Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 26 (2023), No. 1, pp. 189–199
DOI : 10.47974/JSMS-958

Liver disease prediction using ensemble model

Bichitrananda Patra *
Sudhansu Sekhar Bisoyi †
Triloknath Pandy §
Gayatri Nayak @
Faculty of Engineering and Technology
Institute of Technical Education and Research
Siksha 'O' Anusandhan (Deemed to be University)
Bhubaneswar
Odisha
India

Abstract

Despite the fact that medical science has progressed to the point where it is capable of identifying and prescribing the exact cure for known diseases, traditional medicinal practices are outdated and slow and it can take so long to identify a disease that it becomes fatal to patients who seek immediate attention. Because the liver plays such a crucial role in eliminating toxins from our bodies, this research focuses mostly on the prediction of liver disorders such as jaundice and various forms of hepatitis. As a result, early detection and treatment of liver illnesses are critical for successful diagnosis and recovery. For the purpose of demonstrating our work, we used examples of various liver illnesses. We used a disease dataset with a variety of symptoms as features to predict the disease in this paper. Before fitting the data into the Machine Learning model, various data pretreatment techniques and a filter-based feature selection strategy are employed on it. Three different supervised machine learning models, SOM, Back propagation and Ensemble model are used to train the data. The accuracy of these entire machine learning classifier models was evaluated and it was discovered that the ensemble Classifier produced the best results, with nearly 99.18 percent accuracy.

Subject Classification: *68.*

Keywords: *Machine learning, SOM, Back propagation, Ensemble, Classifier.*

1. Introduction

The liver is the largest internal organ in the human body, weighing an average of 1.6 kg (3.5 pounds). The main job is to convey blood from

*E-mail: bichitranandapatra@soa.ac.in (Corresponding Author)
† E-mail: sudhanshushekharbisoyi@soa.ac.in
§ E-mail: trilokpandey@soa.ac.in
@ E-mail: gayatrinayak@soa.ac.in

the digestive tract to the remainder of the body before it is transferred. The liver also creates proteins that are required for blood coagulation and other processes. Carbohydrates are broken down in the liver and converted to glucose, which is then injected into the bloodstream to maintain normal glucose levels. The liver is the only visceral organ that regenerates.

Cirrhosis claimed the lives of 1.32 million people worldwide in 2017, accounting for 2% to 4% of all deaths [1]. Every year, morbidity rises and liver disease-related mortality is estimated to be roughly 2 million fatalities per year worldwide, with 1 million deaths owing to viral hepatitis and hepatocellular carcinoma [2]. According to Wang and coworkers, liver transplant trends in the United States indicate the rise of NAFLD as a major cause of end-stage liver disease and hepatocellular carcinoma [3].

Recognizing chronic liver illnesses and determining which condition a person has is difficult. This research primarily focuses on a Machine Learning (ML)-based model that may help humans distinguish seven different forms of liver illnesses, including jaundice, hepatitis A, B, C, D, E and alcoholic hepatitis. The suggested approach can determine the type of liver illness a person is suffering from just by providing their symptoms as input. As a result, a person will be much more aware of their liver health and will be able to take the appropriate steps. The Ensemble Classifier model was employed in this model to achieve the highest level of accuracy for detecting liver illness.

People with liver disorders used to have to wait for doctors to examine them and treat them. The number of hospitals and doctors has grown throughout time. However, as the world's population grows, so does the number of sick individuals, who are treated by a limited number of doctors. It is literally impossible for everyone to afford skilled doctors at the outset of a disease's symptoms. We are well aware that computers were created to lessen man's effort while working in order to overcome this situation. So why not develop a liver disease prediction algorithm that will not only save time but also assist people in the early stages of symptoms when they suspect something is wrong with their health?

The key goals of this paperwork are to provide a liver disease detection model that is as accurate as possible for a person who is suffering from it and to recommend certain prescriptive measures and drugs and then to advise the individual to seek medical advice as soon as feasible.

The research efforts related to this issue are mentioned in Section 2. The dataset collection and preprocessing are detailed in Section 3. The proposed model and result analysis were reported in Sections 4 and Section

5, respectively. Section 6 brought the study to a close by highlighting the work's future prospects.

2. Literature Review

Some of the work on liver disease prediction models has been covered in this section. Many studies have been done in this topic, according to our literature review and some of those efforts are discussed here. A vector machine method was utilised to diagnose liver illness in a recent study [3]. Support Vector Machine (SVM) was utilised in their study to diagnose liver illness using a diabetic dataset and two datasets from liver patients.

