CLASSIFICATION RISKS OF HEPATITIS C USING PREDICTIVE MACHINE LEARNING ALGORITHMS

By: Gideon Ovuzorie (20033557)
Supervisor: Jeremy Harwood
School of Physics, Astronomy and Mathematics
University of Hertfordshire

Introduction

Hepatitis C is a liver infection caused by the hepatitis C virus (HCV). For some people, hepatitis C is a short-term illness, but for more than half of people who become infected with the hepatitis C virus, it becomes a long-term, chronic infection. While there are several studies to put an end to this nauseating menace and the mortality it leaves in its wake, early diagnosis is considered the best approach to finding a lasting cure. Liver biopsy has been the most prevalent mode of staging liver fibrosis in those who are infected with HCV until recently, when non-invasive models that leverage evolving imaging technologies, serum markers and machine learning models were proven to be better alternatives. The proper classification of the chronic HCV is an indispensable step in designing the most effective regimen to cure the HCV infection.

The Dataset

The dataset that will be used is made up of 1,741 HCV genotype 4 patients collected from an HVC Liver Fibrosis database available at the UCI machine learning repository. The dataset was acquired from El Demerdash Hospital, Egypt.

The Dataset contains 31 features which includes Age, Gender, BMI (Body Mass Index), Fever, Nausea, Headache, Diarrhea, Jaundice, WBC (White Blood Cell), RBC (Red Blood Cell) etc. The dataset will be thoroughly refined at the preprocessed stage and divided into train, test and validation. 70% of the observations would be used to train the model, while 30% would be reserved for testing

Methodology

The classification risk of HCV, will be achieved in three stages:

- Data Collection,
- Data Preprocessing, and
- Classification.

Existing Machine Learning Models used in predicting Liver Cirrhosis and Fibrosis includes:

- Decision Trees,
- Particle Swarm Optimization (PSO),
- Particle Swarm Optimized Gaussian Process Classifier, etc.

I will compare the results of these machine learning models and use the best or a combination of these or other models to implement a ML model specifically for classifying the risk of Hepatis C Virus.

