A Survey of the Various Machine Learning In Liver Disease Prediction

[1]Pardeep Kaur, [2]Dr. Rakesh Kumar ,[3]Ms.Harinder Kaur
[1]M.Tech (Scholar), Department of Computer Science and Engineering
Sachdeva Engineering College for Girls , Gharuan , Mohali (Punjab)
[2]Principal, Department of Computer Science and Engineering
Sachdeva Engineering College for Girls , Gharuan , Mohali (Punjab)
[3]Assistant Professor, Department of Computer Science and Engineering
Sachdeva Engineering College for Girls , Gharuan , Mohali (Punjab)
([1]Pari.billing@gmail.com,[3]Hdeep1829@gmail.com)

Abstract- The liver is a large organ in the human body. In simple terms, the liver is segmented into two different glands. The first one is the secretory gland that has a dedicated structure to make it capable to send a secrete bile in the bile duct. The second endocrine gland creates the chemical which is directly flow in the blood and helped to digestion of food and absorption of fat in a human body. The position of liver is just below the diaphragm that is a muscle which partioned the chest from the abdomen in body. It played out some major functions. Liver disease is considered as a major disease that affect the million of people in this world. The common signs of disease are jaundice, dark urine color, vomiting, loss of appetite and itchy skin. The causes are infection, abnormality in immune system, genetics, cancer and other growths. The detection and diagnosis of liver disease are performed by the use of several methods such as blood tests, MRI, etc. Machine learning is an emerging field which gained a lot of attention in medical field. In this research work, various methods based on machine learning are obtained to detect and diagnose the liver disease patients. In survey, Naïve Bayes classifier (NB Classifier), Hidden Markov Model (HMM), Linear Discriminant Analysis (LDA)and C5.0 boost up algorithms are briefly studied and described. From deep research, the best technique is C5.0 Boostup which is an optimization algorithm due to the vast advantages that overweigh the features of other techniques.

Keywords- *CAD (Computer Aided Diagnosis); NB Classifier (Naïve Bayes Classifier); HMM (Hidden Markov Model); LDA (Linear Discriminant Analysis).*

I. INTRODUCTION

The liver is the most essential organ in the human body. The weight of a liver organ is around 3 pounds. The liver composed of two fundamental portions known as left and right projection. The major function of the liver is to uncontaminated the harmful substances in the blood and it worked with other organs to process and ingest. If the liver is not healthy then, it affects the other organs in the body. Therefore, the survival in this condition becomes harder. The liver is observed as a reddish brown gland that performed many functions in the human body. Nowadays, the liver disease becomes a common disease in all over the world. There are some basic causes that damaged the liver. The common causes are like fatty liver, liver fibrosis, cirrhosis and infections. In the liver disease it is hard to search out each issue normally therefore, doctors and experts are facing many difficulties in the diagnosis of liver disease [1].

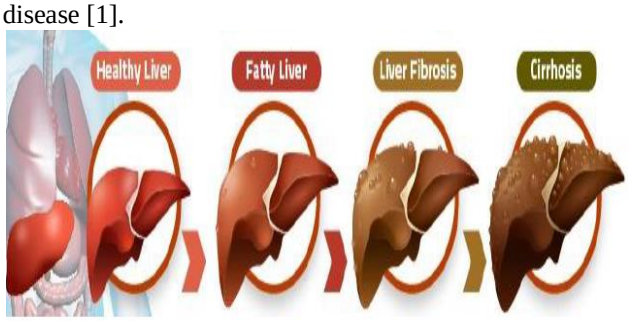

Fig.1: Stages of liver disease [1].

The detection of liver disease at an early stage is difficult because the tissues of liver expelled slowly. It is essential to diagnose the problem of liver to save the life of patients. The increase of chronic liver disease is described by various stages that has pathological properties. The fatty liver is considered as the initial stage of liver disease that is occurring due to the presence of large fat tissues in the body. The fibrosis stage occurred in the situation when damage of organs and injuries are happened. The progression of liver fibrosis is dependent upon the causes of diseases such as chronic hepatitis [2]. The common symptoms of liver disease are fatigue, anorexia, bad taste and smell, vomiting, body weight gain and loss, abdominal pain, jaundice, change in the bowel function, EDEMA and ascites, fever, muscle weaknesses, sleep disturbance, etc. [3].

