J. Mantas et al. (Eds.)

doi:10.3233/SHTI230543

Non-Alcoholic Fatty Liver Disease Diagnosis with Multi-Group Factors

Afrooz ARZEHGAR[a], Raheleh Ghouchan NEZHAD NOOR NIA[a],
Vajiheh DEHDELEH[a], Fatemeh ROUDI [b] and Saeid ESLAMI[a,1]
[a] *Department of Medical Informatics, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran*
[b] *Department of Nutrition, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran*
ORCiD ID: Afrooz Arzehgar https://orcid.org/0000-0003-0459-3395
Raheleh Ghouchan Nezhad Noor Nia https://orcid.org/0000-0001-5697-9041
Vajiheh Dehdeleh https://orcid.org/0000-0003-2767-7321
Fatemeh Roudi https://orcid.org/0000-0002-3570-3246
Saeid Eslami https://orcid.org/0000-0003-3755-1212

Abstract. Although various clinical factors affect the diagnosis of Non-alcoholic Fatty Liver Disease (NAFLD), most studies only use single-source data such as images or laboratory data. Nevertheless, using different categories of features can help to get better results. Hence, one of the most important purposes of this paper is to employ a multi-group of effective factors such as velocimetry, psychological, demographic and anthropometric, and lab test data. Then, some Machine Learning (ML) methods are applied to classify the samples into two healthy and patient with NAFLD groups. The data used here belongs to the PERSIAN Organizational Cohort study at Mashhad University of Medical Sciences. To quantify the scalability of the models, different validity metrics are used. The obtained results illustrate that the proposed method can lead to an increase in the efficiency of the classifiers.

Keywords. NAFLD Diagnose, Classification, Multi-Group Factors

1. Introduction

Diagnostic technologies such as imaging tests, blood tests, and liver biopsies can help diagnose NAFLD [1]. However, these tests are often invasive, expensive, and time-consuming. Some strategies have been presented by applying ML algorithms using non-invasive methods such as image-based methods and other biomarkers [2,3] but other important factors are not considered. Nevertheless, using different groups of features can help to have an efficient model. Hence, we aim to use multi-source features. To handle the behavior of outliers and missing data, we use some techniques in the preprocessing. Then some ML methods are employed to classify the samples into two healthy and patient with NAFLD groups. The rest of this study is organized as follows. The proposed approach is explained in Section 2. The experimental results and conclusions are presented in Sections 3 and 4 respectively.

[1] Corresponding Author: Saeid ESLAMI, E-mail: S.Eslami.H@gmail.com.

2. The Proposed Method

The proposed strategy consists of two main phases: feature engineering and classification. In the first phase, some new groups of features are introduced. In the second phase, some ML methods are employed to classify the data into two healthy and patient with NAFLD groups. Liver function is associated with obesity and metabolic syndrome, and aging [4,5]. Therefore, we use anthropometric information such as BMI and demographic information like gender and age. In addition, it is shown in [6] that there is a relationship between the liver condition and the psychological situation. For instance, depression has been reported to be correlated with NAFLD [7,8]. Hence, such a relationship can be considered as an additional feature. To pre-pare psychology data, the questionnaire is self-administered by the participants and includes the following 3 subscales: Depression, anxiety, and stress (DASS-21), Sleep quality (PSQI), and Occupational stress (administrative personnel SOS and medical personnel HSS-35). Any disruption in liver function can affect blood flow and alter the velocimetry data. For example, patients with NAFLD often exhibit multiple cardiovascular risk factors [9] which can lead to changes in hepatic blood flow velocity [10]. In order to conduct the velocimetry test, a trained physician used the SphygmoCorXCEL electrical module (Illinois, USA). Compared with the image-based diagnosis alone, the image and laboratory test-based combination model is more effective [11,12]. In laboratory tests, to perform cell counting, biochemistry, hormonal, urinalysis, and rapid fecal occult blood (FOB) tests, sampling is done after 10 to 12 hours of fasting. In such cases, Liver Craniocaudal Diameter is measured using imaging. After collecting the data, outlier data are detected based on comparing percentiles for each predictor and the median imputation is used to deal with missing data. A total of 47 variables are used as predictors. To find the best model, XGboost, Support Vector Machines (SVM), and Neural Networks (NNs) are used [13,14]. The efficiency of the algorithms is compared based on the confusion matrix and some indicators including precision, accuracy, recall, F-1. The Flowchart of the proposed diagnosis model is shown in Figure 1.

