9 features (Ascites_N, Ascites_Y, Hepatomegaly_N, Hepatomegaly_Y, Spiders_N, Spiders_Y, Edema_N, Edema_S, and Edema_Y). Moreover the target variable was processed with OHE to get three target variables; Status_C, Status_CL, and Status_D.

| Variable name | Role | Type | Description | units |
|---|---|---|---|---|
| ID | ID | Integer | unique identifier | |
| N_days | Other | Integer | number of days between registration and the earlier of death, transplantation or study analysis time in july 1986 | |
| Status | Target | Categorical | Status of the patient (C, CL, D) | |
| Drug | Feature | Categorical | Type of drug: D-penicillamine or placebo | |
| Age | Feature | Integer | age | days |
| Sex | Feature | Categorical | male (M) of female (F) | |
| Ascites | Feature | Categorical | presence of ascites N (No) or Y (Yes) | |
| Hepatomegaly | Feature | Categorical | presence of hepatomegaly N (No) or Y (Yes) | |
| Spiders | Feature | Categorical | presence of spiders N (No) or Y (Yes) | |
| Edema | Feature | Categorical | presense of edema N (No), S (present without diuretics) or Y (present with diuretics) | |
| Bilirubin | Feature | Continuous | serum bilirubin | mg/dl |
| Cholesterol | Feature | Integer | serum cholesterol | mg/dl |
| Albumin | Feature | Continuous | albumin | gm/dl |
| Copper | Feature | Integer | urine copper | ug/day |
| Alk_Phos | Feature | Continuous | alkaline phosphatase | U/liter |
| SGOT | Feature | Continuous | SGOT | U/ml |
| Tryglicerides | Feature | Integer | tryglicerides | |
| Platelets | Feature | Integer | Platelets per cubic | ml/1000 |
| Prothrombin | Feature | Continuous | prothrombin time | s |
| Stage | Feature | Categorical | histologic stage of disease (1,2,3,4) | |

Figure 5: Features

After preprocessing and feature selection, the final dataset had 312 instances with 3 target variables and 12 out of 17 features which corresponded to 17 predictors.

### 4.2. Software

Python 3 was used in jupyter notebook to construct the model. The notebooks are shared in GitHub [8]. Scikit-learn, matplotlib, pandas and numpy libraries for python were used in the project for visualisation, model construction, evaluation and data processing.

### 4.3. Hardware

Authors personal computer was used for the project (No notable hardware used).

## 5. Results and Discussion

### 5.1. Parameter tuning

The model parameters were tuned with GridSearchCV like discussed before. The grid search was done individu-

ally for both models. Parameter grid created for the search included hand-picked values for the parameters to be optimized. Parameter grid can be seen in the figure 6

```
# Define the parameter grid to search
param_grid = {
    'n_estimators': [50,100, 200, 300, 500],
    'max_depth': [None, 10, 20, 30, 40],  #
    'min_samples_split': [2, 5, 10, 15, 20],
    'min_samples_leaf': [1, 2, 4, 6, 8]
}
```

Figure 6: Parameter grid

Best values from the grid search are shown in the table below.

| Parameters | | |
|---|---|---|
| Parameter | Best value (with CL) | Best value (without CL) |
| max_depth | 20 | 30 |
| min_samples_split | 1 | 2 |
| min_samples_leaf | 15 | 2 |
| n_estimators | 100 | 100 |

Biggest difference from the grid search appeared to be in minimum samples required to be at a leaf node parameter and max depth of the trees. Overall the removal of class CL doesn't have a big effect of the tuning of the random forest classifier. It is still important step to reduce the risk of over and underfitting the model.

## 5.2. Evaluation

With the tuned parameters the models were trained and evaluated with accuracy (AC), recall (REC) and ROC AUC metrics. Also, mean confusion matrices were constructed to better visualize the distribution of the predictions. In the tables below the calculated AC, REC and ROC scores for each class and the mean are shown.

From the results it is clear that the dataset was imbalanced because it overfitted on the occasion of class CL with AC and REC being 93,9% and 100% respectively. ROC AUC scores for the 3-class model shows that the model is no better than a random guess when it comes to predicting a case of CL. What comes to the other classes it didn't really matter if the CL class was in the training or not.