For the diagnosis and predictions of liver disease, several techniques and algorithms are accessed to find out the disease causes in the early stages. The medical images are useful tools to detect any kind of disease. The computer aided diagnosis (CAD) is one of the most preferred technique which helped to the radiologists to interpret the images effectively and to search out the incorrect interpretations. It offers the cross-section images and highly accurate results are formed due to the increased signals of

noise ratio (SNR) [4]. The detection of liver disease is performed using back propagation neural networks, machine learning, ANN (Artificial neural networks), SVM (Support vector Machines) and so many other approaches which are described the better results for the diagnosis and detection of liver disease at the early stages. All the above mention techniques and algorithms are accessed by the CAD for the detection and treatment of liver disease. The CAD systems are the segmentation of an image, capture the unique features and explained the disease issues with the use of a classifier [5].

The remaining part of the research paper is partitioned in different sections. Section I is all about the basic introduction of the liver disease. Section II described the previous work related to the diagnosis and predictions of liver disease. It will help to understand the concepts briefly. In section III, the datasets are explained and Section IV deeply describes the used techniques. Section V included the final summary of the research and described the future scope of new techniques.

II. LITERATURE SURVEY

Hassoon, M., et al., (2017) [6] proposed a deep research on rule optimization of C5.0 classification. The proposed approach was genetic algorithm specifically designed for the predictions of liver disease. Currently, the interesting and crucial field of medical and science was diagnosis of illness by the use of characteristics that influenced the recognition process. The new term Medical Data Mining (MDM) was introduced which was composed of clustering and classification to diagnosis of the diseases. The primitive approach helped the researchers and doctors to find out the disease symptoms and to prevent patients from deaths. The outcomes of research increased the diagnosis time and accuracy. In this research work, genetic algorithm expelled the rules of other algorithms particularity to enhance the performance and throughput. The comparison demonstrated that the proposed technique was better than the existing ones mainly in terms of accuracy from 81% to 93%.

Saba, L., et al., (2016) [7] worked on an automated stratification of liver disease in ultrasound images. Fatty liver disease is one of the common diseases of liver. Mostly ultrasound was preferable for FLD (Fatty Liver Disease) because of low cost, non-radiation and easy process. The new approach was successful to detect and classify the fatty liver disease and it was benchmarked opposite to existing technique. The comparison created an enhanced set of features in the Liebenberg-Marquardt back propagation neural networks.

Abdar, M., et al., (2017) [8] described a performance analysis of classification techniques for the early detection of liver disease. Basically, a human liver was chief organ in the body and if it worked not well then several kinds of problems generated in the human life. The early detection and treatment of liver disease was predicted the disease and stop its affects. A dynamic approach was proposed known as decision tree. This technique considered huge factors and make predictions with high accuracy in the liver disease. Consequently, the proposed method was described the boosted C5.0 and CHAID algorithm that capable to generate new rules in one class mainly for liver disease. It was also considered the gender in disease. It was also proved that, the low chances of females than mans in liver disease. The primitive method worked as an expert and intelligent system which influenced on liver disease detection. The performance was better rather than existing techniques to detect the liver disease.

Acharya, U.R., et al., (2016) [9] give a brief description of an integrated index mainly to identify the fatty liver disease. Nowadays, the alcoholic and non-alcoholic liver diseases were one of the greatest causes of liver diseases which were progressively inclined almost in Asia and western countries. The treatment of this disease through biopsies was expensive and intrusive. The advancement of image processing and data mining approaches make it quick processed, highly efficient, objectives and decision support systems for liver disease. The primitive technique was based on feature extraction which further dependent upon the radon transform, discrete cosine transform. By this dynamic approach the average accuracy, sensitivity and specificity occurred 100% in the detection of non-alcoholic fatty liver disease. Subsequently, two significant locality sensitive discriminant analyses utilized specifically to describe the normal and FLD (fatty liver disease) class with its relative numbers.

