INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 2)
Available online at: www.ijariit.com

Liver disease prediction using machine learning

Vasan Durai
vasan1993@gmail.com
SRM Institute of Science and Technology, Chennai, Tamil Nadu

Suyan Ramesh
suyanramesh@gmail.com
SRM Institute of Science and Technology, Chennai, Tamil Nadu

Dinesh Kalthireddy
kalthireddy.dinesh@gmail.com
SRM Institute of Science and Technology, Chennai, Tamil Nadu

ABSTRACT

Data Mining technologies have been widely used in the process of medical diagnosis and prognosis, extensively. These data mining techniques have been used to analyze a colossal amount of medical data. The steep increase in the rate of obesity and an unhealthy lifestyle eventually reflects the likelihood and the frequent occurrence of liver-related diseases in the mass. In this project, the patient data sets are analyzed for the predictability of the subject to have a liver disease based purely on a widely analyzed classification model. Since there are pre-existing processes to analyze the patient data and the classifier data, the more important facet here is to predict the same the conclusive result with a higher rate of accuracy. There are 5 distinct phases in this process. First, the min-max algorithm is applied to the original liver patient data set that could be collected from the UCI repository. In the second phase, significant attributes are demarcated by the use of PSO feature selection. This helps to bring out the subset of critical data, from the whole normalized datasets of liver patients. After this step, the third phase involves the usage of classification algorithms for comparative analysis and categorization. Accuracy Calculation is the fourth phase. It involves the usage of Root Mean Square value and a Root Error value. The fifth phase is the evaluation phase. Depending on the studies, a simple evaluation process is executed to preserve the integrity of a precise result reflection. J48 algorithm is considered to be a better performing algorithm when it comes to feature selection with an accuracy rate of 95.04%.

Keywords— *Chronic diseases, Classification schemes, Training datasets, Machine learning, Classifiers, Algorithms, Classification models*

1. INTRODUCTION

Digital Technological revolution is marking its potential in the scope of disruptive innovation. With Nanotechnology and Genetic, marking the upward surge of medical technologies - the sky seems to be the limit when it pertains to conceiving the different ways to use the immense potential of the digital marketing era, ineffective prognosis, diagnosis, treatment and healthcare monitoring. A considerable large amount of data is periodically dealt, with respect to the passage of every instance of a medical process functioning. These data sets could be inferential in nature, referential in nature or could be raw enough to be conclusive of further relevant sets of meaningful medical information. This information is not only diversely sources but is also diversely utilised. They can be used for the prognosis, diagnosis and treatment of diseases. The study of the same could accelerate the pace of research projects in the same context. It could help in statistical inferences when it comes to projecting different patterns that could help in the process as a whole. In data mining, classification techniques are much popular in medical diagnosis and predicting diseases [1]. The liver is the second largest internal organ in the human body, playing a major role in metabolism and serving several vital functions, e.g. Decomposition of red blood cells, etc.,. [6] Its weight comes around three pounds. The liver performs many essential functions related to digestion, metabolism, immunity and the storage of nutrients within the body. These functions make the liver as an important organ, without this, body tissues would quickly die from lack of energy and nutrients. Traditionally, liver disease can be diagnosed clinically by analyzing the levels of enzymes in the blood [7]. In this research work, a combination of Naïve Bayes and Support Vector Machine (SVM) classifier algorithms are used for liver disease prediction.

