HEPATITIS C DETECTION

USING

MACHINE LEARNING

Submitted By:

Livya Eldho

MAC23MCA-2036

Faculty Guide:

Prof. Sonia Abraham

Assistant Professor

MCA Department, MACE

INTRODUCTION

Hepatitis C is a blood-borne infection of the liver caused by the hepatitis C virus (HCV). Hepatitis C virus is a major cause for happening liver disease all over the world. In this work, a machine learning based model has been proposed that can classify hepatitis C virus infected patient's stages. I gathered the instances of hepatitis c disease of Egyptian patients from UCI machine learning repository. This study explores the effectiveness of machine learning techniques in the prediction of treatment response in hepatitis c patients.

The dataset is taken from the UCI machine learning repository. The dataset contains 1385 sample observations and has 29 columns including 1 class variable and 28 features.

The performance of machine learning algorithms such as Random Forest, ANN and Naïve Bayes are compared to identify the stages of liver fibrosis (1-4) due to hepatitis c and it is classified into no fibrosis/minimal damage, mild fibrosis, moderate fibrosis and cirrhosis.. The most significant features, which can be used to detect hepatitis c diseases more precisely are identified using Pearson Correlation, ANOVA, Random Forest, correlation matrix (library "pandas" and "seaborn"), Rank Feature Importance method with Extra Classifier Tree model (RFI-ECT), recursive feature elimination method with Logistic Regression model (RFE-LR), and SelectKBest (score_func=chi2, k=5) are addressed using Python-based scikit-learn package. The selected features are then used by the algorithms to build the models. Performance is evaluated and the best model is selected based on accuracy. Among the algorithms, Random Forest is found to be best in terms of computational time and accuracy score, which make it significant for the proposed approach.

A blood test is one of the ways to diagnose Hepatitis C disease. But after a lab blood test, a medical expert needs to examine the test stats to diagnose the disease. There is very little difference in the blood test stats, which refer to different hepatitis c stages. Such minor differences can lead to the wrong diagnosis even by medical experts as human error is expected. Incorrect diagnosis may lead to wrong medication and further complexities. So, an automated system can be very helpful to assist medical experts and even make automated disease predictions without any human mistakes.

LITERATURE SUMMARY

Paper 1 : Effective factors in diagnosing the degree of hepatitis c using machine learning

The paper “Effective factors in diagnosing the degree of hepatitis c using machine learning” explores the application machine learning techniques for detection and classification of hepatitis c disease using selective features.

In this study, a dataset was used which contains 27 features and 1385 records of patients with different grades of HCV, which can be categorized into three categories: Demographic, Symptom, Liver Function Tests (LFTs) and Blood Panels before the start of treatment. The dataset was clean and preprocessed to ensure accuracy and consistency. To reduce the dimension of the dataset and determine the effective features three feature selection, Pearson Correlation, ANOVA, and Random Forest, were applied. Among all the algorithms, KNN, random forests, and Deep Neural Networks were selected to be utilized, and then their evaluation metrics, such as Accuracy and Recall.

Various machine learning algorithms were applied to build predictive models. In all the modeling algorithms, Holdout was used to split up the dataset into "Train" and "Test" sets. The models were trained on the training set and evaluated on the testing set. This study utilizes a retrospective observational design which has been conducted to find the effective factors for diagnosing the degree HCV. This study showed that using data-mining algorithms can be helpful to for HCV diagnosing. The proposed model in this study can help physicians diagnose the degree of HCV at an affordable and with high accuracy.

Performance evaluation of these models based on accuracy showed that Deep Learning with Accuracy = 92.067 had the highest performance. Random Forest had almost the same performance as Deep Learning. This performance was achieved on dataset containing features that were selected by ANOVA feature selection.