The author applied supervised machine learning for automatic white area identification in liver biopsies in another research [4]. The liver photos were examined using supervised machine learning classifiers based on annotations provided by two expert pathologists, according to the paper.

The author of a study [5] utilises Bayesian Classification to predict liver illness. They employed Bayesian classification methods in that study, which is one of the most common classification models.

In a 2012 study, data mining was utilised to diagnose fatty liver disease in ultrasonography [6]. The goal of this study was to develop one of the Computer-Aided Diagnostic (CAD) tools for accurately distinguishing between healthy livers and Fatty Liver Disease-affected livers (FLD).

A survey on data mining classification strategies for analysing liver disease disorder was undertaken in 2016 [7]. Data mining is the process of extracting valuable information from big databases. The application of a data mining classification technique to the research of liver disease disorder is described in this paper. Predicting Mouse Liver Microsomal Stability [8] was done using "Pruned" Machine Learning Models and Public Data in another research. They talked about how to create machine learning models that can help with MLM stability recognition. Hepatitis is a viral infection that affects the liver. Artificial Neural Networks (ANNs) were employed in a study to diagnose hepatitis virus-related liver damage. [9,16]. The results of prior hepatitis diagnosis investigations that employed the same database were also utilised in this paper.

3. Methodology

3.1 *Dataset collection*

The collection and preparation of data sets is the first stage in any Machine Learning model. The data for our experiment came from a liver illness prediction dataset accessible on Kaggle5. Kaggle is a fantastic

website that was founded in 2010 by a group of data scientists. In an online data science environment, Kaggle allows users to search for datasets, explore, create and submit Machine Learning projects. The data in our dataset can be downloaded as CSV files. The dataset file has a total of 133 columns, with 132 of them describing the various symptoms. The prognosis information appears in the last column. These signs and symptoms are linked to seven different liver disorders.

3.2 *Rough Set Feature Selection*

In the microarray gene expression data analysis [17], the data is represented as a collection of real valued vectors. The large dimension of these vectors makes it difficult to detect cancer-causing flag genes.

As a result, a technique of dimensionality reduction of the attribute for semantic preservation of the dataset is simply desirable. There is a permanent change of data in the semantic-destroying dimensionality reduction technique also known as attribute selection [15] that does not change the meaning of the original attribute. The rough set theory can be the approach to uncover the data dependencies by reducing the attributes of the dataset without utilizing any other information than the accessible dataset.

The null values were replaced by the most frequent value or the mode value of that column. As the values of the input variables were either 0 or 1, so there were no such outliers in the dataset. Feature selection helped in highly improving the overall accuracy of our ML model.

3.3 *Proposed Model*

Figure 1 specifies the steps that are used in our proposed model.

3.3.1 Self-Organizing Map (SOM)

The Map That Organizes Itself The Self-Organizing Map (SOM) was invented by T. Kohonen [14]. The approach of learning the set of pattern distribution without any prior information or knowledge of the class is referred as unsupervised learning. It has the property of maintaining topology. The neurons compete for the opportunity to be engaged or fired. As a result, only one neuron wins the competition and is dubbed the winner-takes-all neuron. The SOM can be one dimensional or multidimensional. However, the maps of one dimensional or two dimensional are the most prevalent. The number of categorical attribute determines the number of connections to be employed in the input. During training, the neuron with the weights closest to the input data vector is deemed the winner. The

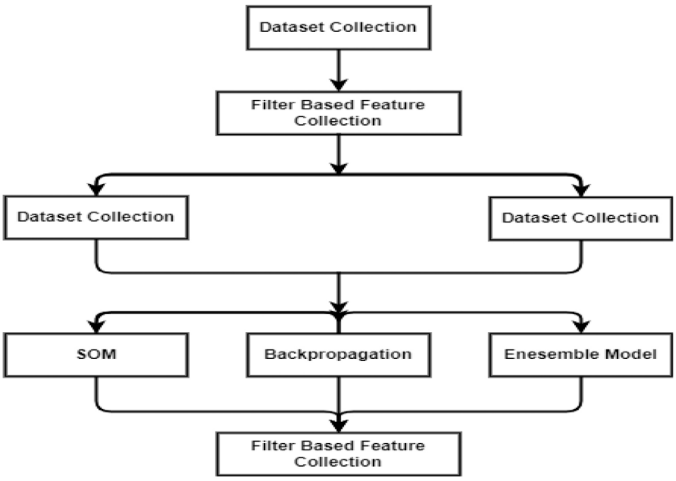

Figure 1

Flowchart of the proposed model

weights of all the neurons in the winning neuron's neighbourhood are then changed by an amount that is inversely proportionate to the distance. Based on the set of attributes employed, it clusters and classifies the data set. The following is a summary of the algorithm:

Step 1: Initial weight vectors $wj(0)$ to be chosen from the random values, with separate weight vectors for j =1,2,...l, where total number of neuron is represented by I.