2. REVIEW OF TRENDS AND RELATED LITERATURE

Anu Sebastian, Surekha Mariam Varghese, "Fuzzy Logic for Child-Pugh classification of patients with Cirrhosis of Liver" [2]: Survival Analysis is an extensively used procedure in the field of medical science. The idea of being able to predict the life expectancy of the subject is of immense value and utility to both, the doctors and the patients. There are three preliminary steps that serve as the elementary foundation of any medical treatment paradigm. The diagnosis stage, the classification stage, the assessment stage, the conclusion stage and finally the treatment stage. All these stages are expected to be accurate to the parameters and effective in their measure to distinctly reflect the quantified magnitude and the intensity of the study of the disease in the context. One of the most widely used classification methodologies that have been used for an extensive assessment of liver diseases, particularly cirrhosis is the Child-Pugh classification method. It is understandable from the extensive study of a voluminous set of cases that the life expectancy of different patients, suffering from different intensities and kinds of liver cirrhosis, is different. Fuzzy Logic, for instance, suits the context like a tailor-made technique.

Insha Arshad, Chiranjit Dutta, "Liver Disease Detection Due to Excessive Alcoholism Using Data Mining Techniques" [3]: Liquor is expended in overabundance by a large number of individuals over the world. Liquor utilization is legitimately connected to perilous liver maladies, for example, cirrhosis which may at last lead to death. Early location of liver illness brought about by overutilization of liquor would help in sparring existences of numerous individuals. By distinguishing liver ailment in its beginning time, it very well may be analyzed in time and may prompt full recuperation in certain patients. This paper proposes identification just as to foresee the nearness of liver sickness utilizing information mining calculations. We will settle on a choice tree for the dataset and afterwards the principles will be created. Subsequent to deciding the principles, we will utilize diverse information mining calculations to prepare and test the dataset to distinguish the liver sickness. The information was gathered from UCI storehouse and our preparation dataset was created. It comprises of 7 unique qualities having 345 occurrences. In the dataset, distinctive classes of blood tests are taken into contemplations which are straightforwardly connected to liver illnesses that may emerge because of unnecessary liquor utilization alongside recurrence of liquor utilization. In light of the sort of liver sickness recognized, the forecast might be proposed.

N. Ramkumar, S. Prakash, S. Ashok Kumar, K Sangeetha, "Prediction of liver cancer using Conditional probability Bayes theorem" [4]: Malignant growth is the one of the unsafe infection on the planet. Malignant growth spreads in lungs, liver, bosom, bones and so forth. Liver malignancy is the most hazardous and it will proceed with long-lasting. The side effects of a liver malignant growth are Jaundice, loss of weight, yellow shaded pee, spewing, torment in the upper right stomach area, sweats, fever and amplified liver. The liver malignancy which starts in the liver separated from moving from another piece of the body is called as essential liver disease. A disease which spreads all other pieces of the body lastly it achieves liver is called an auxiliary liver malignant growth. The liver is one of the critical pieces of the human. WHO reviews state out of 100,000 individuals, around 30 individuals have experienced liver malignant growth and generally it influences the African and Asian nations prior. These days it turned into a well-known ailment. The most widely recognized sort of a liver malignant growth is called hepatocellular carcinoma, this specific influences male as opposed to female. The liver malignant growth happens for the most part because of the more liquor utilization. Numerous information mining calculations, artificial insight ideas are utilized to anticipate liver disease. The likelihood of anticipating the liver malignant growth is performed utilizing the Bayes hypothesis with the WEKA tool.

Mafazalyaqeen Hassoon, Mikhak Samadi Kouhi, Mariam Zomorodi Moghadam, Moloud Abdar, "Rule Optimization of Boosted C5.0 Classification Using Genetic Algorithm for Liver disease Prediction" [5]: One of the fascinating and vital subjects among scientists in the field of therapeutic and software engineering is diagnosing disease by considering the highlights that have the most effect on acknowledgements. The subject talks about another idea which is called Medical Data Mining (MDM). Undoubtedly, information mining techniques utilize diverse ways, for example, characterization and grouping to arrange maladies and their indications which are useful for diagnosing. This paper presents another technique for liver illness analysis to help specialists and their patients in finding the sickness side effects and decrease quite a while of diagnosing and counteract passings. The proposed strategy will streamline the tenets discharged from Boosted C5.0 grouping technique with the Genetic Algorithm (GA), to expand the determination time and exactness. So as opposed to utilizing a transformative calculation for creating rules, the hereditary calculation is utilized for improving and diminishing tenets of another calculation. We demonstrate that our proposed methodology has better execution and throughput in correlation with other work in the field. The precision is improved from 81% to 93% in our work