Table 1: Related Work Survey In The Various Techniques And Parameters

Author Name	Title Name	Technique Used	Problems/ Gaps
Hassoon, M et al., 2017	Rule Optimization of Boosted C5. 0 Classification Using Genetic Algorithm for Liver disease Prediction Rule Optimization of Boosted C5. 0 Classification Using Genetic Algorithm for Liver disease Prediction	C5.0 Boost up optimization Technique	Time Consuming /decrease the accuracy rate
Saba, L., Dey et al., 2016	Automated stratification of liver disease in ultrasound: an online accurate feature classification paradigm. Computer methods and programs in biomedicine.	Levenberg–Marquardt back propagation neural network and Artificial Neural Network	It mostly utilized for training small and medium sized issues in neural network.
Abdar, M., Zomorodi-Moghadam,et al., 2017	Performance analysis of classification algorithms on early detection of liver disease.	Tree based algorithm , Boosted C5.0 and CHAID algorithms	Fatty Liver, Cirrhosis, Liver Cancer over-fit or under-fit challenges
Acharya, U. R., Fujita et al., 2016	An integrated index for identification of fatty liver disease using radon transform and discrete cosine transform features in	DCT (Discrete Cosine Transformation), Locality Sensitive Discriminant	Reduce the accuracy rate, Highly error rate to detect the liver disease.

	ultrasound images	Analysis, Support Vector Machine (SVM) and Fuzzy Sugeno	

III. LIVER DISEASE DATASETS

The data sets were obtained from the machine learning techniques which were accessed by the Online UCI Machine Learning Repository Site (Google). The data was gathered from Indian liver patients. The data was composed of 583 entiries and 416 out of them were the suffering patient parameters. The data was un-balanced and the effectiveness was highly recommended for it. In the data sets, the miner classes were required to the replicate several times to make a difference in the number of affected livers or healthy livers.

The arrangement of datasets is organized as follows:-

The datasets of Indian patients were arranged from Andhra Pradesh. The information included 441 man and 142 women. Both kind of patients were partitioned into two groups. In group 1 people was with healthy liver and the other group 2 people had the affected liver. The common features were considered in the datasets was mentioned in the below section.

While experiment, the attributes were like the age of patient, gender, total bilirubin, direct bilirubin, Alkaline phosphatase, Alamine aminotransferase, Aspartate Aminotransferase, total protein, Albumin and Albumin to Globulin Ratio.

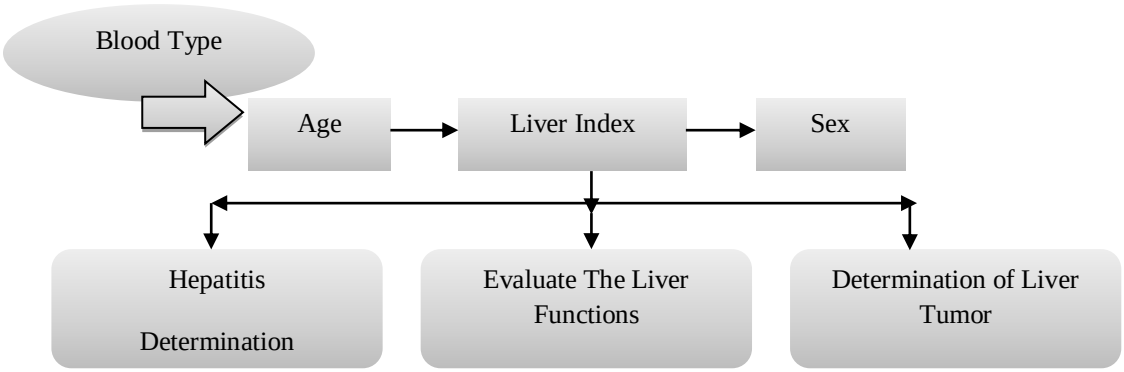

Fig.2: Data Set Structure for the detection of liver disease [11].

Basically, bilirubin is a yellowish pigment that is present in the blood and stool. The excess amount of bilirubin increased the chances of jaundice. It is found in two categories- first one is a form of a protein which is known as unconjugated and the other one is known as direct bilirubin that has the capability to flow directly in the blood cells. Alkaline phosphatase is an enzyme which is present in the blood and help to break down the proteins. Its main purpose is to check the functions of gall and liver. Alamine Aminotransferase is other kind of enzyme that also present in the blood cells and a good indication for the verification of cirrhosis and hepatitis in the liver. The lowest levels of enymes are Asparatate Aminotransferease that described the damages in an organ such as liver or a heart. The total proteins are the albumin and globumin. The albumin is the protein which has the capability to prevent the fluid in the blood cells. Tha Albumin to Globulin ratio describes the good state of the liver and it is nearly around 0.8- 2.0. [10].