Step 2: Sampling: With a specific probability, draw a sample x from the input space.

Step 3: Matching Similarities: Using the smallest distance, find the best matching (winning) neuron I (x) at time steps n. Criteria of Euclid

Step 4: Keep it up to date.

Step 5: Repeat Steps 2–5 until the feature map shows no discernible changes.

3.3.2 Back Propagation Algorithm

For a given collection of input patterns with established classifications, the back propagation algorithm trains a feed-forward multilayer neural

network. The network checks its output response to the sample input pattern when each entry of the sample set is provided to it. The error value is obtained after comparing the output response to the known and desired output. The connection weights are modified based on the mistake. The weight adjustment is done through mean square error of the output response to the sample input "Vel98 Simulation Program" in the back propagation process, which is based on the Widrow-Hoff delta learning rule. The network is repeatedly presented with this set of sample patterns until the error value is minimized. Neurons have L+1 layers and synaptic weights have L layers. We want to adjust the weights W and biases b to bring the actual output closer to the desired output.

The training of the feed-forward multilayer neural network for a given pattern as a set of input with known classification is carried out through the back propagation learning algorithm. The network response is examined to the specified sample pattern of input by presenting each entry of the sample set to the network. The error value is calculated with the help of a comparative analysis of output response to the known and desired output. The adjustment of the connection weight are done based on the error value. The Widrow-Hoff delta learning rule is used in the back propagation algorithm where weight adjustment is done through mean square error of the output response to the sample input "Vel98 Simulation Program".

3.3.3 Stacking Ensemble Model

When compared to a single classifier, ensemble approaches aggregate numerous classifiers to produce superior predictive efficiency [2]. The ensemble model's main idea is to combine a set of weak learners to form a powerful learner, thus increasing the model's precision. The key sources of the disparity between real and predicted values when using any machine learning approach to estimate the target variable are noise, variation and bias. Ensemble aids in the reduction of these factors (with the exception of noise, an irreducible mistake).

Stacking [11] is another ensemble model in which a new model is trained using the mixed results of two (or more) previous models. The model predictions are utilised as inputs for each sequential layer and merged to create a new set of predictions. Because all figures are blended to form a forecast or classification, ensemble stacking is also known as mixing. When compared to separate learning methods, this model enhances performance. The stacking of Probability Distributions (PDs) and Linear Multi-Response (MLR) are the state-of-the-art methods [12]. It

is generally knowledge that a combination of different base-level classifiers (with weakly correlated predictions) produces good results. Merz [13] proposes the SCANN stacking approach, which uses correspondence analysis to find correlations between base-level classifier predictions.

The data is prepared for usage in a Machine Learning model after the data cleaning and preprocessing process, which was mentioned earlier in this section. For training and testing reasons, the dataset is split into two portions. The data for training and testing is separated into 841 rows and 8 rows, respectively. The data is loaded into three separate Machine Learning models after it has been divided. Self-Organizing Map(SOM) model, Back propagation model and Stacking Ensemble Classifier model are among the models. The performance or accuracy of these models was then compared, with the model with the highest accuracy score being used in subsequent steps. Because this model will be used in the field of healthcare, its accuracy cannot be compromised, hence generating the correct outcome is critical. So, the goal of this performance evaluation was to determine the model that best matches our dataset and produces the greatest results. In this example, it was discovered that the Ensemble Classifier model provides the highest level of accuracy, hence it was chosen. The Liver Disease Prediction Model was created after we chose our machine learning model. When a patient describes their symptoms, our model can accurately identify the type of liver disease they have, the prescription measures they should take and the medications they should take after consulting with a healthcare expert.

4. Result Analysis

Three distinct ML classifier models have been developed and their performances have been tested, as in this study [18,19]. As a result, the first graph in Figure 2 compares the models' accuracies. The SOM model has a 90.94 percent accuracy, the Back propagation model has a 88.72 percent accuracy and the Ensemble Classifier model has a 99.18 percent accuracy. As a result, we chose Ensemble Classifier as our final liver disease prediction model because it has the greatest accuracy rate.