2.1 Findings

The mechanisms that are currently used in the prediction of liver disease are prone to have different levels of accuracy and effectiveness. The sense of importance, though, is determined by the need of the hour. Different diseases demand accuracy of a different set of parameters and might not demand the same set of inferences, throughout more than a single case. In the near future, the study reflects that there was a decent amount of accuracy that was achieved. However, the agenda of this paper is to improvise on those lines and come up with better accuracy standards. The slack in the accuracy in the recent cases had been tackled by designating different combinations to be considered, while the case study is being considered. The existing models are also reflective of certain issues that pertain to the handling of the training dataset and data elements. The following are some of the clear limitations that have been observed, in order to account for innovation in this paper, having brought about the connotation of improvising on these lines.

- When it comes to the classification process, it is not necessary that the cohesion that a classifier shares with a particular set of data should stand viable for the rest of the training set. This is to imply that there are some classifiers that don't stand fit to the data set in the context.
- Some of the machine learning approaches that are being considered, do not stand viable for a large volume of data. The due consideration is given to the process, owing to the fact that the methodology suits the conditions where there are smaller volumes of data.
- There are certain methodologies that are incompatible and non-cohesive when it comes to the collection of real-time data and the implementation procedures of the same.

3. ANALYSIS OF FACTORS AFFECTING ACCURACY

When it pertains to machine learning, the inferential information sets are a product of the commonly observed observations that often reflect this sense of pattern in the collection of data or the respective data set. There is a process of characterization, which is reaffirmed by the study and the accounting of a voluminous amount of data that relates to the context, and the study of the same. On the same lines, when the data is of limited volume and is curtly presented for consideration in the machine learning algorithm, it is difficult to come up with accurate and seemingly glaring inferential patterns and thereby the result of predictive analysis of a set of data. There are a set of issues that continue to challenge the accuracy of the machine learning algorithms that are used for predictive analysis.

(a) The Quantity of data that is involved: The concise, yet precise nature of this argument being - the more the data, the more accurate the result of the predictive analysis. With lesser data, the accuracy and effectiveness of the predictive process decline.

(b) Scope of the issue: With machine learning paradigms demanding a huge collection of data for analysis, it is important to give due importance to the selectiveness of the features, that would pivotally define the boundaries of context, in any given problem. It is important to maintain relevance and sync with the issue/problem statement.

(c) Parameters that are involved as a part of the method: The study and analysis of the algorithm and the larger system, as a whole should also be feasible to be executed by non-technicians and absolute rookies with the basic understanding of the functioning of the system. In modern machine learning algorithms, the sense of innovation is reassured with the involvement of more than a single parameter that is involved in the analysis of the scope. These multiple parameter settings are induced and boundary by only the user's understanding, experience and the knack of being about to intuit the necessary parameters that need involvement/tweaking.

(d) Features in the data: It is imperative for any machine learning algorithm or the developer/data analyst to be able to sparsely collate the raw data and project the potentiality in the rich feature space. This is expected to accelerate the learning process of a machine learning system.

(e) Quality of Data: Any data that is to serve as a template for critical studies, fabrication, analysis and research of any subject - needs to be thoroughly checked on qualitative grounds. This is because even the slightest sense of lethargy can vandalize the integrity of the process and compromise on the potential and the expectancy, to be able to deliver.

