Machine Learning Algorithms Based Non Alcoholic Fatty Liver Disease Prediction

Bindu Bhargavi Munukuntla, Mrutyunjaya S Yalawar

***Abstract:** The early stage liver diseases prediction is an important health related research and using this kind of research easily can predict the diseases and take the remedies. The liver diseases are classified into different types such as liver cancer, liver tumor, fatty liver, hepatitis, cirrhosis etc. Non-Alcoholic Fatty Liver Disease is a kind of chronic disease which rigorous prediction is quite difficult at early stages. The prediction of fatty liver plays significant role in treating the disease and also constraining the next health consequences. This paper presents Machine Learning Algorithms based Non Alcoholic Fatty Liver Disease (NAFLD) prediction. The main objective of this project is to identify the potential factors causing NAFLD by using Machine Learning algorithms like Decision Tree (DT) classifier, Support Vector Machine (SVM) classifier, Random Forest (RF) classifier, Logistic regression (LR). Accuracy is used parameter for performance analysis evaluation. The findings of this paper show that random forest model accurately predicts a non-alcoholic fatty liver disease patient.*

***Keywords:** Liver Disease, Classification, Machine Learning, NAFLD, Electronic Health Records.*

I. INTRODUCTION

Liver disease is any disturbance of liver function that causes illness. The liver is responsible for many critical functions within the body and should it become diseased or injured, the loss of those functions can cause significant damage to the body [1]. The development and the fundamental need of exchanging information between various academic disciplines provided an opportunity to make considerable progress in diagnosing and treating chronic disease. In this case, the research was carried out in the field of liver disorders and predicting the grades of fatty liver by using the blood test features instead of ultrasound images [2]. Non-alcoholic fatty liver disease (NAFLD) is a common and progressive disease that can lead to many liver diseases. Its prevalence has increased due to the emergence of obesity, diabetes, and hypertension. In these diseases, liver damage is not related to alcohol consumption. As it can be fatal in some cases, there are many ways to catch this condition early so that it can be treated and healed quickly. Therefore, using machine learning classification algorithms (using various biological diseases of patients to help identify and predict the nature of NAFLD) is a good way to solve these problems, and in this project we use 6 features to identify this condition in patients [3].

Metabolic syndrome, body mass index (BMI), gender, lipoprotein cholesterol (high and low), total and direct bilirubin, fibrosis formation, age, etc. The following risk factors have been found for NAFLD, such as This project will compare the performance of several of the above methods. Since NAFLD estimation is important to prevent its complications, we propose to evaluate whether a combination of blood-based biomarkers and human-centered parameters can be used to predict NAFLD in adult obesity and obesity [4].

This information is used to derive criteria for estimating outcome on a patient or out-group basis. The model then draws rules to divide the data into four groups based on the level and severity of liver disease. The collected data was used for the first time to determine the level of liver disease by taking only blood samples instead of ultrasound images. Therefore, liver doctors will have the right to early knowledge to provide liver patients with the best diagnosis and correct treatment, not to mention reducing the cost and overhead of health problems.

Remaining paper is organized as follows: Section II explains the literature survey, Section III explains described methodology, results and discussions are illustrated in Section IV and finally paper is concluded with section V.

II. LITERATURE REVIEW

R Bharath, P Rajalakshmi et al. and others. [5] proposed a new computer aided algorithm based on compressed variation-distribution coefficients and a subspace KNN classifier. Diagnosis of fatty liver is usually made with ultrasound scans. According to the fat density, the liver is divided into four groups. Sonographic tissue of the liver parenchyma varies in fat concentration and is therefore used by radiographs as a tool for fat classification. Classification of non-alcoholic fatty liver disease is very difficult for radiographers because of the slight differences in the tissue properties seen.

Manuscript received on 20 July 2023 | Revised Manuscript received on 28 July 2023 | Manuscript Accepted on 15 September 2023 | Manuscript published on 30 September 2023.

