Prediction of Risk Factor of the Patient with HepatoCellular Carcinoma using Machine Learning

Devi. E[1], Lavanya. R[2], Malar. S[3], M. Revathi, B.Tech, M.E[4]

[1,2,3]UG Student, Department of CSE, Agni College of Technology Chennai, India
[4]Assistant Professor, Department of CSE, Agni College of Technology Chennai, India

---***---

Abstract - Hepatocellular Carcinoma mostly affects people who are having liver disease. It causes because of heavy drinking. Having more than two alcoholic drinks a day for many years' raises the risk of getting disease. We collected the Hepatocellular Carcinoma affected patient's medical test report details. Then, we have come up with a system which can predict the risk factor based on the patients test results. The risk is classified into two types. They are high risk and low risk. The patients who are predicted to have high risk, the probability of their lifetime is below one year and the patient with low risk will have the lifetime of more than one year.

I. INTRODUCTION

Machine Learning is a technology that builds intelligence system. These systems have ability to learn for experience and provide results according to it. Machine Learning came up to reduce human work. A task can be done without human involvement. A large area of artificial intelligence is machine learning. This enables systems to introduce new knowledge from experience. The system can learn things without being explicitly programmed.

Liver cancer is a life threatening malignant disease, and the number of new cases of liver cancer increased by 75% between 1990s and 2015 according to the Global Burden of Disease study (GBD 2015). In 2015,854000 new cases of liver cancer and 810000 deaths were reported worldwide, making liver cancer the fourth leading cause of cancer-related death, amounting to disease burden of 20578000 disability-adjusted life-years Hepatocellular carcinoma(HCC) accounts for 75%-80% of all cases of liver cancer. The five years overall survival of the HCC patients is 3%-5% across all countries. Patients with stage A HCC (BCLC) have a 5-year OS rate of 0%-75%, with different co morbidities 2.

Hepatocellular Carcinoma (HCC) is a liver disease and it is nearly binded with abnormal DNA methylation process.In this project, we analyzed 450K methylaion chip datas from 377 HCC samples and 50 adjacent normal samples in the TCGA database. In this project we collected and tested 47,000 differentially methylated sites using Cox regression technique as well as SVM-RFE and Few-SVM algorithms, and develop a model using two risk categories to predict the overall survival of the patients based on the 134 methylation sites. The model shows a 10-fold cross- validation score of 0.93 and satisfactory predictive power and SVM-RFE (Support Vector Machine-Recursive Feature Elimination) algorithm was used o predict the risk categories. They are correctly classified into 26 out of 33 samples in testing set obtained by stratified sampling from high intermediate and low risk groups.

In our system, we will predict the risk factor of the HCC affected patients using Machine Learning. Logistic classification is used to classify the processed dataset. Pre- processing and correct replacement of missing value with different attributes type will improve the model.

To achieve the goal, Data Engineering is the first step. Data Engineering consists of two process, they are Data Collection and Data pre-processing. Data Collection will be collected with meaningful parameters like blood, age, test and so on. First we convert the raw data into a clean data set using data preprocessing technique for achieving better results in machine learning projects. Collected data will be pre-processed which means encoding the categorical information in the data. The format of the data should be in a correct order. It is used to dropping unwanted parameters, scaling the parameters values to achieve normal distribution (Zero mean and standard deviation as one) handling missing values and so on.

After the Data Engineering process, Feature Engineering will be done. Feature Engineering is an important step to predict our output. The advantage of Feature Engineering is minimizing the parameter. For example, if our whole dataset contains 10 parameters, after feature engineering only three parameters enough to predict the output with high efficiency. Feature Engineering based on correlation, co-variance, co-linearity and etc. Feature Engineering has many algorithms to predict correct correlated parameters.

After the Feature engineering process, with selected features several Machine learning algorithm will be tested. Machine Learning algorithms are Support Vector Machine, K-NN, Random Forest, Decision Tree and logistic. Accuracy of the model calculated by the confusion matrix. The final model will be optimized by selecting the best accuracy of the algorithm. This model helps us to find the future prediction of survival of liver cancer affected patients.