The datasets were organized in two datasets as AP dataset and BUPA dataset. In AP dataset the data was considered from 416 patients and 167 were the patients who had no liver disease. The data gathered from Andhra Pradesh, India. The variable selection was a class label made by the experts to partition data in groups. The BUPA dataset consists of 345 records of liver patients of America. There were some chances of similar attributes between both datasets. The earlier five attributes were the blood cells and the other were sensitive to the liver disorders. The selection field in both cases was class labels. The true diagnosis come under class I and the false diagnosis seen in class II. [12].

Table .2 : Ap Datasets

Attributes	Types
Gender	category
Age	Positive Number
Total bilirubin	Positive Number
Direct bilirubin	Positive Number
Total protein	Positive Number
Albumin	Positive Number

Globulin	Positive Number
A and G ratio	Positive Number
SGPT	Numeral
SGOT	Numeral
Selected Field	Class

Table .3: Bupa Datasets

Attributes	Types
Attributes	Type
Mcv	Numeral
Alkphos	Numeral
SGPT	Numeral
SGOT	Numeral
Gammagt	Numeral
Selected Field	Binomical Class

IV. TECHNIQUES FOR DETECTION AND DIAGNOSIS OF LIVER DISEASE

The evolution of artificial intelligence becomes higher and it has enormous applications in the information technology. Therefore, the medical field also gaining benefits from artificial intelligence and other machine learning techniques. These techniques are playing a crucial role for the detection and diagnosis of various diseases. Some of the major techniques are described in the following section IV, which is carried out to perform and detect the diagnose liver diseases [5].

A. NaiveBayes Classifier: The naïve bayes classifier is a simple probabilistic method which is accessed to perform a set of probabilities through the determination of frequency and combination of various values in a particular data set. Its based on the Bayes theorem and follows up the rules. The classifier predicts the attributes and become independent, when given to the values of a class variable. The assumption of Naïve bayes is true for real-time applications. The performance is well and has the capability to quickly learn the supervised classification issues. The probability of a document D with vector W = (w1, w2…..,wn.) belong to the hypothesis Q.

$$R(Q_1|W_j) = \frac{R.Wj|Q_1 .R.(Q_1)}{R.Wj|Q_1 .R.Q_1.+R.Wj|Q_2 .R.(Q_2)}$$

Where $R(Q_1|W_j)$ is the probability and $R(Q_1)$ is the initial probability accessed with the hypothesis Q1 [13].

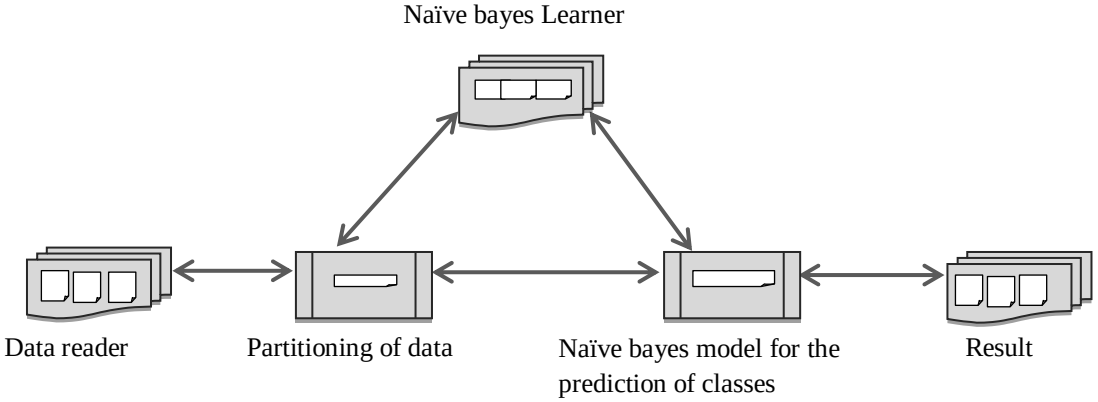

Fig.3:Naïve Bayes Classifier [14]

B. LDA (Linear Discriminant Analysis): LDA method is applied to reduce the extremely dimensional data. The LDA is associated in various applications such as face recognition, statistics analysis and machine learning to search the linear combination of features.The major aim is to maximize the ratio of class scatter to the average class scatter in the presence of low dimensional space. It is considered as an optimization issue that is sorted out with the help of an eigen value de-composition. LDA performed separately and it obtained the results on three different axes. All the vector pixels of liver and kidney are placed on the axis and this will create a spread for three fundamental classes. In this way, the evaluation of probabilistic for the values along with histogram analysis. Afterwards, the probabilistic map is obtained and the process has completed. It has also

tendency to determine the tissue probabilities. LDA come under the category of fisher faces methods which are the holistic methods [15].

