WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Liver Disease Diagnosis and Detection Using Machine Learning Classifiers

Sanchit Sharma

Dept. Of Computer Science Engineering, Medi-Caps University, Indore, India Sanchitsharma151102@gmail.com

Abstract—

In India, the burden of liver diseases is enormous, and in 2017, based on the most recent data from the World Health Organization, liver diseases contributed to 259,749 deaths, which was equivalent to 2.95% of all deaths. Shockingly, India currently accounts for nearly one-fifth of all cirrhosis deaths, which is approximately equivalent to 18.3% of global cirrhosis deaths. Considering the constantly changing economic scenery in India, lifestyle changes and a shift in dietary patterns, there is a distinct possibility that etiological factors causing liver cirrhosis over the past few years may have deviated.

This paper presents our research findings and the development of an algorithm capable of detecting liver disease in patients using blood sample results. We have utilized Python and ML classifiers for algorithm generation. Our results show that the program works well and gives fast results, which is really promising.

Keywords— Liver Disease, Machine learning, Hepatitis, SVC, Classifier

Introduction

Liver disease is a growing concern for public health, affecting millions of individuals worldwide and creating a systemic burden on healthcare systems. Therefore, the prompt and accurate diagnosis of liver disease is essential for successful care processes and achievement of favorable outcomes. Advances in computer sciences and machine learning, especially in the context of present-day healthcare, have encouraged the development of various predictive models for liver disease diagnosis. The current study seeks to support the creation of the Liver disease prediction System using Machine Learning Classifiers that would assist in the identification of liver diseases in their earlier phases. The use of computational algorithms and analysis of patient data can advance the capabilities of medical practitioners for the accurate evaluation of liver health and decision-making.

It is especially critical to detect liver diseases early because these disorders have a wide range of causes that can have severe consequences. There are various types of liver diseases, including hepatitis due to viruses, alcoholic liver disease due to heavy drinking, non-alcoholic fatty liver disease due to obesity and metabolic syndrome, and cirrhosis characterized by irreversible liver tissue scarring. In addition, liver cancer, autoimmune hepatitis, and genetic conditions such as hemochromatosis and Wilson's disease are life-threatening.

Early detection of liver diseases helps to take clinical action that slows down the pathological process. In addition to stopping the disease's progression, this contributes to optimizing treatment results. The patient is directed to preventive measures, a definite way of living, and treatment approaches. It is possible because early diagnosis allows at-risk people to be identified.

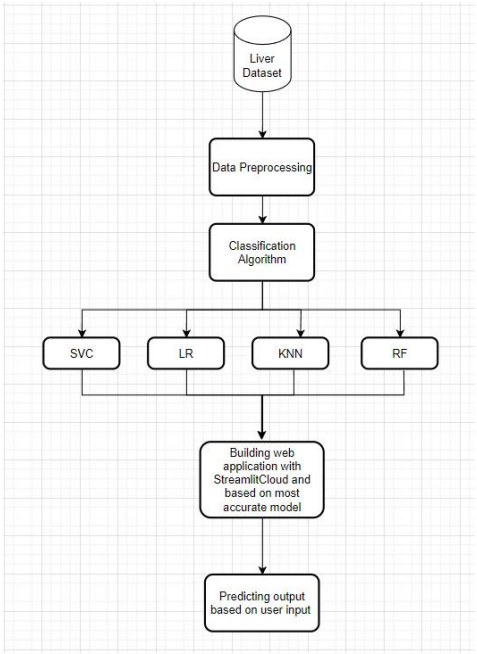

Fig. 1. Proposed System

Methodology

In this study, the dataset of Indian liver patients are utilized for training and testing of the classification model. The dataset used in this research is taken from UC Irvine, Machine Learning Repository. This data set contains records of 416 patients diagnosed with liver disease and 167 patients without liver disease. This information is contained in the class label named 'Selector'. There are 10 variables per patient: age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT and Alkphos. Of the 583 patient records, 441 are male, and 142 are female. The current dataset has been used to study - differences in patients across US and Indian patients that suffer from liver diseases, gender-based disparities in predicting liver disease, as previous studies have found that biochemical markers do not have the same effectiveness for male and female patients.

The aim of the model is to predict the liver disease. For model building and training, python libraries are used like pandas, matplotlib, and etc. Library is used which is an open-source library for high-performance numerical computation. For training the model, Jupyter notebook environment is also used. The proposed model was trained with 80% train dataset and the remaining 20% data were used to validate the performance of the model after it has been trained. The dataset contains attributes like patients age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT and Alkphos. Of the 583 patient records, 441 are male, and 142 are female. The current dataset has been used to study - differences in patients across US and Indian patients that suffer from liver diseases. - gender-based disparities in predicting liver disease, as previous studies have found that biochemical markers do not have the same effectiveness for male and female patients. Then the model was trained using svc architecture and predict the disease. Training accuracy and loss were obtained for each epoch. One epoch is when an entire dataset is passed forward through the neural network only once.
After each training iteration, the model performance was evaluated with the validation set to get validation accuracy and loss. Binary cross entropy measures how far away from the true value the prediction is for each of the classes and then averages these class wise errors to obtain the 14 final loss. Early stopping is employed to stop the model when validation loss is the lowest. It is a form of regularization used to avoid overfitting when training a learner with an iterative method. Thus, the trained model is evaluated with the validation set for computing the accuracy and average loss. The trained model is applied for predicting new input patient data.