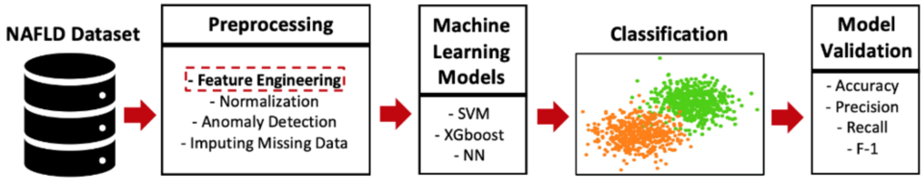

Figure 1. Flowchart of the diagnosis model.

3. Results

The dataset belongs to Mashhad Cohort Center in Iran. Ethics committee approval was received for this study from Mashhad University of Medical Sciences. Ethics committee approval number is IR.MUMS.REC.1401.354 and Dr. Mohammad Ali Kiani is the chairperson. The total number of samples in this case study is 1118, 42.94% (480) of whom have NAFLD. Before applying the models to the data, we use anomaly detection technique and two strategies removal and imputation (median) are applied to deal with NaN values. To diagnose NAFLD, we use a NNs, XGboost, and SVM. In the NN model, we use Grid Search to find the optimal parameters. The results of these tests are listed in the next table. As can be seen from Table 1, the XGboost model with the imputation strategy has the best result. By employing all features as predictors for XGboost, it has the best accuracy (Table 2).

Table 1. The accuracy of ML methods according to different ways of dealing with missing data

Ways	Outlier removal	NN	XGBoost	SVM
Impute-Median	Yes	0.82	0.85	0.72
	No	0.73	0.72	0.71
Missing values removal	Yes	0.80	0.83	0.71
	No	0.79	0.71	0.70

Table 2. Evaluating different categories of predictors on XGboost

Predictors	Macro average			
	Precision	Accuracy	Recall	F1
Demographic and anthropometric	0.74	0.73	0.76	0.73
Laboratory	0.63	0.63	0.63	0.62
Psychology	0.45	0.46	0.49	0.41
Velocimetry	0.67	0.67	0.65	0.65
All	0.84	0.85	0.89	0.83

4. Conclusions

It is crucial to discuss how to optimally apply different non-invasive assessment techniques and present a comprehensive assessment to diagnose NFLD. Most studies in this application only use images or laboratory data. Nevertheless, using different groups of features can help to have a more applicable model. Hence, in this study, we utilize a multi-group of effective factors. The data belongs to the Cohort study at Mashhad University of Medical Sciences. Then some ML models are employed to classify the data into two healthy and patient with NAFLD groups. Since XGBoost is a gradient boosting algorithm that uses an ensemble of decision trees trained in parallel, resulting in faster and more efficient convergence, superior handling of missing values, and easier parameter tuning compared to SVM and NNs in specific situation. The results show that XGBoost has the highest accuracy by applying the proposed categories of features.

References

[1] Byrne CD, Targher G. How should endocrinologists diagnose and treat non-alcoholic fatty liver disease? Lancet Diabetes Endocrinol [Internet]. 2022 Jul 1 [cit-ed 2023 Mar 22];10(7):478–80. Available from: http://www.thelancet.com/article/S221385872200167X/fulltext

[2] Yu JH, Lee HA, Kim SU. Noninvasive imaging biomarkers for liver fibrosis in nonalcoholic fatty liver disease: current and future. Clin Mol Hepatol [Internet]. 2023 Feb 28 [cited 2023 Mar 22];29(Suppl). Available from: https://pubmed.ncbi.nlm.nih.gov/36503205/