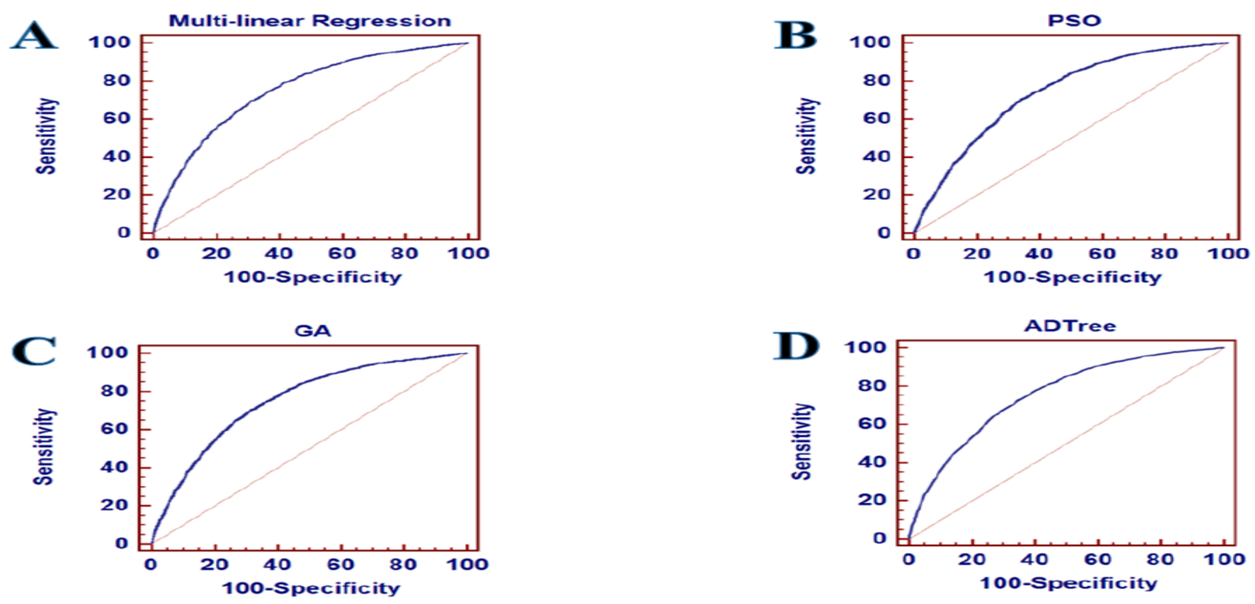
The Project

This project primarily aims to classify the risks of HCV viral infection in an infected body using the best machine learning model while equally predicting liver fibrosis in such patients. Various machine learning approaches have been adapted to the medical field to lead the course of noninvasive methods for predicting the infection or invasion of diseases, detecting the presence or absence of diseases, diagnosing such diseases, determining the best routine for treatment, monitoring response to treatment, predicting if a regimen should be discontinued of not intelligently and with unassailable accuracy.

For HCV in particular, various machine learning models have been developed to predict liver cirrhosis and fibrosis, giving numerical analysis on data analyzed in real-time. Some of them include Support Vector Machine (SVM), Particle Swarm Optimization (PSO), Artificial Neural Networks (ANN), Decision trees, Particle Swarm Optimized Gaussian Process Classifier, Genetic Algorithms, and so on.

These models have to varying degrees been able to foretell if a patient has liver fibrosis or not accurately, and this project aims to .implement Machine Learning Model that best classifies the risk of HCV using one or a combination of these models to archive the best result.

Preliminary Result and Analysis



Source: Hashem, S., Esmat, G., ...ElHefnawi, M., 2018. Comparison of Machine Learning Approaches for Prediction of Advance Liver Fibrosis in Chronic Hepatitis C Patients. IEEE/ACM Transaction on Computational Biology and Bioinformatics 15, 861-868. doi:10.1109/TCBB.2017.2690848

The figure above shows the Receiver Operating Characteristic (ROCs) curve values of four distinct ML techniques applied for predicting advanced liver fibrosis. The ROC curves and their results are:

(A) Multi-Linear Regression Model with a value of 0.76. This had the best result.

(B) Particle Swarm Model with a value of 0.73. The value is good, but not the best

(C) Genetic Algorithm Model with a value of 0.75. The value is good, but not the best

(D) Alternating Decision Tree Model with a value of 0.75.The value is good, but not the best

Receiver Operating Characteristic curve is a common tool used to evaluate the performance of a model. Apart from ROC curves, other tools for evaluating the performance for Machine Learning Models include accuracy, specialties, predictive values as well as sensitivity.

The Next Steps

Once the data have been properly refined at the preprocessing stage, classification of the data to identify inherent risks would be done. This could be accomplished using Particle Swarm Optimization, Particle Swarm Optimized Gaussian Process Classifier, Related Objects Skipper RCS algorithm, Decision trees, support vector machine, Subsumptions Rule Based Classifier, or a combination of these and other techniques.

References

CDC, 2020. Hepatitis C. April, 2020. Available at: https://www.cdc.gov/hepatitis/hcv/pdfs/hepcgeneralfactsheet.pdf (Accessed: 26 May 2022).

Hashem, S., Esmat, G., ...ElHefnawi, M., 2018. Comparison of Machine Learning Approaches for Prediction of Advance Liver Fibrosis in Chronic Hepatitis C Patients. IEEE/ACM Transaction on Computational Biology and Bioinformatics 15, 861-868. doi:10.1109/TCBB.2017.2690848

Badillo, S., Banfai, B., Birzele, F., Davydov, I. I., Hutchinson, L., Kam-Thong, T., and Zhang, J. D., 2020. An Introduction to Machine Learning. Clinical Pharmacology and Therapeutics, 107(4), 871–885.

University of Hertfordshire UH