A. Data Preprocessing:

In this phase, two datasets are utilized which are BUPA Liver Disorders and another is Indian Liver Patient Dataset (ILPD). These datasets are then analyzed in jupyter notebook using many python libraries like pandas, matplotlib, seaborn etc. After which they are classified into train and test data for classification model.

B. Model Training:

In this phase, model training for liver disease prediction utilizes a Support Vector Classifier (SVC) to analyze patient data attributes, including age, gender, total bilirubin, liver enzymes, and albumin levels. SVC, a supervised learning algorithm suitable for binary 10 classification tasks, is trained to classify patients into two categories: those with liver disease and those without. To ensure uniform contribution to the classification process, input features are scaled, and kernel functions (e.g., linear, polynomial, radial basis function) are employed to transform features into a higher-dimensional

space. Hyperparameters such as the choice of kernel, regularization parameter (C), and kernel coefficient (gamma) are tuned for optimal model performance.

During training, SVC learns the optimal decision boundary to separate different classes in feature space, aiming to minimize classification errors while maximizing the margin between support vectors. Trained SVC model predicts the likelihood of liver disease for each patient, assigning a class label (positive or negative) based on their proximity to the decision boundary. Model performance is evaluated using metrics like accuracy, precision, recall, F1-score, and confusion matrix, providing insights into the model's ability to correctly classify patients. SVC's interpretability is enhanced by identifying support vectors, allowing clinicians to understand which features are most influential in predicting liver disease. In conclusion, the described SVC model architecture offers a framework for leveraging machine learning in liver disease diagnosis and management, empowering healthcare professionals with predictive insights derived from patient data analysis

C. Support Vector Machine (SVM):

Classification algorithms are extensively employed across various medical domains, aiming to develop robust models for predicting class labels of unknown data instances. These models are trained using a dataset comprising input data points and their corresponding class labels. One such popular algorithm is the Support Vector Machine (SVM), which operates by separating data into distinct categories through the construction of an N-dimensional hyperplane. SVMs are akin to classical multilayer perceptron neural networks, as they also construct hyperplanes in high-dimensional or infinite-dimensional spaces. The effectiveness of an SVM is determined by its ability to create a hyperplane with the largest distance to the nearest training data point of any class, known as the functional margin. Different kernel functions, such as linear, polynomial, sigmoid, and radial basis function (RBF), can be utilized in SVMs, with the choice depending on the application and determined through cross-validation. There are many possible kernel functions and the most common kernel are: Linear, polynomial, sigmoid and radial basis function (RBF).

In this paper we use linear kernel function shows in equation 1: $K(xi,xj) = xiTxj$

Feature selection plays a crucial role in SVM modeling, where predictor variables, termed attributes, are transformed into features used to define the hyperplane. The aim is to identify the optimal hyperplane that separates clusters of vectors, with cases belonging to one category of the target variable positioned on one side, and cases of the other category on the opposite side. The vectors close to the hyperplane are referred to as support vectors.

D. Streamlit Cloud :

We Deployed the application using Streamlit Cloud, an Open-source app framework for Machine Learning and Data Science which automatically gets deployed as the GitHub repository is updated. Streamlit Cloud is a platform that allows developers and data scientists to build and share data applications1. It's designed to turn data scripts into shareable web apps in minutes. You can deploy, manage, and share your apps with the world directly from Streamlit. The deployment process is simple: sign in with Github or SSO, pick a repo, branch, and file, and then click Deploy.

Streamlit Cloud has revolutionized the way we build and share data applications. It's a powerful platform that allows developers and data scientists to turn data scripts into shareable web apps with ease.

Fig. 2. Correlation Matrix

Results

The model has been trained with two datasets(BUPA dataset and ILPD) and is giving an accuracy of 99%.[6] This accuracy is very high as compared to other machine learning algorithms like NB or KNN. Timely detection is very important in case of a liver disease which can save a patient life. So, if it is possible for us to judge quickly whether any patient have liver disease or not. So that, we can take further steps to give the treatment. Even though it is non-invasive and blood test report is required, quickly monitering can bridge the gap of find a deadly liver disease.[4]

TABLE I. RESULT OF ALL ML ALGORITHMS [4]

	Average Accuracy	Average Precision	Average Recall	Average F1 score
Naive Bayes	0.9120	0.9917	0.8876	0.9359
KNN	0.9882	0.9844	1	0.9921
SVB	0.9882	0.9847	0.9894	0.9917
SVM	**0.9922**	**0.9895**	**0.9920**	**0.9907**
Decision Tree	0.9863	0.9974	0.9894	0.9933
Random Forest	0.9901	0.9974	0.9920	0.9946

Conclusion

Predicting liver diseases through non-invasive methods provides a convenient and accessible means for early detection and monitoring. By leveraging machine learning, we can enhance pediatric healthcare practices, ultimately improving infant health outcomes.

In summary, the provided code marks a significant advancement in non-invasive liver disease detection, highlighting technology's potential to revolutionize pediatric healthcare.

References

[1] Kawano, T. T. Zin and Y. Kodama, *"A Study on Non-contact and Non-invasive Neonatal Jaundice Detection and Bilirubin Value Prediction,"* 2018 IEEE 7th Global Conference on Consumer Electronics (GCCE), Nara, Japan, 2018, pp. 401-402, doi: 10.1109/GCCE.2018.8574674.

[2] UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/datasets/Hepatitis, 29/09/2017,7:48.

[3] W. Hashim, M. Alkhaled, A. Al-Naji and I. Al-Rayahi, "A Review on Image Processing Based Neonatal Jaundice Detection Techniques," 2021 7th International Conference on Contemporary Information Technology and Mathematics (ICCITM), Mosul, Iraq, 2021, pp. 213-218.

[4] Fatema-Tuz-Zohra Khanam, Ali Al-Naji, Asanka G. Perera, Danyi Wang & Javaan Chahl (2023) Non-invasive and non-contact automatic jaundice detection of infants based on random forest, Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 11:6, 2516-2529, DOI: 10.1080/21681163.2023.2244601

[5] K. Srividya, K. Renganathan, M. S and Y. U, "Review on Jaundice Detection in Neonates Using Image Processing," 2022 International Conference on Communication, Computing and Internet of Things (IC3IoT), Chennai, India, 2022, pp. 1-5, doi: 10.1109/IC3IOT53935.2022.9767938.

[6] NJN: A Dataset for the Normal and Jaundiced Newborns, https://doi.org/10.3390/biomedinformatics3030037

[7] Abdulrazzak AY, Mohammed SL, Al-Naji, "NJN: A Dataset for the Normal and Jaundiced Newborns," 2023 BioMedInformatics, 3(3):543-552, doi: 10.3390/biomedinformatics3030037

[8] Biomarkers for prediction of liver fibrosis in patients with chronic alcoholic liver disease written by Sylvie Naveau and Bruno Runyard.

[9] Strategic analysis in prediction of liver disease using classification algorithms written by Piyush Kr Shukla and Binish Khan.

[10] Software based prediction of liver disease with feature selection and classification techniques witten Jagdeep Singh, Sandeep Bagga and Ranjodh Kaur.

[11] Liver disease prediction using SVM and Naïve Bayes algorithm written by S Dhayanand.

[12] Prediction and analysis of liver diseases using data mining written Shambel Kefelgen, Pooja Kamat.

[13] Amin, Md Nurul, and Md Ahsan Habib. "Comparison of different classification techniques using WEKA for hematological data." American Journal of Engineering Research (AJER) e-ISSN: 2320-0847 p-ISSN: 2320-0936 Volume-4, Issue-3, pp-55-61 www. ajer. org (2015).

[14] Karthikeyan, T., and P. Thangaraju. "Analysis of classification algorithms applied to hepatitis patients." International Journal of Computer Applications 62.15 (2013).

[15] Ba-Alwi, Fadl Mutaher, and Houzifa M. Hintaya. "Comparative study for analysis the prognostic in hepatitis data: data mining approach." spinal cord 11 (2013).

[16] Panchal, Gaurang, et al. "Behaviour analysis of multilayer perceptrons with multiple hidden neurons and hidden layers." International Journal of Computer Theory and Engineering 3.2 (2011): 332.

[17] H. M. Mudawi, "Epidemiology of viral hepatitis in Sudan," Clinical and Experimental Gastroenterology, vol. 1, pp. 9–13, 2008.

[18] Mitra, Malay, and R. K. Samanta. "Hepatitis disease diagnosis using multiple imputation and neural network with rough set feature reduction." Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014. Springer, Cham, 2015.

[19] C. Barath Kumar, M. Varun Kumar, T. Gayathri, S. Rajesh Kumar. Data Analysis and Prediction of Hepatitis Using Support Vector Machine (SVM). International Journal of Computer Science and Information Technologies, Vol. 5 (2), ISSN: 0975-9646. (Dec, 2014).

[20] Behera, N. K. S., et al. "Bird mating optimization based multilayer perceptron for diseases classification." Computational Intelligence in Data Mining-Volume 3. Springer India, 2015. 305-315.

[21] Pushpalatha, S., and Jagdesh Pandya. "Data model comparison for Hepatitis diagnosis." India, July (2014).