AFrica Leadership

X - ACADAMY

ALX

Predicting liver disease : using Machine Learning Algorithm

Submitted by: - Mubarik Meka

Submitted to: -

Dr.Kunamar(Phd)

Submission date: June 25, 2024

Abstract

Introduction: Liver diseases such as hepatitis, cirrhosis, and liver cancer are major global health concerns, contributing to a significant percentage of mortality. Early and accurate diagnosis remains challenging due to the limitations of conventional testing. This study explores the use of machine learning algorithms to improve liver disease detection using clinical parameters. By classifying patients based on liver function indicators, we aim to develop a cost-effective, reliable prediction model that supports early diagnosis and reduces reliance on invasive procedures.

Objective: The objective of the research was developing robust classification model for prediction of liver disease with Machine Learning Approaches in kaglgle data

Methodology: This study used WEKA 3.9.6 to develop a liver disease prediction model based on a Kaggle dataset with 30,691 records. After removing 5,425 entries with missing values, the data was cleaned, normalized using min-max scaling, and balanced. Ten features were selected, with one target variable. Classification algorithms—including Decision Tree, Bayesian Network, Random Forest, and Bagging—were applied. Model performance was evaluated using precision, recall, F1-score, and confusion matrix. Python 3.2 was used for visualization

Result and descussion: The study assessed the performance of various machine learning classifiers for liver disease prediction using accuracy, precision, F1-score, and other metrics. Random Forest and Bagging achieved the highest accuracy (0.998 and 0.994), followed by Decision Tree (0.990). Bayesian Network (Bayes.nt) showed the lowest performance, with an accuracy of 0.879 and F-measure of 0.88. Overall, ensemble methods proved more effective in delivering accurate and reliable predictions for liver disease classification.

Conclussion: This study applied and evaluated several machine learning algorithms for liver disease prediction. Random Forest and Bagging outperformed other models, with Random Forest achieving the highest accuracy (99.8%) and lowest error. Bayes.net showed the weakest performance, highlighting its limitations for this dataset. The results confirm the strength of ensemble methods in healthcare classification tasks. Future work may explore real-time deployment and model interpretability.

TABLE OF CONTENT

LIST OF TABLE

LIST OF FIGURE

1. Introduction

The liver, the largest organ in the human body, plays a crucial role in digesting food and removing toxins. Damage to the liver, often caused by alcohol consumption or infections, can lead to serious and life-threatening conditions. Various liver diseases exist, including hepatitis, liver cancer, liver tumors, and cirrhosis, with cirrhosis and related disorders being the most common causes of death. [1]

Together with other organs, the liver plays a key role in processing, absorbing, and digesting food. Its primary function is to filter blood coming from the digestive tract before circulating it to the rest of the body. The liver also breaks down medications and detoxifies harmful substances. During this process, it produces bile, which is later returned to the intestines. In addition, the liver synthesizes essential proteins necessary for blood clotting and various other vital functions. [2]

Seventy percent of deaths worldwide are caused by liver disease [4]. More accurate methods of detecting and diagnosing liver disease must be developed. Above all, patients must have access to and be able to afford liver function tests. Applying statistical machine learning algorithms to CMP findings to extract information for a physician may be useful for diagnosis in order to avoid the costly and invasive tests [3,5]. Machine learning holds great promise for enhancing disease detection and prediction, which has attracted significant attention in the biomedical field. It also contributes to making the decision-making process more objective [6].

improving the accuracy of outcome prediction are the primary goals of this study. As a result, we classified patients according to whether they had liver disease or not using various algorithm.

1.1 Objective

1.1.1 General objectives

The main objective of this thesis is:

- ✓ To develop an accurate and reliable machine learning-based prediction system for liver disease diagnosis by evaluating and comparing the performance of the model.

1.1.2 Specific objectives

- ✓ To preprocess and analyze clinical liver disease data, addressing missed values
- ✓ To implement and train four machine learning algorithms on the curated dataset
- ✓ To assess and compare model efficacy using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

2 Literature Review

The accuracy of the machine learning techniques discussed in earlier research has been assessed using a mix of k-fold cross-validation, receiver operating characteristic under area under curve, and confusion matrix.In order to forecast the risk of liver illness from a data set containing liver function test results, Singh et al. created software based on classification methods, such as logistic regression, random forest, and naive Bayes [7].

Vijayarani and Dhavanand discovered that SVM outperformed naive Bayes in predicting cirrhosis, acute hepatitis, chronic hepatitis, and liver malignancies [8]. Compared to SVM, random forest, Bayesian network, and an MLP-neural network, SVM with particle swarm optimization (PSO) predicted the most crucial features for liver disease detection with the highest accuracy [9]. Compared to Bayesian and other previously used models, SVM predicted drug-induced hepatotoxicity more correctly with fewer molecular descriptors [9].

