Predicting Chronic Hepatitis B Disease Progression and Outcomes in the Gambia using Machine Learning Algorithms

Dorcas Asare 1,4 Ismaïl Koné 1,3 Bubacarr Bah 2 Gibril Ndw 2 Maud Lemoine 2 Rohey Bangura 4 Lamin Bojang 4

1Stellenbosch University 2MRC Unit The Gambia at LSHTM 3Université Virtuelle de Côte d’Ivoire(UVC) 4African Institute for Mathematical Sciences

Abstract

Hepatitis B virus (HBV) infection remains an important global health problem with high mortality and morbidity. There is a problem in the early detection of HB patients at high risk and its contributing factors. The study is aimed to predict the mortality outcome of hepatitis B patients in The Gambia, and also, to identify the significant associated factors. Six (6) machine learning models were used to predict the mortality of HB patients namely Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Naive Bayes (NB), Support Vector Machines (SVM), and K-Nearest Neighbor (K-NN). The F1-score for LR, DT, RF, NB, SVM, and K-NN models were 0.51, 0.96, 0.97, 0.53, 0.53, and 0.75 respectively with RF showing the best performance for predicting mortality outcome. The model can be deployed in clinical routine to further test its effectiveness and adjust it to better detect high-risk patients.

Introduction

- HBV disease which is inflammation of the liver puts people at high risk of death in its progression. According to WHO, Hepatitis B affects approximately 257 million people (1.7% of the world’s population) globally of which 60 million are from the Africa region [3].
- HBV causes approximately 800,000 deaths every year [3].
- The project is part of the PROLIFICA (Prevention of Liver Fibrosis and Cancer in Africa) program in The Gambia which seeks to alleviate the burden of HBV disease. The longitudinal study of HBV patients was set up in 2011 with approximately 2,600 adults.
- The objective of the study is to use different machine learning models to predict mortality and find the most important associated factors based on the PROLIFICA data.

PROLIFICA Dataset

- A follow-up study was done in 2018-present of which 52% (2116) patients died out of 1545
- The dataset has over 100 features that contain information on patients’ demographic, clinical, and family history of HBV.

Methods

After pre-processing of the data, 40 features remained and 681 patients were obtained for further analysis. 371 patients were alive whiles 191 (47%) were dead. Feature selection techniques like correlation and mutual information were used to filter out important features.

- Correlation measures the strength of association between features [1] whiles mutual information explains the amount of information obtained from the target variable [4].
- The data was split into training and validation for modeling.

- Logistic regression is used to estimate probabilities by using a logistic function.
- A decision tree models decisions to predict the value of the target variable by using a flowchart-like tree structure.
- Random forest consists of multiple decision trees to select output by averaging all of the tree outputs.
- Support vector machine classifies data points into the target classes with the help of a hyperplane.
- The k-Nearest Neighbors algorithm uses proximity to make classifications about the group a data point belongs to.
- Naive Bayes is a probabilistic classifier which uses Bayes theorem to assign class labels to instances.

Results

- Table 1. Summary of the predictive performance of each model

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>Sensitivity</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>0.66</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td>Decision tree</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td>Random forest</td>
<td>0.97</td>
<td>0.98</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>SVM</td>
<td>0.66</td>
<td>0.44</td>
<td>0.66</td>
<td>0.53</td>
<td>0.66</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.66</td>
<td>0.44</td>
<td>0.66</td>
<td>0.53</td>
<td>0.66</td>
</tr>
<tr>
<td>K-nearest neighbour</td>
<td>0.77</td>
<td>0.76</td>
<td>0.77</td>
<td>0.75</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Conclusion

The best-performing model was the Random Forest with an F1 score of 0.97. The 5 most contributing factors were fibroscan results, patients with a normal ultrasound scan, albumin levels, patients with HCC, and space-occupying lesions. We plan to study HB patients’ evolution and driving factors as well.

Future work

Aside from predicting the mortality outcome in hepatitis B patients and finding significant factors related to the outcome, we also want to predict HBV disease progression (cirrhosis, hepatocellular carcinoma (HCC), and end-stage liver disease (ESLD)). We want to identify the significant factors associated with the progression of CHB disease.

References