In the second graph, the accuracy of prediction for specific liver illnesses such as Jaundice, Hepatitis A, Hepatitis B, Hepatitis E and Alcoholic Hepatitis was evaluated independently using all three models.

This model's major goal is to solve various real-world issues. As a result, this model considers a wide range of real-world circumstances. The patient may neglect to mention a few symptoms while entering them

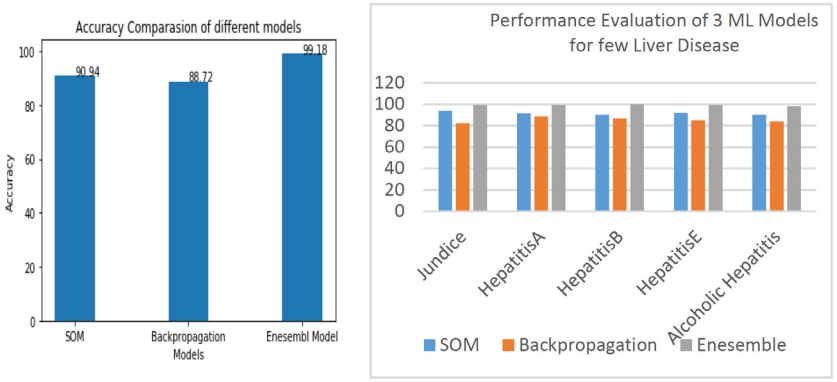

Figure 2

Prediction of three classifier ML models

into the system. It's also possible that the patient is completely ignorant of any symptoms. By grouping every 20 continuous classes, the dataset is separated column by column. According to this graph, the first 20 symptoms in the dataset can provide a high accuracy percent on their own, while the fourth group of symptoms has the lowest accuracy percent. As a result of this finding, it may be concluded that the first set of symptoms is the most important and that even if the user misses some of the other symptoms, they should focus on the first set.

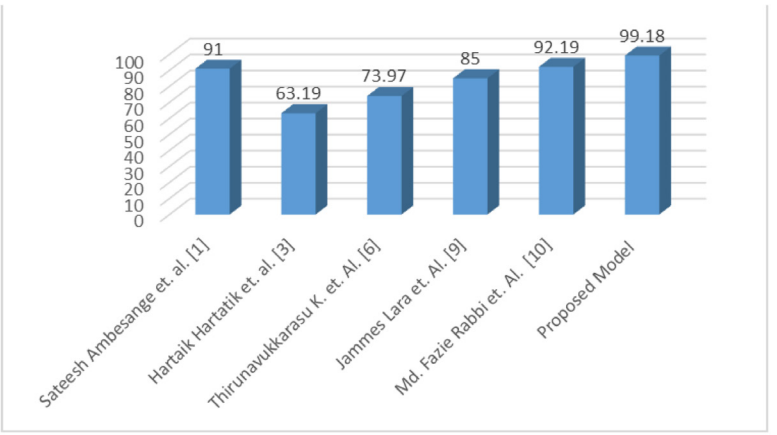

Figure 3

Accuracy prediction of various other similar models

The accuracy of the suggested model is compared to various other similar models in the next graph in Figure 3. The proposed model's accuracy is more percent, which is the greatest among all the other models. Because accuracy cannot be compromised at any cost, particularly in health-related situations, a model with higher accuracy will always be the most efficient for users.

5. Conclusion & Future Works

This model's greatest accomplishment is achieving a near-perfect accuracy of roughly 99.18% while keeping the total approach simple. The main goal has been to make the overall paradigm easier to apply. It translates to less complex code, faster execution times, lower hardware requirements and overall superior performance across all platforms. This code is also incredibly scalable and simple to use, so anyone with a basic understanding of programming may use it. We used a dataset from Kaggle, which is an open source platform that may be changed at any time. When it comes to limitations, the only one that comes to mind is that this research is limited to liver problems. The dataset should be increased and data on a variety of additional disorders should be included, so that this study can reach a wider audience. It's exciting to consider that if the method works with the same, or nearly the same, accuracy and efficiency after being modified and added to numerous different illness types, then this code could be used to solve one of the world's most pressing problems.