The trained mode will be saved and loaded for web development. With the help of build model and with a selected feature. Web development will have an input variable of selected features, by submitting the answer of the selected feature, the prediction will be done.

II. LITERATURE SURVEY

First, we have investigated various papers and discussions on Machine Learning for Liver Cancer.[1]The title of the paper is Hepatocellular Carcinoma current trends in worldwide epidemiology, risk factors, diagnosis and therapeutics, In this paper they described Liver transplantation and surgical resection remains the corner stone of curative treatment. But major advances in loco regional therapies and molecular –targeted therapies for the treatment of advanced HCC have occurred recently.[2]The title of the paper is Computed aided diagnosis system developed for ultrasound diagnosis of liver lesions using deep learning, they described the accuracy of this 2-class classification CADx was 94.8%, the sensitivity was 93.8%, and he specificity was 95.2%. Both 4-class classification and

2-class classification CADx had relatively high accuracy. However in this study they used only a small amount data collected from a single facility.[3] Comparison of the Machine Learning approaches for prediction of advanced liver fibrosis in chronic hepatitis C patients, they described about age, platelet count ,AST and albumin were found to be statistically significant to advanced fibrosis.[4]Developing an expert system for diagnosing liver disease ,they described about 3 categories of disease diagnostic factors, questions corresponding sub section, and the response range are determined on mocklar chart, and finally the tables are designed to express the system interference according to users answers to the questions.

III. METHODS AND ALGORITHMS

A. DATA PREPROCESSING

Data preprocessing technique is used to convert the raw data into clean data. Raw data is nothing but, the data was collected from various sources and that format was not feasible for analysis process. In our dataset they are 50 attributes are present out of that 8 patients details are fully filed. In preprocessing technique we fill the missing values of attributes. It can be predicted by remaining by output variables. The data can be classified into three categories namely nominal, categorical and ordinal. Nominal information was named information that can be isolated into discrete classifications which don’t cover. Example of nominal data is gender, hair color, and eye color. Categorical data is a numerical data, Example of categorical is age. Ordinal data was a data that can be placed in an order or scale. Using preprocessing technique we filled all kinds of data that type has any missing values.