***Correspondence Author(s)**

Bindu Bhargavi Munukuntla*, Department of Computer Science and Engineering, CMR Engineering College, Hyderabad (Telangana), India. E-mail: bhargavimunukuntla26@gmail.com, ORCID ID: 0009-0004-7643-0370

Mrutyunjaya S. Yalawar, Assistant Professor, Department of Computer Science and Engineering, CMR Engineering College, Hyderabad (Telangana), India. E-mail: mrutyunjaya.cmrec20@cmrec.ac.in

The proposed algorithm separates textures with 98.8% accuracy when tested on a large dataset of 1000 images, where each class has 250 images.

Arshad I, Dutta C, Choudhury T, Thakral A, et. al. [6], Naïve Bayes algorithm used to apply on Bupa dataset. Alcohol consumption introduced as main and principle feature of the study. The simulation results represented and checked in weka software. Weka software is tool used for data mining and it is capable of working with any new developed machine learning algorithms.

Talpur N, Salleh MNM, Hussain K, et. al. [7], an investigation over ANFIS algorithm carried on with 6 different datasets including a liver data. The three different genfis methods for programming in Matlab software were evaluated and genfis3 with clustering acted with better performance.

Tavakkoli P, Souran DM, Tavakkoli S, Hatamian M, Mehrabian A, Balas VE, [8], Multi-Layer Adaptive Neuro-Fuzzy Inference System (ANFIS) Architecture Accelerator Accuracy. The simulation is given and compared. This study adopted the ANFIS model with both input and output at different stages, so the calculation and estimation error of each step is small.

Hashem EM, Mabrouk MS. meat. and others. Metin [9] reviewed the DVM algorithm A DVM method was created to maximize the value by dividing the input data into space and separating the plane. The method introduces features to space and sees the general plane, then uses a Gaussian kernel function to classify the data.This work seeks to maximize results and achieve the highest accuracy, efficiency, and complete correct answers.

Bahramirad S, Mustapha A, Eshraghi M, et al. and others. [10], this study provides information about the optimization of feature selection methods and shows how the results can be improved by using the optimization method. It is a comparative study evaluating the results of different methods. The results prove that classification tree regression (CART) and decision tree algorithms gave good results in this study.

III. METHODOLOGY

A. Non Alcoholic Fatty Liver Disease Prediction

Because we have so much data with 605 patients and 27 features, we need to check the data for possible nulls, outliers, and other raw data. The next step is preliminary data involving processing of missing values, after checking the non-value samples, we found 154 missing or worthless entries in some or other specific or all attributes and these were then subtracted from the data to get a higher-than-fact score. The data is now reduced to 451 entries, but 27 have the same characteristics. After subtracting these values, we need to model the data and set it to a certain value (usually 0-1) for faster classification. We use the Pandas python library to assist us in data management.

Feature engineering is the next step in processing the data we collect using domain knowledge (in this case medical or liver knowledge) and raw data to extract features. Consulting a doctor is another way to reduce the hassle of choosing a specialty. cirrhosis, alcohol consumption, etc. Hypothesis generation is also used, which is ambiguous as the main goal is to show humans specific nonalcoholic fatty liver disease (NAFLD), which includes eliminating factors such as In this study, machine learning algorithms such as decision tree (DT) classifier, support vector machine (SVM) classifier, random forest (RF) classifier and logistic regression (LR) were used.

Our sole goal for each algorithm we use is to improve the ROC and the accuracy with which we hyperparameter tuning. Hyperparameters are parameters such as algorithmic complexity, learning speed of the algorithm, and properties of the model. Next, we make sure of these things – we measure the quality of the data, use the data correctly, and uncover the lessons and content of the data, a step called visualization.

Another useful method we use is to visualize the data, for example we have a picture at hand when we are preparing the correlation matrix. The main factors associated with cystic fibrosis are age, weight, diabetes, and waist circumference, suggesting that NAFLD is more common in individuals when these factors are risk factors. Before designing the model, we also prepared scatterplots and histograms that gave us the opportunity to visualize the data. We use the Seaborn Python library to assist us in data visualization.

We observed that metabolic syndrome, body mass index, triglyceride, total cholesterol, age, waist to hip ratio, high density lipoprotein cholesterol, low-density lipoprotein cholesterol might be the risk factors of NAFLD and accuracy is compared for all four Machine learning algorithms (LR, DT, SVM, RF).