Structured data was the study's main focus.For optimal output classification and prediction, logistic regression, random forest, decision trees, and XGBoost Classifier were employed. The XGBoost Classifier and logistic regression both had very good accuracy. The accuracy for the random forest model was roughly 74.57%. [10]

3. Materials and Methodology

In this section, we provide a brief overview of the dataset used, describe the data preparation techniques applied, and explain the role of data normalization in the machine learning algorithms implemented using the **WEKA 3.9.6** open-source software.

3.1 Data Collection

This study used a secondary dataset from Kaggle, with 71.6% male and 25.42% female participants. To develop an effective liver disease prediction model, key preprocessing steps—such as cleaning, feature selection, and class balancing—were applied. From 11 parameters in the dataset, 10 were chosen as input features and 1 as the target variable. Features included age, gender, bilirubin levels, alkaline phosphatase, and GPT.

3..2. Tool and Language

In this study, WEKA 3.96 software as a tool and python 3.2 programming language for visualization is used

3.3 Data Preprocessing

In this study, we removed 5,425 instances with missing values from a total of 30,691 records. The remaining cleaned data primarily consists of patients without liver disease, used for class balancing. After removal, 21,917 instances correspond to liver disease patients, while 877 represent non-liver disease patients

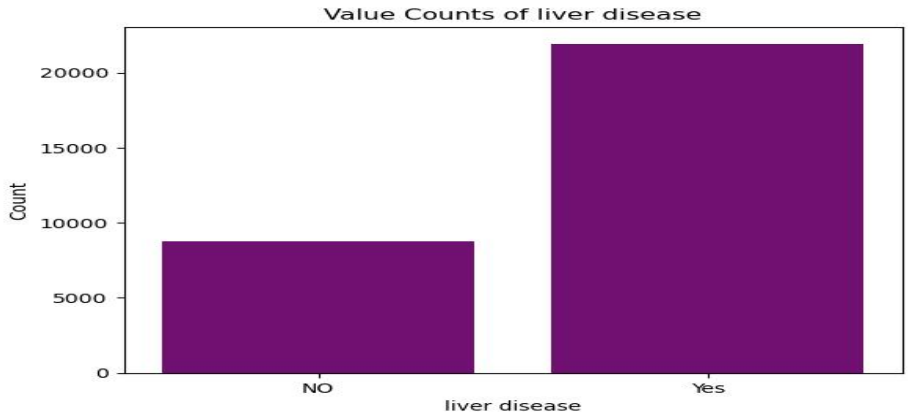

Figure 1 Count Plot shows the ratio of liver patients

3.3.1 Normalization

Normalization scales numeric data to a fixed range, typically 0 to 1, ensuring all features are on the same scale. Using min-max normalization, each value is adjusted based on its feature's minimum and maximum, improving algorithm performance and training speed.

Table 1Min-Max Normalization Analysis in WEKA

Viewer

Relation: liver_disease2-weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0-weka.filters.unsupervised.attribute.NominalToBinary-

No.	1: Age of the patient Numeric	2: Gender of the patient=Male Numeric	3: Total Bilirubin Numeric	4: Direct Bilirubin Numeric	5: ?Ikphos Alkaline Phosphotase Numeric
1	0.23255813953488372	0.0	0.004021447721...	0.005102040816...	0.06106497313141182
2	0.2558139534883721	1.0	0.004021447721...	0.005102040816...	0.06106497313141182
3	0.5465116279069767	1.0	0.004021447721...	0.005102040816...	0.06106497313141182
4	0.5465116279069767	1.0	0.004021447721...	0.005102040816...	0.06106497313141182
5	0.2558139534883721	1.0	0.004021447721...	0.005102040816...	0.06106497313141182
6	0.4418604651162791	1.0	0.004021447721...	0.005102040816...	0.06106497313141182
7	0.20930232558139536	1.0	0.004021447721...	0.005102040816...	0.06106497313141182
8	0.20930232558139536	1.0	0.004021447721...	0.005102040816...	0.06106497313141182
9	0.4186046511627907		0.004021447721...	0.005102040816...	0.06106497313141182
10	0.2441860465116279	1.0	0.004021447721...	0.005102040816...	0.06106497313141182
11	0.16279069767441862	1.0	0.004021447721...	0.005102040816...	0.06106497313141182

3.4 Model Building and Optimization

Different classifier models are going to be utilized for experimentation in model building and hyperparameter tuning within the prepared data at hand.

3.4.1 Random Forest Model

Random Forest builds multiple decision trees using randomly selected data samples. Each tree makes a prediction, and the final output is based on the majority vote. More trees generally improve accuracy and reduce errors. [11].