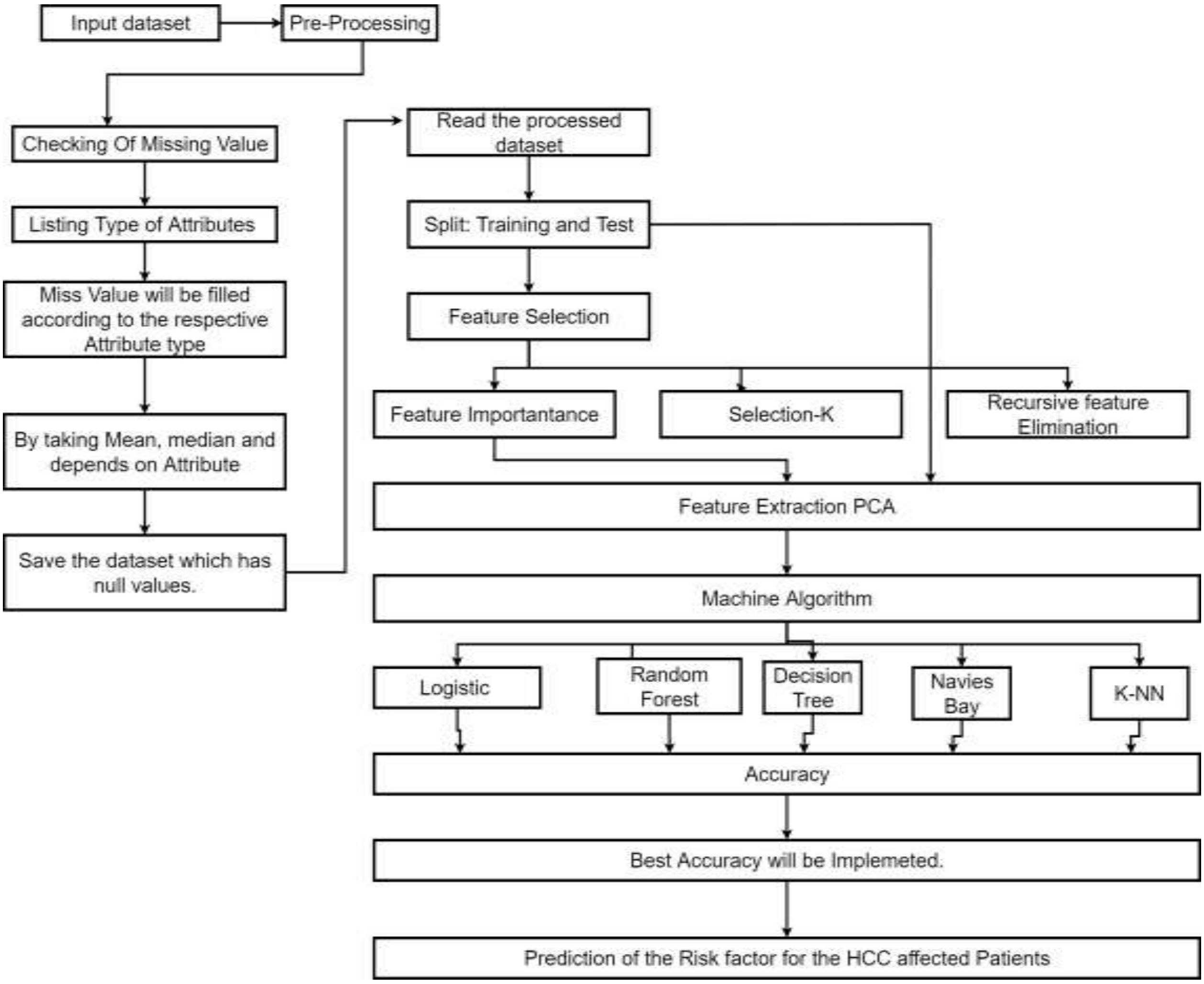
B. FEATURE SELECTION

Feature selection is the method of decreasing the number of input variables when developing a predictive model. It is difficult to analyze every attribute so feature selection select the most important attribute for prediction. It is also increase the performance of model. In feature selection we use Select-K Best algorithm, it was simply retains the first K features of X with highest scope. And also we used RFE (Recursive Feature Selection) algorithm to remove the less importance features.RFE plays a main role in prediction model.

C.FEATURE EXTRACTION

Feature Extraction is the process of reduce the number of features and also it create a new feature from existing one. It also extracts the new patterns available in features. In feature extraction we use PCA (Principal Component Analysis), it combines the same attributes and also create a new patterns which is superior to original attributes. It does not combine the attribute it just evaluates the quality, predictive power and selects the best one for model.

SYSTEM ARCHITECTURE:


Input dataset
Pre-Processing
Checking Of Missing Value
Listing Type of Attributes
Miss Value will be filled according to the respective Attribute type
By taking Mean, median and depends on Attribute
Save the dataset which has null values.
Read the processed dataset
Split: Training and Test
Feature Selection
Feature Importantance
Selection-K
Recursive feature Elimination
Feature Extraction PCA
Machine Algorithm
Logistic
Random Forest
Decision Tree
Navies Bay
K-NN
Accuracy
Best Accuracy will be Implemeted.
Prediction of the Risk factor for the HCC affected Patients

D. MACHINE LEARNING

In machine learning algorithms we use several algorithms like Random Forest, Decision Tree, KNN, SVM and Logistic Classification. Applied all these algorithms and find the accuracy of the model. Random Forest technique was learning strategy for order, relapse that worked by developing a choice tree on preparing time and outputting the class which the model of the classes or mean. In this algorithm it randomly select K features from total n features and also calculate the decision node and daughter node using best split, this process is looped until it reach the result.

Decision tree is a flowchart structure in that node represents the test on a attribute and each branch represent the outcome of the test. It only contains conditional and control statements. This algorithm split the datasets into smaller subset at the same time a associated decision tree incremently developed. Next algorithm is K-nearest algorithm; it is a non-parametric method for classification and regression. This algorithm stores all available cases and also classifies new cases on similarity measures. Support Vector Machine is a supervised learning model with associated learning algorithms that analyze the data. It is simply the co-ordinates of individual observations. Logistic regression is predictive analysis algorithm and based on the concept of probability.