Title of the paper	"Effective factors in diagnosing the degree of hepatitis c using machine learning" Mohammadjavad Sayadi, Elahe Gozali, Malihe Sadeghi.
Area of work	Detection and classification of hepatitis c disease using selective features.
Dataset	Dataset was taken from UCI repository. The dataset contains 1385 sample observations and each sample is represented by 29 features.
Methodology / Strategy	The dataset was clean and preprocessed to ensure accuracy and consistency. Three feature selection techniques were first used to select the most useful features for HVC analyzing. Then for each selected feature set, three machine-learning algorithms were applied, and a diagnosing model for detecting Hepatitis C was reported.
Algorithm	RF, KNN, DNN
Result/Accuracy	Results indicate that DNN and RF achieved the highest accuracy of 93.51 and 93.029 respectively. DNN – 93.51 RF – 93.029 KNN – 89.283

Paper 2 : Diagnosing the Stage of Hepatitis C Using Machine Learning

The paper " Diagnosing the Stage of Hepatitis C Using Machine Learning " aims to precision performance evaluation of the proposed Intelligent Hepatitis C Stage Diagnoses System (IHSDS) empowered with machine learning is presented to detect the Stage of Hepatitis C in a human using Artificial Neural Network (ANN).

In this research, the Hepatitis C patient dataset titled "HCV-Egy-Data" is obtained from the UCI machine learning repository. It includes 1385 observations, where each sample has 29 properties, out of which 19 properties are selected. The value of attribute "histological staging" in this dataset indicates the Stage of the patient. There are336(24.26%) cases in class 1, 332 (23.97%) cases in class 2, 355 (25.63%) cases in class 3, and 362 (26.14%) cases in class 4.

969 samples (70% of the dataset) are used for training the model, and the remaining 416 samples (30% of the dataset) are used for validation purposes. This research work predicts the Stage of Hepatitis C by using the back-propagation algorithm of the Artificial Neural Network model. This model is used to gain the maximum precision in the prediction of the Stage of Hepatitis C.

Different stages involved in the back-propagation algorithm include reading the training data, building and connecting the ANN layers (this includes preparing weights, biases, and activation function of each layer), predicting error, updating parameters, and prediction precision.

To evaluate the performance of the proposed model, precision, miss rate, and mean square error (MSE) are calculated. The proposed IHSDS was trained and validated using the dataset and shows 98.89% and 94.44% precision during training and validation phases, respectively. The validation precision value of previous methods is compared with the proposed model to state that the results provided by the proposed IHSDS empowered with machine learning are better than the results provided by other proposed methods(RF,SVM,LR) with regard to precision.

Title of the paper	"Diagnosing the Stage of Hepatitis C Using Machine Learning" |*Journal of Healthcare Engineering*, Muhammad Bilal Butt, Majed Alfayad, Shazia Saqib, M. A. Khan, Muhammad Adnan Khan ,and Nouh Sabri Elmitwally, Munir Ahmad | Volume 2021, Article ID 8062410
Area of work	Performance evaluation of the proposed Intelligent Hepatitis C Stage Diagnoses System (IHSDS) empowered with machine learning in detection of stages of Hepatitis C using ANN.
Dataset	Dataset was taken from UCI repository. The dataset contains 1385 sample observations and each sample is represented by 29 features.
Methodology / Strategy	This model is used to gain the maximum precision in the prediction by using the back-propagation algorithm of the Artificial Neural Network model. It includes reading the training data, building and connecting the ANN layers (this includes preparing weights, biases, and activation function of each layer), predicting error, updating parameters, and prediction precision.
Algorithm	ANN(back-propagation algorithm)
Result/Accuracy	The proposed IHSDS shows 98.89% and 94.44 % precision during training and validation phases

Paper 3 : Hepatitis C Virus (HCV) Prediction by Machine Learning Techniques

This study aims to know the performance comparisons between multi and binary class labels of the same dataset, not limited to tool comparison, and to know which selected features play a key role in the prediction of HCV by using Egyptian patient's dataset. The management of HCV infected patients can be monitored by the assessment of liver fibrosis staging in Chronic Hepatitis C (CHC).

In this paper, our objective is to identify the best model, the selected features which play a key role in the prediction of HCV disease, and to compare the performances of Python and R tools by using Hepatitis C Virus (HCV) from Egyptian patient's dataset from UCI in the content of multi and binary class labels. The multivariate datatype consists of 1385 instances with 29 attributes. The multiclass label (i.e., Baseline histological staging) dataset instances consist of distinct values and frequencies, which are as follows F1, F2, F3 and F4. The binary class label dataset consists of mild to moderate fibrosis as class label = 0 (F1 and F2) and advanced fibrosis as class label =1 (F3-F4).

A total of seven machine learning techniques followed by different feature selection methods were used to evaluate the performance of the classifiers on dataset. The analytical tools such as Python-based Scikit learn and R based CARET package were used to perform data analysis. In Spyder (Python IDE)-based scikit-learn package, before to classification model building, the pre-processing steps such as data normalization, total dataset split into 70–30% as a training and testing data respectively with set.seed was used.

In Python, the multiclass dataset shows the highest accuracy (28.36%) in a random forest, followed by KNN with 26.44%. The similar accuracy performance of KNN accuracy has been observed in NN, but when coming to precision and recall the KNN shows a better performance. Similarly, the binary dataset shows the highest accuracy (53.12%) in NN with the exemption to precision in comparison with SVM where the accuracy shows 52.64%. The individual performances of each classifier to accuracy, precision, and recall in binary class label almost shows the double score of multiclass label respectively.

On the other hand, the R multiclass dataset shows the highest accuracies with similar performances in SVM and RF (51.31%) followed by KNN (50.83%), and NB (50.65%). In the binary class label, boosting shows the highest accuracy (54.23%), followed by KNN (53.06%). Similar performance of accuracy also been noticed in NN and Bagging (51.73%) respectively. The accuracies of multiclass and binary class labels show similar performance, but whereas precision and recall show different performances. Within the comparison of 3 datasets features, the 12 selected features

overall average accuracy, precision, and recall shows similar performances to 29 and 21 selected features in both class labels and tools. Thus, indicating 12 selected features can be useful for the prediction of the HCV dataset. Despite class labels and tools used, average performances of evaluation metrics of the classifiers are in the order of SVM, RF, KNN, NN, and NB.

Title of the paper	Satish CR Nandipati, Chew XinYing, Khaw Khai Wah. ”Hepatitis C Virus (HCV) Prediction by Machine Learning Techniques”. *Applications of Modelling and Simulation* VOL 4, 2020, 89-100
Area of work	Performances of classifiers and tools on multi and binary class labels of the same HCV datasets.
Dataset	The data was taken from the UCI machine learning repository. Dataset contains 1385 instances with a total of 29 features.
Methodology / Strategy	The selected targeted dataset is imbalanced, so to make the dataset balanced, resampling techniques are used. Feature selection methods like univariate feature selection approach and the feature importance method is implemented and the best set of characteristics for building effective models are selected.
Algorithm	Random Forest Classifer, Naive Bayes Classifer, Adaboost, K-Nearest Neighbor, SupportVectorMachine, NN, Bagging
Result/Accuracy	Performance comparison of multi and binary class labels Despite class labels and tools used, average performances of evaluation metrics of the classifiers are in the order of SVM, RF, KNN, NN, and NB. 12 selected features can be useful for the prediction of the HCV dataset.

PROJECT PROPOSAL

The project aims to enhance the early detection of Hepatitis C using machine learning algorithms. From the above three papers, we get to know that different models were used for the detection of hepatitis c disease. First paper is the detection and classification of hepatitis disease using selective features. Second paper focuses on performance evaluation of the proposed Intelligent Hepatitis C Stage Diagnoses System (IHSDS) empowered with machine learning in detection of stages of Hepatitis C using ANN. Third paper aims at performances of classifiers and tools on multi and binary class labels of the same HCV datasets.

The proposed system is the comparative study of three algorithms Random Forest, ANN, SVM. The models will classify hepatitis c disease under four classes which are no fibrosis/minimal damage, mild fibrosis, moderate fibrosis and severe fibrosis/cirrhosis. Patients can diagnose their condition without the assistance of a medical expert.

DATASET

In this study, the dataset is taken from the UCI machine learning repository.

Feature name	Feature Category	Description	Data type
Age	Demographic	Age of patient	Numeric
Gender	Demographic	Gender of patient (1: Male, 2: Female)	Numeric (bool)
BMI	Symptom	Body Mass Index of patient	Numeric
Fever	Symptom	Presence of Fever (1: Absence, 2: Presence)	Numeric (bool)
Nausea/Vomiting	Symptom	Presence of Nausea/Vomiting (1: Absence, 2: Presence)	Numeric (bool)
Headache	Symptom	Presence of Headache (1: Absence, 2: Presence)	Numeric (bool)
Diarrhea	Symptom	Presence of Diarrhea (1: Absence, 2: Presence)	Numeric (bool)
Fatigue & generalized bone ache	Symptom	Presence of Fatigue & generalized bone ache (1: Absence, 2: Presence)	Numeric (bool)
Jaundice	Symptom	Presence of Jaundice (1: Absence, 2: Presence)	Numeric (bool)
Epigastric pain	Symptom	Presence of Epigastric pain (1: Absence, 2: Presence)	Numeric (bool)
WBC	Liver Function Tests	White Blood Cell count	Numeric
RBC	Liver Function Tests	Red Blood Cell count	Numeric
HGB	Liver Function Tests	Haemoglobin	Numeric
Plat	Liver Function Tests	Platelet count	Numeric
AST 1	Blood Panels	Aspartate Aminotransferase Enzyme Level at one week	Numeric
ALT 1	Blood Panels	Alanine Aminotransferase Enzyme Level at one week	Numeric
ALT4	Blood Panels	Alanine Aminotransferase Enzyme Level at four weeks	Numeric
ALT 12	Blood Panels	Alanine Aminotransferase Enzyme Level at 12 weeks	Numeric
ALT 24	Blood Panels	Alanine Aminotransferase Enzyme Level at 24 weeks	Numeric
ALT4 36	Blood Panels	Alanine Aminotransferase Enzyme Level at 36 weeks	Numeric
ALT 48	Blood Panels	Alanine Aminotransferase Enzyme Level at 48 weeks	Numeric
ALT after 24 w	Blood Panels	Alanine Aminotransferase Enzyme Level after 24 weeks	Numeric
RNA Base	Blood Panels	RNA at the start of treatment	Numeric
RNA 4	Blood Panels	RNA at four weeks	Numeric
RNA 12	Blood Panels	RNA at 12 weeks	Numeric
RNA EOT	Blood Panels	RNA at the End of Treatment	Numeric
RNA EF	Blood Panels	RNA Elongation Factor	Numeric
Baseline histological Grading	Target Column	Baseline histological Grading (0-13)	Numeric

The dataset contains 29 attributes including 1 class variable and 28 features and has 1385 records of patients with different grades of HCV, which can be categorized into different categories before the start of treatment such as:

(1) Demographic

(2) Symptom

(3) Liver Function Tests (LFTs)

(4) Blood Panels

The features are age, gender, BMI, fever, nausea, headache, diarrhea, fatigue & generalized boneache, jaundice, epigastricpain, WBC, RBC, HGB, Plat, AST1, ALT1, ALT4, ALT12, ALT24, ALT36, ALT48, ALT, RNA Base, RNA 4, RNA12, RNA EOT, RNA EF, Baseline histological Grading and Baseline histological staging.

Class variable : Baseline histological staging

The class variable include numbers from 1 to 4 which indicates different stages in hepatitis c disease which are :

1. no fibrosis/minimal damage
2. mild fibrosis
3. moderate fibrosis
4. severe fibrosis/cirrhosis.

Dataset:https://archive.ics.uci.edu/dataset/503/hepatitis+c+virus+hcv+for+egyptian+patients