International Journal of Computational Science and Engineering Research
ISSN: XXXX- XXXX(Online) , http://www.ijcser.com/
Regular Issue , Vol. 2, Issue. 1 (January - March), 2025 , Pages: 1 - 6
Received: 05 January 2025 ; Accepted: 20 February 2025 ; Published: 25 February 2025.
Research Paper , https://doi.org/10.XXX/XXXX.XXXX.XX.XXXX

Logistic Regression based Hepatocellular Carcinoma Liver Cancer Detection

Kuppireddy Krishna Reddy

Department of EEE, Mother Theresa Institute Of Engineering and Technology (A),Palamaner, JNTUA, AP, India ; krishnareddy206@mtieat.org

Abstract: One of the most typical malignancies of the liver is called hepatocellular carcinoma. This is a very dangerous illness that can even be fatal. Surgical removal of the tumor or a liver transplant may be effective treatments for hepatocellular carcinoma. In this article, we cover the topic of Hepatocellular carcinoma (HCC) liver cancer prediction models. We have suggested a simple method for predicting HCC liver cancer from a freely available dataset. In order to enhance the quality of the dataset, we are applying several pre-processing methods. We are using the Statistical model which is Logistic Regression for the classification, to predict the HCC Liver cancer disease at the early stage.

Keywords: Liver, Neural Network, Cancer , Machine Learning , HCC, DNA.

1. Introduction

The liver cancer known as hepatocellular carcinoma (HCC) is one of the most common forms of cancer. It's a diverse collection of tumours with a wide range of potential causes and outcomes. It can be caused by a number of different things, such as viruses, chemicals, and both inherited and acquired metabolic disorders. There is strong evidence linking HCC to hepatitis B virus and, possibly in other parts of the world, hepatitis C virus. There are many important effects on cell survival, growth, transformation, and maintenance that are triggered by viral factor-induced liver injury. These processes include cell signalling, apoptosis, transcription, and DNA repair. People with chronic liver diseases (Figure.1)are at a greater risk of developing hepatocellular carcinoma, the most common form of liver cancer. Scarring of the liver from either hepatitis B or hepatitis C infection also increases the risk.

Hepatocellular carcinoma occurs more frequently in heavy alcoholics and those with liver fat accumulation. Loss of appetite, yellowing of the skin and eyes (jaundice), Fever, general weakness and bloating of stomach and throwing up are all signs of this illness. Surgery or a transplant may be successful treatments if caught in time[1]. Even though there is no cure for advanced stages, treatment and support can extend and improve quality of life. Keep in mind that you are the one who ultimately decides how you will be treated and how you will live your life. Recent clinical studies have shown, for instance, that patients infected with hepatitis C virus (HCV) have a risk of developing chronic hepatitis, cirrhosis, and HCC, all of which can lead to the eventual development of end-stage liver disease and death. Consequently, early detection of HCC and the administration of appropriate treatment are of the utmost importance. This project will use machine learning methods on actual laboratory data to investigate potential causes of HCC beyond HCV, such as fatty liver.

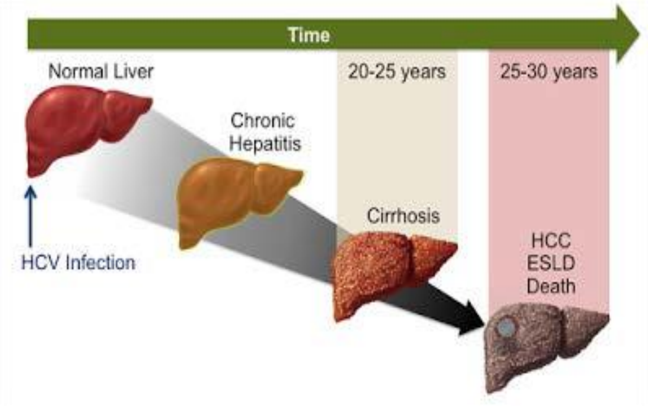

Figure.1: HCV infection progresses to HCC

HCC Liver Cancer occurs when a person is suffering from hepatitis B or hepatitis C. It leads to develop the cancer tumour in the liver.