C. HMM (Hidden Markov Model): It is defined as a category of the statistical model which referred to describe the progression of stochastic model. The states are not observed directly. To deeply describe a HMM model is done by defining it's attributes like N (count of states in model), M for the distinct observations of states, A is the transition probability and B considered as a observational probability of state. Pie (π) initial state distribution. The diagnosis process using HMM needs to pass through the two phases of HMM as a learning phase and exploitation phase. In both phases the description of diagnosis and prognostic is performed [16].

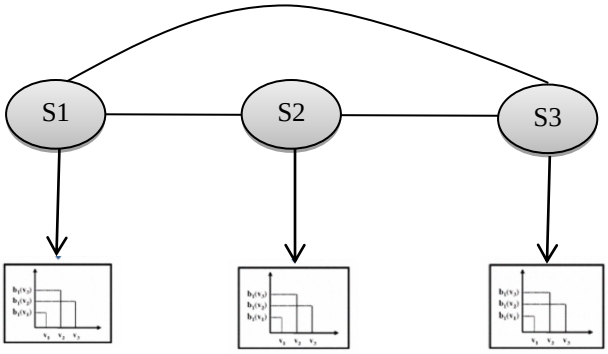

Fig.4: A three state HMM Process [16]

D. C5.0 Boostup Optimization Algorithm: This is an optimization algorithm which has the tendency to manage both continuous and discrete attributes. First of all, it creates a threshold value, then make a list of all attributes that has less value as compared to the threshold value. The biggest advantage is to manage the missing or forgotten attribute values. It considered them as a question mark (?). It is an enhanced version of C4 and it has the greatest properties that overweight the previous version. C5.0 just access the small decision tree and due to the small datasets the error rates are automatically declined which helped to gain the more accurate results.

The classification of unseen data is becoming an easy task with the use of C5.0. it follows up the basic theory of ID3 algorithm. The processing time is low as compare to C4.0 and other algorithms [17].

Table .4: Comparison Of Various Techniques

Technique Name	Description	Benefits	Drawbacks
Naïve Bayes Classifier	A probabilistic method to evaluate the set of probabilities. It based on the frequency and combined values.	Good performance Learning capability is higher.	Dependency upon the the interpretation of the issue.
Linear Discriminant analysis (LDA)	It is a type of holistic approach which mainly foxcus to expel the highly dimentionaldata.	Fastest technique It is superior as compared to PCA (Principal component analysis)	High cost
HMM (Hidden Markov Model)	A statistical process to define the stochastic procedures. The states are hidden.	Dynamic tool Good performance	Need of training data.
C5.0 Boost up Algorithm		Memory is highly efficient. Highly accurate results. Error pruning issue is solved.	Issues arised due to the missing values.

V. CONCLUSION

To summarize, the liver disease is an essential to detect and the diagnosis of disease is crucial to save the life of millions of people. Therefore, in this research, the focal point is at the detection and diagnose of liver disease. Machine learning is a major development in science and played a pivotal role in the medical area. For the detection and treatment, several methods are utilized. Naïve bayes classifier is reliant on the bayes theorem and it is applied when the dimensions of images are of higher rate. It has enormous advantages like robustness and manage missing values in an image. It has the drawback of loss of accuracy in class independence. Linear discriminanat analysis (LDA) is other method which is a category of fishers linear discriminant and used for the recognition of patterns and for the machine learning approaches. It is of high accuracy and has the capability to discriminant easily on different sets of data. It fails when subsets of a set become stronger. Hidden markov model (HMM) is a finite machine and obtained a sequence of amino acids in the form of series states. Eventually, C5.0 boostup is considered as an optimization algorithm that overweight the properties of other methods due to the higher rate of performance parameters.