Speaking of the Limitations, it can be said that the only limitation found here is that this work is limited only to liver diseases. The dataset needs to be expanded and data related to several other types of diseases are needed to be included, so that this work can reach more and more people. It is fascinating to realize that, if the algorithm, after modification and addition of several other disease types, works with the same, or nearly the same accuracy and efficiency, then this code can be the tool to solve one of the greatest real-life problems. It goes without saying that our long-term goal is to continue working on the same project, expanding the dataset to cover more and more diseases and adapting the algorithm to work with the same efficiency and accuracy.

References

[1] S. Ambesange, R. Nadagoudar, RUppin, V. Patil, S. Patil and S. Patil, "Liver Diseases Prediction using KNN with Hyper Parameter Tuning

Techniques," 2020 IEEE Bangalore Humanitarian Technology Conference (B-HTC), 2020, pp. 1-6.

[2] Patra, B., Mondal, S. (2022). Neural Ensemble Recognition for Lung Cancer Credentials. In: Gunjan, V.K., Zurada, J.M. (eds) Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications. Lecture Notes in Networks and Systems, vol 237. *Springer, Singapore.* https://doi.org/10.1007/978-981-16-6407-6_59.

[3] H. Hartatik, M. B. Tamam and A. Setyanto, "Prediction for Diagnosing Liver Disease in Patients using KNN and Naïve Bayes Algorithms," 2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS), 2020, pp. 1-5.

[4] Sepanlou SG, Safiri S, Bisignano C, Ikuta KS, Merat S, Saberifiroozi M, et. al. (2020) The global, regional and national burden of cirrhosis by cause in 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet Gastroenterol Hepatol. 5(3):245–266. pmid:31981519.

[5] Sumeet K Asrani, Harshad Devarbhavi, John Eaton, Patrick S Kamath (2019) Burden of liver diseases in the world.pmid:30266282.

[6] Wang S, Toy M, Hang Pham TT, So S (2020) Causes and trends in liver disease and hepatocellular carcinoma among men and women who received liver transplants in the U.S., 2010–2019. PLoS ONE 15(9): e0239393. pmid:32946502.

[7] Patra, Bichitrananda, et. al. "Analysis and Prediction of COVID-19 Pandemic." Deep Learning, Machine Learning and IoT in Biomedical and Health Informatics. CRC Press, 2022. 69-88.

[8] T. Kannapiran, A. S. Singh and A. Chowdhury, "Prediction of Liver Disease using Classification Algorithms," 2018 4th International Conference on Computing Communication and Automation (ICCCA), 2018.

[9] Tripathy, H. K., Mishra, S., Thakkar, H. K., & Rai, D. (2021). CARE: A Collision-Aware Mobile Robot Navigation in Grid Environment using Improved Breadth First Search. Computers & Electrical Engineering, 94, 107327.

[10] Seto WK, Lo YR, Pawlotsky JM, Yuen MF (2018) Chronic hepatitis B virus infection. Lancet 392(10161):2313–2324. pmid:30496122.

[11] Lara, James, et. al. "Computational models of liver fibrosis progression for hepatitis C virus chronic infection." BMC bioinformatics 15.8 (2014): 1-9.

[12] Rabbi, Md Fazle and Debakanta Mishra. "Using FWD deflection basin parameters for network-level assessment of flexible pavements." *International Journal of Pavement Engineering* 22.2 (2021): 147-161.

[13] Deepa Mathur & Deepak Bhatia (2022) Gait classification of stroke survivors - An analytical study, *Journal of Interdisciplinary Mathematics*, 25:1, 163-181, DOI: 10.1080/09720502.2021.2006332.

[14] Two Spatial Population Dynamics Models. *International Journal of Applied and Computational Mathematics* 8:3.

[15] Ruchi Kaushik, Vijander Singh & Rajani Kumari (2021) Multi-class SVM based network intrusion detection with attribute selection using infinite feature selection technique, *Journal of Discrete Mathematical Sciences and Cryptography*, 24:8, 2137-2153, DOI: 10.1080/09720529.2021.2009189.

[16] Shalu & Dinesh Singh (2021) Artificial neural network-based virtual machine allocation in cloud computing, *Journal of Discrete Mathematical Sciences and Cryptography*, 24:6, 1739-1750, DOI:10.1080/09720529.2021.1878626.

[17] Patra, Bichitrananda. "Reliability Analysis of Classification of Gene Expression Data Using Efficient Gene Selection Techniques." *International Journal of Computer Science Engineering & Technology* 1.11 (2011).