The causes of this disease are excessive consumption of alcohol, subsequent use of tobacco, exposure to toxic chemicals leads to the damage of liver. Due to delay in recognizing the disease in human body, sometimes it leads to death [2]. By Using our model which consists of more features (Figure.2) and dimensionality reduction technique will give accurate result in which the disease can be detected at early stages.

2. Problem Analysis

2.1 Existing System

- The current system makes use of a dataset from the UCI machine learning repository known as HCC. This dataset is also available for use on Kaggle under the same name.
- The disease HCC is to be predicted on the basis on three types of attributes Nominal, Integer, Continuous and Ordinal. Gender, symptoms, alcohol consumption, hepatitis B virus infection, and surface antigen status are some of the features [3].
- They have used five machine learning approaches like KNN, SVM, Random Forest, Decision Tree and Navi Bayes classifiers to get highest accuracy.
- As a form of classification measure, they have employed the use of confusion matrices. By far the most accurate classifier they tested was the Random Forest algorithm, with an accuracy of 80.64 percent.

2.2 Challenges Faced

- Much Noisy and Irrelevant data in the process of data analysis.
- The data results acquired are less accurate.
- The SVM classifier uses the most time in computation, which directly correlates to a high demand on both the computer's processor and memory.
- Accuracy is less
- Specifically, they employed a two-classification-algorithm strategy (i.e. K-Nearest Neighbor and Naive Bayes.

2.3 Proposed System

In this proposed System we predict the patient's survival or death from Hepatocellular Carcinoma (HCC) using the data included as features of the dataset. Specifically, we used the HCC Dataset, a dataset hosted on Kaggle that is openly available for experimentation. As recommended by the European Association for the Study of the Liver and the European Organization for Research and Treatment of Cancer, the dataset includes 49 features (EASL-EORTC). Logistic Regression was chosen because it is simple to implement and interpret, and produces good training results with little effort. It quickly and accurately categorises new data. Its use enhances precision. We have used a confusion matrix as the disease classification measure matrix to establish some metrics that will aid in the prediction of the HCC disease.

2.3.1 Advantages

The Accuracy is increased

- This is a much faster approach
- It derives confidence level (about its prediction).
- It makes sure final model is parsimonious and balanced.

3. Proposed System Model

The architecture of a system is its conceptual model, which includes the system's structure, behavior, and other perspectives. System components and developed sub-systems that cooperate to implement the entire system might be part of a system's architecture(Figure.2). Formal documentation outlining the specifics of a system or its individual parts in order to facilitate construction. Objects in a system architecture diagram are represented by symbols like circles, squares, and arrows.

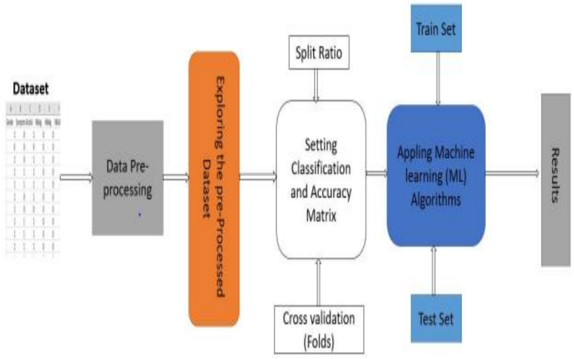
Dataset
Data Pre-processing
Exploring the pre-Processed Dataset
Split Ratio
Setting Classification and Accuracy Matrix
Cross validation (Folds)
Train Set
Appling Machine learning (ML) Algorithms
Test Set
Results

Figure.2: Architecture of Proposed System

a) Data Pre-processing

Note that the flaws and dispersed data in the aforementioned dataset are not exaggerations. We have pre-processed the dataset to make it usable and to extract the information from it. As a means of dealing with inaccurate data, we have examined the dataset to identify outliers and rectified them by hand. To handle missing values, we find the median of that characteristic and fill in the blanks with that number.