4. EXPERIMENTAL STUDY: INFERENCE FROM LITERATURE REVIEW

There are some logically strong inferences that can be made from the literature review. Since the thesis is to composite the ideology of using machine learning algorithms for the prognosis, diagnosis and study of liver diseases and their predictability, it is important to deal majorly with the kind of machine learning algorithms that would suit the purpose and be centric on the major objectives - being able to predict the presence of a liver disease in the most accurate possible way. The literature surveys conclude the use of Naive Bayes and Support Vector Machine algorithms for the prediction of liver diseases. There are two major parameters that are involved in understanding the suitability of the respective methodologies and they are - the time taken to execute the prediction process and the accuracy of the predictive result. It is clear through various studies and experimentations that SVM classifier is the best of all the algorithms owing to the extremely high accuracy rates. But when it comes to the time taken to execute the predictive process, the Naive Bayes classifier reflects higher suitability since it takes the least possible time to execute the process.

4.1 Objective

The objective of this paper is to be able to predict the occurrence of liver disease in a sample dataset/ training data set, in order to be able to calculate the predictability to the greater magnitude of accuracy using the appropriate machine learning algorithm.

4.2 Present System

The present system shares the same objective but encompass different methodologies to arrive at a relatively less accurate conclusion. The qualitative superiority that these methods have over one another is dependent on the accuracy of the results produced. There are different aspects of the data that are used in order to parametrically come to a definite conclusion over the prediction of liver disease. Fuzzy logic has been developed for the classification of patients with liver cirrhosis. In gastroenterology, the Child-Pugh score is used to assess the prognosis of chronic liver disease, mainly cirrhosis. It was originally made to predict the mortality during surgery.

It is now used to determine the prognosis, the required strength of treatment and the necessity of liver transplantation.

The following score is instrumental in utilizing the five clinical measures of liver disease and each of these measures are scored between 1 and 3, with 3 indicating a serious condition of organ deterioration.

Measure	1 point	2 points	3 points
Total bilirubin, μmol/L (mg/dL)	<34 (<2)	34–50 (2–3)	>50 (>3)
Serum albumin, g/dL	>3.5	2.8–3.5	<2.8
Prothrombin time, prolongation (s) **OR** INR	<4.0 <1.7	4.0–6.0 1.7–2.3	> 6.0 >2.3
Ascites	None	Mild (or suppressed with medication)	Moderate to severe (or refractory)
Hepatic encephalopathy	None	Grade I–II	Grade III–IV

Some use a modified version of the Child-Pugh score where there are reflective changes in the fact that these diseases feature high conjugated bilirubin levels. The upper limit for a single point is designated to be 68 μmol/L (4 mg/dL) and the upper limit for 2 points is 170 μmol/L (10 mg/dL).

Chronic liver diseases are classified into Child-Pugh class A to C, employing the reflective score from the table above.

Points	Class	One-year survival	Two-year survival
5–6	A	100%	85%
7–9	B	80%	60%
10–15	C	45%	35%

The systems were designed in a similar fashion with the involvement of a comparative degree. The standardization of preset conditions is prefixed. These conditions act as templates with which the training data sets are compared, depending on the results of which, the conclusions are made. Under some practically feasible circumstances, the results can be inferred through pure mathematical modelling. The Bayes theorem for conditional probability can be put to use in order to predict the predefined scenario, which is contextual, the occurrence of liver disease. There are two critical ends to this methodology of approach. The predecessor end is to come up with the template that distinctly points at the presence of the disease element in the test cases' report. The template is defined with the patterns observed in the physiological conditions that are observed in a number of test cases. Data mining techniques are used to come to one particular conclusion that pertains to the characteristics of patients of all kinds who have liver diseases. The data mining techniques would also take in the data elements of patients who have been subjects of excessive alcoholism, which could have more likely been the causality behind the disease.

4.3 Proposed System

Machine learning is understandably one of the most extensively utilized paradigms of big data management where a significantly high set of distinct raw data can be collated effectively to make appropriate inferences and eventually to come up with a usual collection of contextually useful collection of integrative information. With the onset of the exponential technological

explosion in the field of medicine, there is a felt need to handle a colossal set of data, thereby managing and utilizing the same to make effective and informative inferences for the doctors and patients.