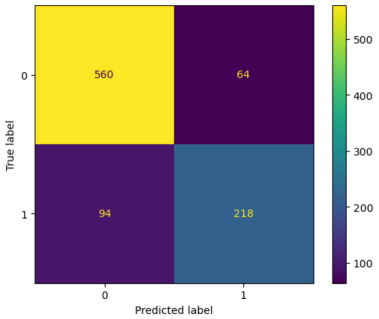


Figure 7: 3-class model

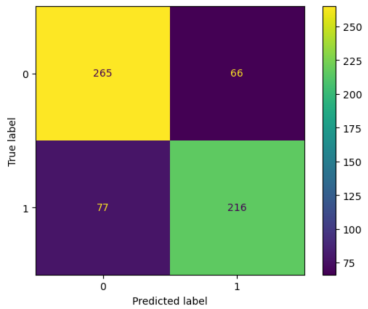


Figure 8: 2-class model

| Accuracy | | |
|---|---|---|
| Class | 3-class model | 2-class model |
| C | 78,5% | 77,9% |
| D | 76,9% | 76,3% |
| CL | 93.9% | - |
| Mean | 83,1% | 77,1% |

| Recall | | |
|---|---|---|
| Class | 3-class model | 2-class model |
| C | 73,6% | 70,8% |
| D | 86,1% | 87,2% |
| CL | 100% | - |
| Mean | 89,7% | 80,1% |

| ROC AUC score | | |
|---|---|---|
| Class | 3-class model | 2-class model |
| C | 0,782 | 0,774 |
| D | 0,746 | 0,736 |
| CL | 0,5 | - |
| Mean | 0,676 | 0,755 |

From the matrices (Figures 7 & 8), with both cases there are many instances with false negative prediction which isn't good when making possibly life threatening predictions. With that in this current stage this model shouldn't be used to predict the survival stage of a PBC patient without more optimization and tuning. Also, other approaches should be investigated to get more comprehensive image of the problem at hand.

In summary the 3-class model can predict the survival stage of a PBC patient with 83,1% ACC, 89,7% REC and ROC AUC score being 0,676. For the 2-class model results: 77,1% ACC, 80,1% REC and ROC AUC score being 0,755. Even though ACC and REC indicates that the 3-class model is "better" it needs to be noted that the metrics biased the imbalances in the dataset which can be seen in the ROC AUC score difference.

## 6. Conclusions

Results show that with some data processing, feature selection and use of random forest classifier, a model that predicts the survival state of PBC patient can be constructed and gives more accurate results than random guess. Absolute "goodness" of this model is hard to tell through this project, because of the lack of comparison to other ML algorithms. More research and implementation of different models could be done to more evaluate the performance better. Also, different combination of features and processing methods could be evaluated to see the effect in robustness of the model.

## 7. Acknowledgements

I thank Mayo Clinic for funding the original creation of the dataset and Fleming et. al. [2] for sharing it as open access. The dataset can be found at [9] . I appreciate the assistance of the course staff during the project in the form of materials and lectures.

## 8. Contributions

The project work was conducted individually by the author.

## References

[1] Dickson, E.R., Grambsch, P.M., Fleming, T.R., Fisher, L.D., and Langworthy, A. (1989). Prognosis in primary biliary cirrhosis: Model for decision making. Hepatology, 10.

[2] Fleming, Thomas R., and David P. Harrington. Counting processes and survival analysis. Vol. 625. John Wiley and Sons, 2013.

[3] Haug CJ, Drazen JM. Artificial Intelligence and Machine Learning in Clinical Medicine, 2023. N Engl J Med. 2023; 388: 1201-8

[4] Purohit, T., Cappell, M. S. (2015). Primary biliary cirrhosis: Pathophysiology, clinical presentation and therapy. World journal of hepatology, 7(7), 926–941. https://doi.org/10.4254/wjh.v7.i7.926

[5] Trivella J, John BV, Levy C. Primary biliary cholangitis: Epidemiology, prognosis, and treatment. Hepatol Commun.2023;7:e0179. https://doi.org/10.1097/HC9.0000000000000179

[6] I. Hanif and M. M. Khan, "Liver Cirrhosis Prediction using Machine Learning Approaches," 2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, NY, USA, 2022, pp. 0028-0034, doi: 10.1109/UEMCON54665.2022.9965718.

[7] Kanwal, F., Taylor, T. J., Kramer, J. R., Cao, Y., Smith, D., Gifford, A. L., El-Serag, H. B., Naik, A. D., & Asch, S. M. (2020). Development, Validation, and Evaluation of a Simple Machine Learning Model to Predict Cirrhosis Mortality. JAMA network open, 3(11), e2023780. https://doi.org/10.1001/jamanetworkopen.2020.23780

[8] https://github.com/KonstaLaurila/ML-in-medicine—project-work

[9] https://archive.ics.uci.edu/dataset/878/cirrhosis+patient+survival+prediction+dataset-1