3.4.2 Decission tree(DT)

The computation of decision trees is a component of supervised learning algorithms [12]. Unlike other supervised learning algorithms, a decision tree approach can also be used to address classification and regression problems.
Making a training model that can be used to forecast class or estimate objective characteristics by using choice criteria established from previous data (training data) is the main idea behind using decision trees.

3.4.3 Bayesian Networks

A Bayesian Network (BN), also known as a belief network, is a probabilistic graphical model that represents a set of variables and their conditional dependencies Each node in the graph represents a random variable, while edges between nodes denote probabilistic relationships. The network encodes the joint probability distribution over all variables through chain rule factorization [13]

$$P(X_1, X_2, \ldots, X_n) = \Pi P(X_i | parents(X_i))$$

3.4.4 Bootstrap Aggregating

It is a machine learning operates as a wrapper method that improves the stability and accuracy of base estimators through:Bootstrap Sampling,Parallel Training and Aggregation[14]

4.RESULTS AND DISCUSSION

About WEKA

WEKA was used as the simulation tool for model evaluation. It offers an easy-to-use interface for data preprocessing, training, and testing without requiring coding, making it ideal for beginners and research-focused analysis

Evaluation Metrics Overview

To evaluate model performance, key classification metrics—accuracy, precision, recall, F1-score, and confusion matrix—were used. Random Forest and Bagging performed best, with accuracy scores of 0.998 and 0.994. Decision Tree followed with 0.990 accuracy, while Bayes.net showed the lowest performance (accuracy 0.879, F-measure 0.88), making it the least effective.

Table 2 the comparative performance metrics

algorithm	Accuracy	precision	F-Measure	ROC Area
Randsom forest	0.998	0.99	0.99	1.000
DT	0.990	0.99	0.99	0.999
Bayes.nt	0.879	0.9	0.88	0.968
bagging	0.994	0.99	0.99	1.000

The following Fig 2 describes an overall classificationalgorithm for accuracy values analysis.

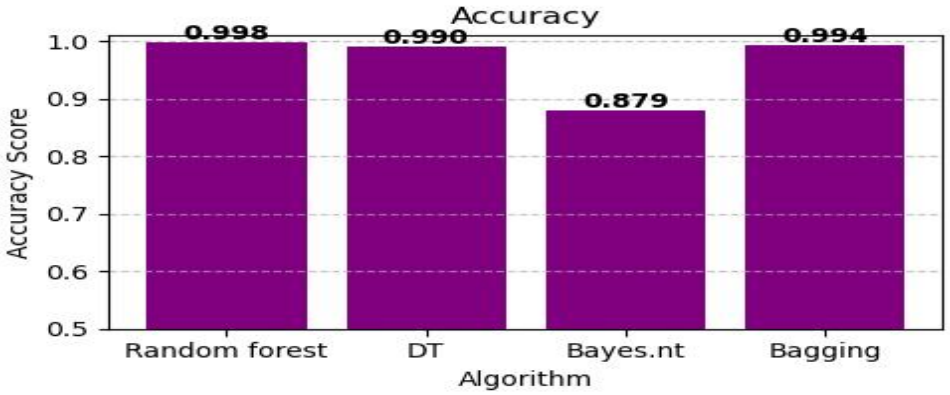

Figure 2 accuraracy measure of the algorithm

Root Mean Square Error (RMSE)

It measures the average prediction error. Random Forest had the lowest RMSE (0.0558), indicating the most accurate predictions. Bagging and Decision Tree had slightly higher errors, while Bayes.net had the highest RMSE (0.2927), making it the least reliable model. Lower RMSE indicates better performance.

Table 3 Root Mean Square Error

Classification of algorithm	Root Mean Square Error
Randsom forest	0.0558
DT	0.0874
Bayes.nt	0.2927
bagging	0.0781

The following figure illustrates the RMSE of each model, helping to identify which algorithm produced more accurate predictions.