IV. RESULT

Feature selection using select-K with chi-square parameters which have selected the highly correlated top five features from 50 feature input. The selected features are blood glucose random level, blood urea, serum creatine, packed cell volume and white blood count. We can able to predict that risk factor of a patient with selected feature values.

Machine Learning Algorithm with selected featured, various Machine Learning Algorithms were tested. The algorithms are Support Vector Machine, Random Forest and Na�ve Bay, Logistic, Decision Tree and K-NN algorithm. The highest accuracy obtained for the selected feature and support vector machine algorithm. The accuracy achieved was 95% for 5 feature input values.

	Select K	Rfe	select Pca	rfepca
SVM	0.95	0.79	0.73	0.80
Random forest	0.71	0.69	0.64	0.67
Logistic	0.59	0.65	0.59	0.56
Decision Tree	0.66	0.75	0.66	0.67
Navies bay	0.69	0.80	0.71	0.57
KNN	0.66	0.56	0.66	0.82

Web Development, the proposed system was deployed using Django with selected feature value as input. This Web Development takes-time real prediction of the risk factor.

V. CONCLUSION

The HCC affected person's risk factor was classified with Support Vector Machine. This was achieved with feature selection method select –K parameter with chi- square. The effective five features were selected from 50 features using feature selection method. The result achieved was 95% accuracy. The trained model SVM for 5 features input are able to predict the low risk or high risk. Advantage of using feature selection has eliminated the unwanted feature which may increase the blood test cost of the person.

VI. REFERENCES

1. Bird A.DNA methylation patterns and epigenetic memory. Genes Dev.2002; 16:6-21.
2. Dhanasekaran R, Limaya A, Cabrera R. Hepatocellular carcinoma: current trends in worldwide epidemiology, risk factors, diagnosis, and therapeutics. Hepat Med. 2012; 4:19.
3. Mizuno Y, Meamura K, Tanaka Y, et al. Expression of delta-like 3 is down regulated by aberrant DNA methlylation and histone modification in hepatocellular carcinoma. OncolRep. 2018; 39:220-2216.
4. Zhang Y, Petropoulos S, Liu J, et al. The signature of liver cancer in immune cells DNA methylation. Clin Epigenetics. 2018; 10:8.
5. Tsukuma H, Hiyama T, Tanaka S, et al. Risk factors for hepatocellular carcinoma among patients with chronic liver disease. N Engl J Med. 1993; 328(25): 1797-1801.
6. Yoshizawa H. Hepaticellular carcinoma associated with hepatitis C virus infection in Japan: Projection to other countries in the foreseeable future. Oncology 2002; 62 Supple 1:8-17.
7. Chen JD, Yang HI, Hoeje UH, et al. Carries of inactive hepatitis B virus are still at risk for hepatocellular carcinoma and liver-related death. Gastroenterology. 2010; 138(5): 1747-1754.
8. Nishida N, Nagasaka T, Nishimura T, Ikai I, Boland CR, et al.(2008) Aberrant methylation of multiple tumor suppressor genes in aging liver, chronic hepatitis, and hepatocellular carcinoma . Hepatology 47:908-918.
9. Feng Q, Stern JE, Haws SE, Lu H, Jiang M, et al.(2010) DNA methylation changes in normal sliver tissues and hepatocellular carcinoma with different viral infection. Exp Mol Pathol 88: 287- 292.
10. Ishak KG, Sobin LH (1994) Histological typing of tumors in the liver (International histological classification of tumors 2nd ed). Berlin: Springer- Verlag.
11. Yeh CC, Goyal A, Shen J, et al. Global Level of plasma DNA Methylation is Associated with Overall Survival in Patients with Hepatocellular Carcinoma. Ann SurgOncol. 2017; 24:3788-3795.
12. Xu R, Wei W, Krawczyk M, et al. Circulating tumor DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. Nat Mater. 2017; 16:1155.