4.4 Advantages of the Proposed System

Considering the certain differences that have been adopted in the current system the following are the distinct advantages that are observed:

- **The performance classification of liver-based diseases is further improved:** with the far deepened understanding of the different kinds of ailments in the field of medicine, the different set of parameters to distinctly determine the kind of liver disease and its occurrence has become a far less complicated task. With advancements in data mining paradigms and software architectures like Hive, R, easing up the data collection process, the preprocessing and evaluation stages are given more attention to.
- **Time complexity and accuracy can be measured by various machine learning models, so that we can measures different parameters, owing to the needs of the user:** Every prediction system is based on the kind of parameters that it is expected to accept, compare and then finally come to a predictive conclusion. Accordingly, there are different algorithms that are used to model the predictive system to suit the context. The different machine learning algorithms judge the kind of disease and the testing parameters.
- **Different machine learning having high accuracy of the result:** In comparison to other methodologies considered, the right machine learning algorithm can aptly increase the efficiency of the results that are expected out of the predictive system.
- **Risk factors can be predicted early by machine learning models:** The machine learning algorithms predict the risk factors through simple methodologies of analysis the inconsistencies in the collective training data set and their respective parameters.

4.5 Advantages of Machine Learning Algorithms

Machine learning is a functionality of a system to be able to learn through the extensive usage of examples that pose a set of conditions that can be incorporated as a part of the self-improvement process without being coded by a programmer. The result, thus obtained is then used by the corporate, in order to make actionable inferences for decision making. It has its roots related to data mining and close association with Bayesian predictive modelling. The data is taken as an input by the machine and the result is formulated as the output. Typical machine learning algorithms are utilized in trying to improve the user experience by providing recommendations using historical data. This would be an opportunistic approach to utilise this unsupervised learning to do the same.

4.6 Machine Learning vs. Traditional Programming

In traditional programming paradigm, the programmer is required to analyse, study and code all the rule subordinations in accordance with the experts and their recommendations on an advisory capacity. These rules act as the logical foundation for the machine. When the system grows, there is a rising need in the complexity of these systems and the need to incorporate more and more rules. This can get too haphazard to maintain. Machine learning replaces the conventional paradigms under these circumstances. The learning systems in the machine learning paradigm are centric on enabling the system to derive theses functional rules in order to inferentially use a set of example patterns to derive those rules and build a solid logical foundation in a system.

4.7 Working of Machine Learning Algorithms

The machine learning component is considered to be the brain of the system where all the learning aspects take place and are controlled centrally. The machine learning algorithms enable the system to learn, similar to how the human brain does. Human brains are used to understanding and making viable inferences using experiences. However, in order for a machine to make an accurate prediction, the following data could be utilized. The core activity phases of a machine learning system would be - learning and inference. The discovery of patterns plays a major role. Feature selection would be the follow-up procedure, where it is decided which of the core values of the field are put to use. The discovery part is facilitated with the collection of data, which is put to use. The right set of data is also critical at the feature selection stage. The list of these attributes is chosen by what is known as an attribute vector.

4.8 Inferences

The functionality of the proposed system has to be tested for the kind of limitations that could put up constraints on the operations of the system. The powerfulness of the system is tested by exploring the limits using data that the system has never been acclimated to or the kind of data that is unexplored at every level. The new data that is incorporated into the system, is incorporated and transformed into a features vector, go through the model and then conclusively come up with a prediction.