IV. RESULTS

While no significant fibrosis was found in 386 patients entered in the previous data, significant fibrosis was detected in 219 patients. This means that 386 of the original 605 patients had NAFLD. When we analyzed the data further, we found that the mean age of the group was 46.30, with 323 patients with a mean age of 46. There are 321 women and 284 men in the database. The average patient weight is as high as 86.40, meaning that heavier patients have a greater risk for the disease, as shown by the mean BMI of 31.87 (including obesity). Machine Learning algorithms like Decision Tree (DT) classifier, Support Vector Machine (SVM) classifier, Random Forest (RF) classifier, Logistic regression (LR) are used in this paper. We used various performance metrics.

Accuracy is simply refers to the correctness of predictions, both positive and negative, in the testing set.

The ROC Curve - AUC-ROC Curve or Receiver Operating Characteristic Curve is a metric that easily tells us that the model can distinguish between different classes. The higher the AUC (or specific area), the better the probability of a standard deviation between detection of disease and absence of disease. The curve is measured with positive and negative values. By analyzing the results of each classification, we determined the classification with the highest accuracy in estimating NAFLD.

Logistic regression is a type of linear classifier. Therefore, it misses the nonlinearity required for the distribution of activities. Thus, the AUC-ROC value is 73.81% and the AUC value is 60.71%.The logistic regression model is slightly more accurate at 77%. Due to the lack of linearity, the area under the curve in the AUC-ROC curve for logistic regression is minimal, meaning that it is possible to discriminate between classes and hence the prediction accuracy is low. However, the importance of these data in the ROC curve suggests that patient age and BMI are the most important factors in predicting NAFLD using this model, following Miletus diabetes.

Decision trees are the simplest interpreters where the inside of the tree is labeled as features. It is a non-parametric algorithm that predicts the value of the target according to the rules and then builds a tree in which the leaf nodes carry the predicted group. The AUC-ROC score of the decision tree is 64.46%. The accuracy of the model is only 58%.twenty four%. Similarly, diabetes Miletus is the most important factor predicting NAFLD in this model. Then thrombosis, age and white blood cell count.

The random forest classifier is an enhanced version of the decision tree classifier. It works by creating multiple decision trees and extracting classes based on species or average estimates of individual decision trees. Therefore it gives the best results and we can expect the most accurate results. The AUC-ROC score of 1.0 (100%) is very good. It shows the second accuracy of all the classifications we used as 85%. Thrombosis was the most important factor in the random forest distribution, followed by age, diabetes mellitus, AST, and BMI.

Supervised learning models that analyze data for classification were also used in this project. This classifier had the second highest AUC-ROC at 96.15%. The accuracy was the best at 85.7%.

Table 1: Performance Analysis

S. No.	Classifier	AUC-ROC score (%)
1	LR	73.81
2	DT	64.46
3	SVM	96.15
4	RF	100

Finally, research was conducted on the importance of Diabetes Miletus, in which we examined 225 items related to Diabetes Miletus. As 219 of these products contained fibrosis, this suggests that diabetes is an important factor in NAFLD patients. The results of this study show that machine learning classification models, especially random forest models, can predict patients with non-alcoholic fatty liver disease.

V. CONCLUSION

In this paper, Machine Learning Algorithms based Non Alcoholic Fatty Liver Disease (NAFLD) prediction is described. Non-Alcoholic Fatty Liver Disease is a kind of chronic disease which rigorous prediction is quite difficult at early stages. The prediction of fatty liver plays significant role in treating the disease and also constraining the next health consequences. The findings of this project show that machine learning classification models especially the random forest model accurately predicts a non-alcoholic fatty liver disease patient. This project also helped us to find out some undiscovered factors majorly causing NAFLD. This method may lead to greater insights for doctors to effectively identify NAFLD for novel diagnosis, and for preventive and therapeutic purposes to mitigate the global burden of NAFLD.