Pandas [5] and NumPy [6] library were used to manage the dataset's dispersed nature and facilitate easy data processing throughout the experiment, respectively[4].

b) Setting Classification and Accuracy Matrices

Metrics that aid in the prediction of HCC disease need to be established before the disease can be classified. Considering that we are conducting this experiment with the scikit-learn machine learning package [7], we have chosen to use the confusion matrix as our classification measure matrix. Below is a complete breakdown of all the metrics we utilized in our experiment, including Precision, Recall, F1-Score, and Accuracy.

ACCURACY: One way to rank classification algorithms is by how well they perform. Accuracy can be defined informally as the percentage of correct predictions made by our model.

The following equation describes accuracy formally:

$$\text{Accuracy} = \frac{\mathbf{TP+TN}}{\mathbf{TP+TN+FP+FN}}$$

Where, TP=True Positive, TN=True Negative, FP=False Positive, FN=False Negative

PRECISION: To measure how well a machine learning model does its job, we can look at how accurate its positive predictions are. The term "precision" is used to describe the ratio of "real" successes to "total" successes in making forecasts (i.e., the number of true positives plus the number of false positives).

Simply Precision is What partition of positive identification was actually correct.

$$\text{Precision} = \frac{\mathbf{TP}}{\mathbf{TP+FP}}$$

RECALL: The Recall is the proportion of true positives over the total number of samples that were correctly labeled as positive. The proportion of positive samples correctly identified by the model is what we call its recall. The greater the recall, the greater the number of true positives identified. Simply, Recall is defined as what partition of actual positives was identified correctly.

$$\text{Recall} = \frac{\mathbf{TP}}{\mathbf{TP+FN}}$$

F1-SCORE: The F1 score is composed of two components: Precession and Recall. F1 is an attempt to merge the precision and recall measurements into a single indicator of performance.

$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2\frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

Confusion Matrix

A confusion matrix, also known as an error matrix, is a special table structure that permits visualization of the performance of an algorithm, often a supervised learning one, in the subject of machine learning and more specifically the problem of statistical classification (in unsupervised learning it is usually called a matching matrix). There are two possible interpretations of this literature finding: either each row of the matrix represents the occurrences in an actual class(Figure.3) and each column represents the instances in a predicted class, or vice versa [5]. Because it's so simple to tell if the programmer is mixing up different types of data, we call it a "class confusion" indicator (i.e., commonly mislabeling one as another).

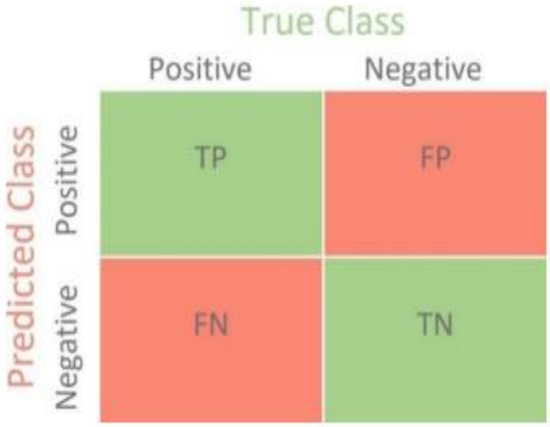

Figure.3: Confusion Matrix

A Confusion matrix will contain 4 fields.

True Positive (TP): True positive mean when the record is actually a positive record and our model also predicted the record as a positive record then it is what we call a "positive" result.

False Positive (FP): False positive mean when a record is actually a negative record and our model predicted the record as a negative record then it is known as false positive.

False Negative (FN): False negative mean when a record is actually a positive record and our model predicted the record as a negative record then it is known as false negative.

True Negative (TN): True negative mean when a record is actually a negative record and our model predicted the record as a negative record then it is known as true negative [6].

4 . Experimental Results

Using the KNN Algorithm, we were able to get an accuracy of 93%, and using the Logistic Regression Algorithm, we were able to get an accuracy of 98%. To see how the dataset performs on various classifiers and to obtain a more nuanced set of results, we have selected a number of different classifiers(Figure.4). In Table II, we display the values of the various performance metrics we've measured for each of the classifiers we've employed[7]. Now, with the aid of scikit-accuracy learns score, we were able to calculate the accuracy of each classifier without employing cross-validation in our experiment.