The life of Machine Learning programs is straightforward and can be summarized in the following points:

- Define a question
- Collect data
- Visualize data
- Train algorithm
- Test the Algorithm
- Collect feedback
- Refine the algorithm
- Loop 4-7 until the results are satisfying
- Use the model to make a prediction

4.9 Supervised learning

A calculation utilizes preparing information and criticism from people to get familiar with the relationship of offered contributions to a given yield. For example, an expert can utilize showcasing cost and climate gauge as information to foresee the offers of jars. You can utilize administered realizing when the yield information is known. The calculation will foresee new information. There are two classifications of regulated learning:

- **Classification task:** Imagine you need to foresee the sexual orientation of a client for a business. You will begin gathering information on the tallness, weight, work, pay, obtaining crate, and so forth from your client database. You know the sexual orientation of every one of your client; it must be male or female. The target of the classifier will be to dole out a likelihood of being a male or a female (i.e., the name) in light of the data (i.e., highlights you have gathered). At the point when the model figured out how to perceive male or female, you can utilize new information to make an expectation. For example, you just got new data from an obscure client, and you need to know whether it is a male or female. On the off chance that the classifier predicts male = 70%, it implies the calculation is certain at 70% that this client is a male, and 30% it is a female. The mark can be of at least two classes. The above precedent has just two classes, yet on the off chance that a classifier needs to anticipate object, it has many classes (e.g., glass, table, shoes, and so forth each article speaks to a class)
- **Regression task:** When the yield is persistent esteem, the assignment is a relapse. For example, a money-related

investigator may need to gauge the estimation of a stock dependent on a scope of highlights like value, past stock exhibitions, and macroeconomics record. The framework will be prepared to gauge the cost of the stocks with the least conceivable blunder.

5. CONCLUSION

Information mining is the technique to recover an example from huge informational collection regarding AI, information base, and insights. An information mining strategy such as grouping, order and affiliation which is suitable for medicinal finding. Prominent order calculation, for example, SVM, NB and others considered for execution in assessment in liver issue infections forecast. In liver issue infections there are 500 informational indexes with 10 traits. The qualities are Total Bilirubin, Direct Bilirubin, Total Proteins, Albumin, A/G proportion, SGPT (Alanine Aminotransferase), SGOT (Aspartate Aminotransferase) and Alkaline Phosphatase. Future work we can utilize the blend and increasingly Hybrid way to deal with improve execution exactness for liver issue maladies forecast with their reasonable informational collections

6. ACKNOWLEDGEMENTS

The authors acknowledge the support from SRM Institute of Technology, Ramapuram, Chennai for providing and sustaining necessary domain insights and enabling a conducive learning environment.

7. REFERENCES

[1] Bendi Venkata Ramana, Surendra. Prasad Babu. M, Venkateswarlu. N.B, A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis, International Journal of Database Management Systems (IJDMS), Vol.3, No.2, May 2011 page no 101-114

[2] Sebastian, Anu, and Surekha Mariam Varghese. "Fuzzy logic for child-pugh classification of patients with cirrhosis of the liver." 2016 International Conference on Information Science (ICIS). IEEE, 2016.

[3] Arshad, Insha, et al. "Liver disease detection due to excessive alcoholism using data mining techniques." 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE). IEEE, 2018.

[4] Ramkumar, N., et al. "Prediction of liver cancer using Conditional probability Bayes theorem." 2017 International Conference on Computer Communication and Informatics (ICCCI). IEEE, 2017.

[5] Hassoon, Mafazalyaqeen, et al. "Rule optimization of boosted c5. 0 classification using a genetic algorithm for liver disease prediction." 2017 International Conference on Computer and Applications (ICCA). IEEE, 2017.

[6] Karthik. S, Priyadarshini. A. Anuradha J. and Tripathi B. K, Classification and Rule Extraction using Rough Set for Diagnosis of Liver Disease and its Types, Ad.

[7] Thapa, B. R., and Anuj Walia. "Liver function tests and their interpretation." The Indian Journal of Pediatrics 74.7 (2007): 663-671.

[8] Sullivan, Tim. "Blitzscaling." Harvard business review 94.4 (2016): 15.

[9] Jae-Young Lim, "The Prospect of the Fourth Industrial Revolution and Home Healthcare in Super-Aged Society", https://www.researchgate.net/profile/Jae-Young_Lim2.