REFERENCES

[1] Kumar, M.K., Sreedevi, M., Reddy & Y.C.A. Padmanabha. (2018). Survey on machine learning algorithms for liver disease diagnosis and prediction. International journal of engineering and technology, 7 (1.8), 99-120. 2

[2] Kumar, S. S., & Devapal, D. (2014, July). Survey on recent CAD system for liver disease diagnosis. In Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014 International Conference on(pp. 763-766). IEEE. 1

[3] Aziz, K., & Israel, J. (1998). Signs and Symptoms of Liver Disease. In Diseases of the Liver and Bile Ducts (pp. 3-14). Humana Press, Totowa, NJ. 9…

[4] Hassan, T.M., Elmogy, M., & Sallam, E. (2015, May). Medical image segmentation for liver disease: A survey. International journal of computer applications (0975-8887), Volume 118-No.19. 3

[5] Hassan, T. M., Elmogy, M., & Sallam, E. (2015, December). A classification framework for diagnosis of focal liver diseases. In Computer Engineering & Systems (ICCES), 2015 Tenth International Conference on (pp. 395-401). IEEE. 6.

[6] Hassoon, M., Kouhi, M. S., Zomorodi-Moghadam, M., & Abdar, M. (2017, September). Rule Optimization of Boosted C5. 0 Classification Using Genetic Algorithm for Liver disease Prediction Rule Optimization of Boosted C5. 0 Classification Using Genetic Algorithm for Liver disease Prediction. In Computer and Applications (ICCA), 2017 International Conference on (pp. 299-305). IEEE.

[7] Saba, L., Dey, N., Ashour, A. S., Samanta, S., Nath, S. S., Chakraborty, S., ... & Suri, J. S. (2016). Automated stratification of liver disease in ultrasound: an online accurate feature classification paradigm. Computer methods and programs in biomedicine, 130, 118-134.

[8] Abdar, M., Zomorodi-Moghadam, M., Das, R., & Ting, I. H. (2017). Performance analysis of classification algorithms on early detection of liver disease. Expert Systems with Applications, 67, 239-251.

[9] Acharya, U. R., Fujita, H., Sudarshan, V. K., Mookiah, M. R. K., Koh, J. E., Tan, J. H., ... & Ng, K. H. (2016). An integrated index for identification of fatty liver disease using radon transform and discrete cosine transform features in ultrasound images. Information Fusion, 31, 43-53.

[10] Sontakke, S., Lohokare, J., & Dani, R. (2017, February). Diagnosis of liver diseases using machine learning. In Emerging Trends & Innovation in ICT (ICEI), 2017 International Conference on (pp. 129-133). IEEE. 10

[11] Lin, R. H. (2009). An intelligent model for liver disease diagnosis. Artificial Intelligence in Medicine, 47(1), 53-62. 1…..

[12] Bahramirad, S., Mustapha, A., & Eshraghi, M. (2013, September). Classification of liver disease diagnosis: a comparative study. In Informatics and applications (ICIA), 2013 second international conference on (pp. 42-46). IEEE. 2

[13] Patil, T. R., & Sherekar, S. S. (2013). Performance analysis of Naive Bayes and J48 classification algorithm for data classification. International journal of computer science and applications, 6(2), 256-261.

[14] Classification, A. (2018). A simple explanation of Naive Bayes Classification. [online] Stack Overflow. Available at: https://stackoverflow.com/questions/10059594/a-simple-explanation-of-naive-bayes-classification [Accessed 21 Aug. 2018].

[15] Gloger, O., Kühn, J., Stanski, A., Völzke, H., & Puls, R. (2010). A fully automatic three-step liver segmentation method on LDA-based probability maps for multiple contrast MR images. Magnetic Resonance Imaging, 28(6), 882-897.

[16] Tobon-Mejia, D. A., Medjaher, K., Zerhouni, N., & Tripot, G. (2011, May). Hidden Markov models for failure diagnostic and prognostic. In Prognostics and System Health Management Conference (PHM-Shenzhen), 2011 (pp. 1-8). IEEE.

[17] Pandya, R., & Pandya, J. (2015). C5. 0 algorithm to improved decision tree with feature selection and reduced error pruning. International Journal of Computer Applications, 117(16), 18-21.