Classifier	TP	FP	FN	TN	Precision	Recall	F1-Score
KNN	27	6	7	22	0.79	0.82	0.81
SVM	26	7	14	15	0.65	0.79	0.71
NB	5	28	3	26	0.62	0.15	0.24
RF	28	5	7	22	0.80	0.85	0.82
Logistic Regression	9	1	0	22	0.96	0.985	0.99

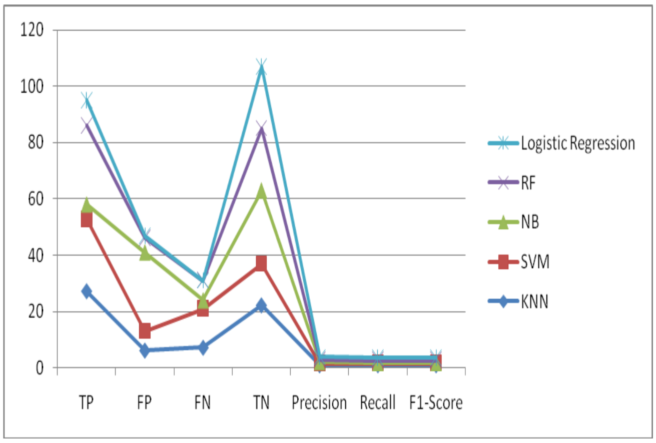

Figure.4: Accuracy graph without cross-validation

Logistic Regression 0.98375

Logistic	Precision	Recall	F1-Score	Support
0	0.96	0.985	0.99	10
1	0.97	0.981	1.00	22
Accuracy			0.98	32
Macro avg	0.96	1.00	0.978	32
Weighted avg	0.971	0.99	0.971	32

Figure.2: Precision, recall, f1-score, support for values for Logistic Regression

6. Conclusion and Future Scope

In this paper, we explore the topic of a model for predicting the development of Hepatocellular Carcinoma (HCC), a particularly aggressive form of liver cancer. We suggested a simple and effective method for predicting liver cancer. For this categorization, we employed the supervised machine learning methods of KNN and logistic regression. In our project by using the Logistic Regression we have achieved the maximum accuracy.

In future research, other machine learning technique such as artificial neural network can be applied with a relatively large data set.

References

[1]. Sanapala Rajesh, Nurul Amin Choudhury, Soumen Moulik have worked on a journal titled “Hepatocellular Carcinoma (HCC) Liver Cancer prediction using Machine Learning Algorithms”, published in IEEE in the year 2021.

[2]. Qiyao Wang, etc have worked on a journal titled "Malignancy characterization of hepatocellular carcinoma using hybrid texture and deep features" at Key Laboratory for Health Informatics, Chinese Academy of Sciences, Shenzhen, China. This is published in 2017 IEEE International Conference on Image Processing (ICIP).

[3]. Yu-Siang Lin, Pei-Hsin Huang, Yung-Yaw Chen have worked on a journal titled "Deep Learning-Based Hepatocellular Carcinoma Histopathology Image Classification: Accuracy Versus Training Dataset Size" at National Taiwan University, Taipei, Taiwan and published in IEEE on 22 February 2021.

[4]. VeleryVirgina Putri Wibowo, Zuherman Rustam, Sri Hartini, QisthinaSyifa Setiawan and Jane Eva Aurelia have worked on a journal titled "Comparison between Support Vector Machine and Random Forest for Hepatocellular Carcinoma (HCC) Classification" at University of Indonesia, Depok, Indonesia and published in IEEE on 15 January 2021.

[5]. Delia Mitrea, Sergiu Nedevschi, Paulina Mitrea, Monica PlatonLupşor and Radu Badea have worked on a journal titled "HCC Recognition Within Ultrasound Images Employing Advanced Textural Features with Deep Learning Techniques" at Technical University of Cluj-Napoca, Cluj-Napoca, Romania and published in IEEE on 23 January 2020.

[6]. S. Singh and D. Hanchate, “Improving disease prediction by machine learning,” 06 ,2018.

[7]. D. Shetty, K. Rit, S. Shaikh, and N. Patil, “Diabetes disease prediction using data mining,” in 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017.