[3] Samaddar P, Mishra AK, Gaddam S, Singh M, Modi VK, Gopalakrishnan K, et al. Machine Learning-Based Classification of Abnormal Liver Tissues Using Relative Permittivity. Sensors (Basel) [Internet]. 2022 Dec 1 [cited 2023 Mar 22];22(24). Available from: https://pubmed.ncbi.nlm.nih.gov/36560303/

[4] Porukala M, Vinod PK. Network-level analysis of ageing and its relationship with diseases and tissue regeneration in the mouse liver. Sci Rep [Internet]. 2023 Mar 21 [cited 2023 Mar 23];13(1):4632. Available from: https://pubmed.ncbi.nlm.nih.gov/36944690/

[5] Buzova D, Maugeri A, Liguori A, Napodano C, Lo Re O, Oben J, et al. Circulating histone signature of human lean metabolic-associated fatty liver disease (MAFLD). Clin Epigenetics [Internet]. 2020 Aug 20 [cited 2023 Mar 23];12(1):1–15. Available from: https://clinicalepigeneticsjournal.biomedcentral.com/articles/10.1186/s13148-020-00917-2

[6] Goulart AC, Bianchi LLT, Bismarchi D, Miname MH, Lourenção ACM, Henares BB, et al. Sex differences in the relationship between hepatic steatosis, mood and anxiety disorders. J Psychosom Res [Internet]. 2023 May 1 [cited 2023 Mar 22];168. Available from: https://pubmed.ncbi.nlm.nih.gov/36913766/

[7] Ntona S, Papaefthymiou A, Kountouras J, Gialamprinou D, Kotronis G, Boziki M, et al. Impact of nonalcoholic fatty liver disease-related metabolic state on depres-sion. Neurochem Int [Internet]. 2023 Feb 1 [cited 2023 Mar 22];163. Available from: https://pubmed.ncbi.nlm.nih.gov/36634820/

[8] Gu Y, Zhang W, Hu Y, Chen Y, Shi J. Association between nonalcoholic fatty liver disease and depression: A systematic review and meta-analysis of observation-al studies. J Affect Disord [Internet]. 2022 Mar 15 [cited 2023 Mar 22];301:8–13. Available from: https://pubmed.ncbi.nlm.nih.gov/34986375/

[9] Zhang S, Mak LY, Yuen MF, Seto WK, Kasper P, Demir M, et al. Screening strat-egies for non-alcoholic fatty liver disease: a holistic approach is needed. Clin Mol Hepatol [Internet]. 2023 Mar 20 [cited 2023 Mar 22]; Available from: http://www.e-cmh.org/journal/view.php?doi=10.3350/cmh.2023.0059

[10] Chipperfield AJ, Thanaj M, Scorletti E, Byrne CD, Clough GF. Multi-domain anal-ysis of microvascular flow motion dynamics in NAFLD. Microcirculation [Inter-net]. 2019 Jul 1 [cited 2023 Mar 23];26(5). Available from: https://pubmed.ncbi.nlm.nih.gov/30803094/

[11] Cen C, Wang W, Yu S, Tang X, Liu J, Liu Y, et al. Development and validation of a clinical and laboratory-based nomogram to predict nonalcoholic fatty liver dis-ease. Hepatol Int [Internet]. 2020 Sep 1 [cited 2023 Mar 23];14(5):808–16. Availa-ble from: https://pubmed.ncbi.nlm.nih.gov/32572817/

[12] Xu B, Zhou NM, Cao WT, Li XJ. Evaluation of elastography combined with sero-logical indexes for hepatic fibrosis in patients with chronic hepatitis B. World J Gastroenterol [Internet]. 2018 Oct 7 [cited 2023 Mar 23];24(37):4272–80. Available from: https://pubmed.ncbi.nlm.nih.gov/30310260/

[13] Avolio M, Fuduli A. A Semiproximal Support Vector Machine Approach for Binary Multiple Instance Learning. IEEE Trans Neural Netw Learn Syst. 2021 Aug;32(8):3566–77.

[14] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. 2016 Mar;13-17-August-2016:785–94. Available from: https://arxiv.org/abs/1603.02754v3.