DECLARATION

Funding/ Grants/ Financial Support	No, I did not receive.
Conflicts of Interest/ Competing Interests	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval and consent to participate with evidence.
Availability of Data and Material/ Data Access Statement	Not relevant.
Authors Contributions	All authors having equal contribution for this article.

REFERENCES

1. A.Jaya Mabel Rani, S. Nishanthini, D.C.Jullie Josephine, Hridya Venugopal, S.Gracia Nissi, V. Jacintha, “Liver Disease Prediction using Semi Supervised based Machine Learning Algorithm”, 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC), Year: 2022
2. Lukas Brausch, Steffen Tretbar, Holger Hewener, “Identification of advanced hepatic steatosis and fibrosis using ML algorithms on high-frequency ultrasound data in patients with non-alcoholic fatty liver disease”, 2021 IEEE UFFC Latin America Ultrasonics Symposium (LAUS), Year: 2021 https://doi.org/10.1109/LAUS53676.2021.9639128
3. Michal Byra, Grzegorz Styczynski, Cezary Szmigielski, Piotr Kalinowski, Lukasz Michalowski, Rafal Paluszkiewicz, Bogna Ziarkiewicz-Wroblewska, Krzysztof Zieniewicz, Andrzej Nowicki, “Adversarial attacks on deep learning models for fatty Liver Disease classification by modification of ultrasound image reconstruction method”, 2020 IEEE International Ultrasonics Symposium (IUS), Year: 2020
4. Golmei Shaheamlung, Harshpreet Kaur, Jimmy Singla, “A Comprehensive Review of Medical Expert Systems for Diagnosis of Chronic Liver Diseases”, 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), Year: 2019 https://doi.org/10.1109/ICCIKE47802.2019.9004438
5. R Bharath, P Rajalakshmi, “Nonalcoholic Fatty Liver Texture Characterization based on Transfer Deep Scattering Convolution Network and Ensemble Subspace KNN classifier”, 2019 URSI Asia-Pacific Radio Science Conference (AP-RASC), Year: 2019 https://doi.org/10.23919/URSIAP-RASC.2019.8738717
6. Arshad I, Dutta C, Choudhury T, Thakral A, editors. “Liver disease detection due to excessive alcoholism using data mining techniques”, 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE); 2018 https://doi.org/10.1109/ICACCE.2018.8441721
7. Talpur N, Salleh MNM, Hussain K, editors. “An investigation of membership functions on performance of ANFIS for solving classification problems”, IOP Conference Series: Materials Science and Engineering; 2017 https://doi.org/10.1088/1757-899X/226/1/012103

8. Tavakkoli P, Souran DM, Tavakkoli S, Hatamian M, Mehrabian A, Balas VE, editors. “Classification of the liver disorders data using Multi-Layer Adaptive Neuro-Fuzzy Inference System (ANFIS)”. 2015 6th International Conference on Computing, Communication and Networking Technologies (ICCCNT); 2015 https://doi.org/10.1109/ICCCNT.2015.7395182
9. Hashem EM, Mabrouk MS. “A study of support vector machine algorithm for liver disease diagnosis”, American Journal of Intelligent Systems. 2014; 4(1):9-14
10. Bahramirad S, Mustapha A, Eshraghi M, editors. “Classification of liver disease diagnosis: A comparative study”, 2013 Second International Conference on Informatics & Applications (ICIA); 2013 https://doi.org/10.1109/ICoIA.2013.6650227

AUTHORS PROFILE

Bindu Bhargavi Munukuntla, Received B.Tech degree in Computer science and engineering from CMR Engineering College, Hyderabad and pursuing M.Tech in Computer science and engineering from CMR Engineering College, Hyderabad. Her area of research interest is machine learning, Artificial intelligence, Deep Learning.

Mr. Mrutyunjaya S. Yalawar working as Assistant Professor at CMR Engineering College, Medchal, Hyderabad, India and having more than 3 years of experience in IT-Industries as Software Developer and more than 7 years of Experience in Teaching Field. His Research area includes Artificial Intelligence, Machine Learning, NLP, Cyber Security, Blockchain. He published about more than 15 Papers in various National and International Journals.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)/ journal and/or the editor(s). The Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP) and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.