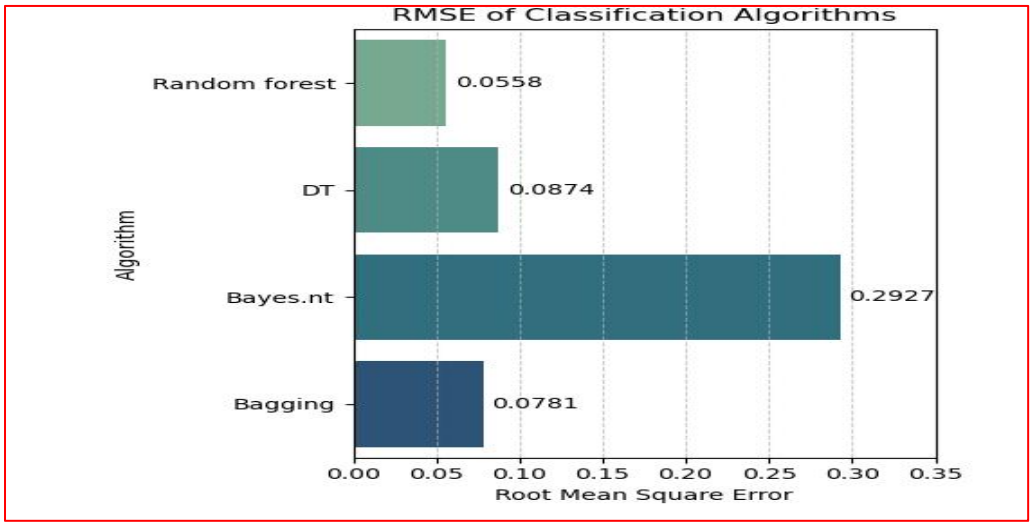

Figure 3 RMSE Comparison of Classification Algorithms

4.9 Discussions

The results show that machine learning models effectively predict liver disease. Random Forest and Bagging performed best, with near-perfect accuracy (0.998 and 0.994) and balanced precision and recall (F-Measure 0.99). Decision Tree followed closely but may suffer from overfitting. BayesNet lagged behind, indicating weaker handling of complex data. Overall, ensemble methods demonstrated superior classification and robustness.

5. CONCLUSION AND FUTURE WORK

In this study, various machine learning algorithms were applied to classify the dataset and their performance was evaluated using key metrics including Accuracy, Precision, F-measure, ROC area, and Root Mean Square Error (RMSE). Among the models tested, Random Forest emerged as the most effective classifier, achieving the highest accuracy (99.8%) and the lowest RMSE (0.0558), followed closely by Bagging. These ensemble models demonstrated strong predictive performance and robustness. In contrast, Bayes.net produced the lowest results across most metrics, indicating its limited suitability for the dataset used in this research. The findings highlight the importance of using ensemble techniques in classification tasks, especially in healthcare-related datasets where high precision and accuracy are critical.

This study validates the efficacy of ensemble learning while underscoring the importance of algorithm selection based on dataset characteristics and performance metrics. Future work could investigate real-time implementation or explainability techniques for these high-performing models.

6. REFERENCES

1) K. Sumeet, J.J. Larson, B. Yawn, T.M. Therneau, W.R. Kim, Underestimation of liver-related mortality in the United States. Gastroenterology;(2013) 145:375–382
2) https://www.medicinenet.com/liver_disease/articl e.htm
3) Borroni, G.; Ceriani, R.; Cazzaniga, M.; Tommasini, M.; Roncalli, M.; Maltempo, C.; Felline, C.; Salerno, F. Comparison of simple tests for the non-invasive diagnosis of clinically silent cirrhosis in chronic hepatitis C. Aliment. Pharmacol. Ther. 2006, 24, 797–804
4) Asrani, S.K.; Devarbhavi, H.; Eaton, J.; Kamath, P.S. Burden of liver diseases in the world. J. Hepatol. 2019, 70, 151–17
5) Udell, J.A.; Wang, C.S.; Tinmouth, J.; FitzGerald, J.M.; Ayas, N.T.; Simel, D.L.; Schulzer, M.; Mak, E.; Yoshida, E.M. Does this patient with liver disease have cirrhosis? JAMA 2012, 307, 832–842
6) S. M. Mahmud, et al. "Machine Learning Based UnifiedFramework for Diabetes Prediction." Proceedings of the 2018 International Conference on Big Data Engineering and Technology. ACM (2018)
7) Singh, J.; Bagga, S.; Kaur, R. Software-based Prediction of Liver Disease with Feature Selection and Classification Techniques.Procedia Comput. Sci. 2020
8) Vijayarani, S.; Dhayanand, S. Liver disease prediction using SVM and Naïve Bayes algorithms. Int. J. Sci. Eng. Technol. Res (IJSETR) 2015, 4, 816–820.
9) Jaganathan, K.; Tayara, H.; Chong, K.T. Prediction of Drug-Induced Liver Toxicity Using SVM and Optimal Descriptor Sets. Int. J.Mol. Sci. 2021, 22, 8073
10) S. Katiyar, “Predictive analysis on diabetes, liver and kidney diseases using machine learning,” InternationalJournal for Research in Applied Science and Engineering Technology, vol. 8, no. 5, pp. 2285–2292, 2020.
11) Y.G.Robi and T.M. Sitote, “Neonatal Disease Prediction Using MLTechniques,” J Healthc Eng,vol. 2023, 2023,doi: 10.1155/2023/3567194
12) Decision Trees, Retrieve from: https://dataaspirant.com/2017/01/30/how-decision-tree algorithm-works/, Last Accessed: 5 Octobor,2019
13) Beinlich, I. A., Suermondt, H. J., Chavez, R. M., & Cooper, G. F. (1989). The ALARM monitoring system
14) Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall