Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth
A. Ugon et al. (Eds.)

doi:10.3233/978-1-61499-852-5-166

Applications of Machine Learning in Fatty Live Disease Prediction

Md.Mohaimenul Islam [a, b], Chieh-Chen Wu [a, b], Tahmina Nasrin Poly [a, b], Hsuan-Chia Yang [b], Yu-Chuan (Jack) Li [a, b, c,1]

[a] *Graduate Institute of Biomedical Informatics, Taipei Medical University, Taipei, Taiwan*

[b] *International Center for Health Information Technology (ICHIT), Taipei Medical University, Taipei, Taiwan*

[C] *Department of Dermatology, Wan Fang Hospital, Taipei, Taiwan*

Abstract: Fatty liver disease (FLD) is considered the most prevalent form of chronic liver disease worldwide. The prediction of fatty liver disease is an important factor for effective treatment and reduce serious health consequences. We, therefore construct a prediction model based on machine learning algorithms. A dataset was developed with ten attributes that included 994 liver patients in which 533 patients were females and others were male. Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN), and Logistic Regression (RF) data mining technique with 10-fold cross-validation was used in the proposed model for the prediction of fatty liver disease. The performances were evaluated with accuracy, sensitivity, specificity, positive predictive value, and negative predictive value. In this proposed model, logistic regression technique provides a better result (Accuracy 76.30%, sensitivity 74.10%, and specificity 64.90%) among all other techniques. This study demonstrates that machine learning models particularly logistic regression model provides a higher accurate prediction for fatty liver diseases based on medical data from electronic medical. This model can be used as a valuable tool for clinical decision making.

1. Introduction

Fatty liver disease (FLD) is one of the major cause of liver disease worldwide which eventually lead to noncholestatic cirrhosis and hepatocellular carcinoma [1]. The prevalence of FLD has been increased and appear to pose a greater economic burden. However, the prevalence of FLD has been increasing in parallel with the prevalence of diabetes, metabolic syndrome and obesity [2]. In the United States, the prevalence of FLD

[1] Corresponding author : jaak88@gmail.com, jack@tmu.edu.tw

by ultrasound is 10 to 46 percent, but most biopsy-based studies reported the prevalence of FLD is only 3 to 5 percent [3, 4]. Although biopsy is considered as a good standard, it might make side effects and sampling errors during the application of this method.

Ultrasonography has now been used as a functional tool for FLD diagnosis with higher accuracy, whereas identifying accuracy is highly operator dependent [5]. Machine learning has been playing a critical role in medical decision making and it specializes in the integration of multiple risk factors into a predictive tool [6, 7]. Being a promising tool for interference in medicine, it is helping physicians to treat patients more precisely. To date, the benefits of utilizing machine learning with data from electronic medical record to predict FDL has not been evaluated on a large scale. We, therefore aimed at constructing a machine learning model to predict fatty liver disease.

2. Methods:

Our method is composed of seven steps.

1. *Data collection* : data were collected from Taipei Medical University Hospital under a liver protection project. database included the information of 994 patients during the study period between January 1, 2012, and December 31, 2013. Patients aged less than 30 years with incomplete examination procedure were excluded from our study. Ultrasonography test was used to identify fatty liver disease patient. This study was reviewed and approved by the institutional ethical committee board of Taipei Medical University and Taipei Medical University Hospital, conducted in accord with the ethical guidelines of the Declaration of Helsinki of the world medical association.
2. *Machine learning strategies* : The aim of the study was to select prognostic factors to predict fatty liver disease using classification machine learning algorithms.
3. *Data preprocessing* : we removed all those variable containing more than 50% missing value. In addition, data imputation and normalization is needed to get a high-quality dataset. SMOTE over-sampling method was used to generate synthesis samples for the minority class and balanced the positive and negative training set.
4. *Machine learning model selection:* four classification algorithms such as RF, SVM, ANN, and LR were applied. We used 10-fold cross-validation to trained and evaluated training datasets.
5. *Feature selection:* Data with extremely high dimensionality has presented serious challenges to existing learning methods [8]. Due to a large number of features, it may tend to over fit that cause decrease performance of the model. Features selection for classification model attempts to select minimally sized subset according to following criteria : (1) The classification accuracy should have to increase; (2) The values for the selected features should have to close as possible to the original class distribution.
6. *Model development and validation:* Weka 3.7 was used to construct data mining algorithms
7. *Model assessment:* The confusion matrix has been used to determine the relationship between the actual values and predicted values [9]. Table 1 shows the structure of confusion matrix.

Table 1: Confusion matrix representation

	Positive	Negative
Predicted true (+)	TP	TN
Predicted false (-)	FP	FN

Following quality parameters were used to evaluate the results

(1) Accuracy = TP+TN/ TP+FP+TN+FN
(2) Sensitivity = TP/ TP+FN
(3) Specificity = TN/TN+FP

3. Results:

3.1. Patient's characteristics:

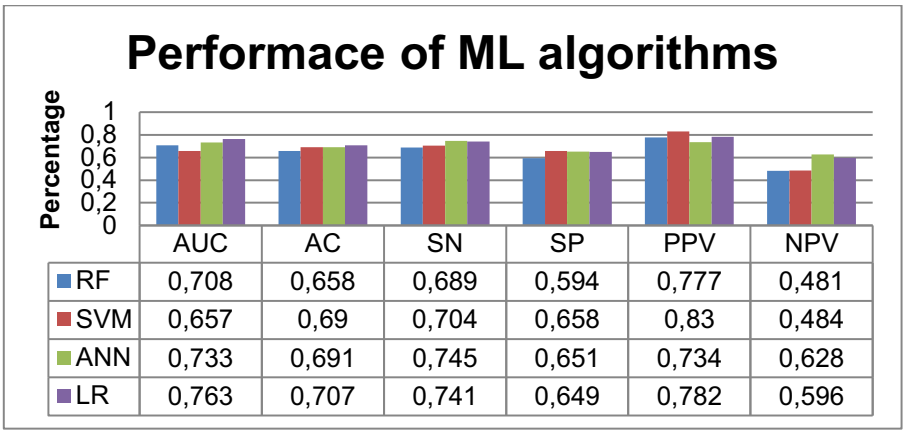
Table 2 shows the demographic and clinical characteristics of overall 994 patients. The age of the patients with FLD was 62.10±12.55 and NFLD was 62.07±13.52. Compared to non-fatty liver group, FL group had higher frequency of BMI, GOT-AST, GPT-ALT, and triglyceride. There were significant differences between two groups in GOT-AST, GPT-ALT, and triglyceride level ($p<0.01$).

Table 2: Demographic characteristics of abdominal ultrasonography diagnostic groups

Patient variable	Fatty liver (593)	Non-fatty liver (401)	P-value
Age (Mean age, y)	62.10±12.55	62.07±13.52	0.949
Gender	M:288/F:305	M: 173/F: 228	0.092
BMI	25.88±3.958	22.64±2.918	<0.001
Cholesterol	184.03±35.21	186.66±34.16	0.243
HDL-C	53.15±11.58	58.02±12.43	<0.001
LDL-C	111.54±22.99	112.49±24.89	0.535
Glucose AC	116.32±33.81	112.19±33.901	0.060
GOT-AST	39.67±101.07	28.99±25.56	0.014
GPT-ALT	41.70±84.97	28.10±31.84	<0.001
Triglyceride	136.38±109.69	106.13±53.71	<0.001

3.2. Performance of machine learning algorithms:

In our proposed model, we predicted the whole dataset using 10-fold cross-validation and evaluated the performance on FLD by AUC, accuracy, sensitivity, specificity, positive predictive value, and negative predictive value. Figure 1 shows the performance of the predictive model using different data mining algorithm techniques. The AUC of random forest (RF), support vector machine (SVM), artificial neural network (ANN), and logistic regression (LR) were 0.708, 0.657, 0.7333, and 0.763 respectively. Logistic regression model showed higher performance (AUC = 0.763, AC= 0.70, SN = 0.741, SP = 0.649, PPV = 0.782, NPV= 0.596) than other machine learning models.



	AUC	AC	SN	SP	PPV	NPV
RF	0,708	0,658	0,689	0,594	0,777	0,481
SVM	0,657	0,69	0,704	0,658	0,83	0,484
ANN	0,733	0,691	0,745	0,651	0,734	0,628
LR	0,763	0,707	0,741	0,649	0,782	0,596

Figure 1 Performance comparison of different machine learning algorithms on fatty liver disease prediction

*Note: RF= Random Forest, SVM= Support Vector Machine, ANN= Artificial Neural Network, LR= Logistic Regression. AUC= Area under the ROC Curve, AC= Accuracy, SN= Sensitivity, SP= Specificity, PPV= Positive Predictive Value, NPV= Negative Predictive Value.

4. Discussion:

In this study, we used various machine learning algorithms to improve prediction of FLD that provided significant insights comparted with traditional statistical models. In this proposed model, logistic regression model showed higher performance with C statistic 0.763. To our knowledge, this is the first study attempted to predict FLD using various machine learning algorithms. There are many kind of machine learning algorithms have been developed along with the most popular Bayesian algorithm, it is hard to make a proper algorithm for clinical decision making and clinical practices [10]. Therefore, the performances of different algorithms are the most important consideration, along with the easy to use and the interpretation of the models. However, our model could effectively detect fatty liver disease (FLD) for anyone by initial screening without using abdominal ultrasonography. In addition, the model could provide an easy, fast, low cost, and non-invasive method to accurately diagnose FLD [11].

As the healthcare data has been increased day by day and machine learning allow massive amounts of data to be analyzed rapidly [12]. Therefore, it is the opportunity to apply machine learning algorithms to the care of individual patients in medical practice. Using various machine learning prediction models, physicians could be able to extract the minimum data necessary to make a therapeutic decision [13]. Our model has the potential to early FLD detection that will help to improve precise and appropriate treatment pattern. It is very important for physicians to know about the most predictive variables for best treatment outcome. Patient's baseline characteristics might be the strongest predictors of FDL for evaluation of the individual patient level [14]. Therefore, we carefully adopted a feature selection strategy and used 10-fold cross-validation to repeatedly screen potential variables. Data were included from a medical center EMR without additional clinical assessments, and our high-performance prediction model could be easily integrated into EMR to identify FLD risk. Our model could help to identify FLD patients that might

significantly impact on treatment pattern. Early prediction using this model might bring benefits from treatment reduction, and medical cost decrease.

5. Conclusion:

In this study, machine learning techniques were used to predict fatty liver disease, and logistic regression model showed better performance than other classification techniques. This prediction outcome has the potential to help clinicians make more precise and meaningful decisions about fatty liver disease diagnosis and treatment.

References:

[1] M. Lazo, J.M. Clark, The epidemiology of nonalcoholic fatty liver disease: a global perspective. Seminars in liver disease (2008), 339-350.
[2] M.H. Le, P. Devaki, N.B. Ha, D.W. Jun, H.S. Te, R.C. Cheung, M.H. Nguyen, Prevalence of non-alcoholic fatty liver disease and risk factors for advanced fibrosis and mortality in the United States, PloS one 12 (2017), e0173499.
[3] C.D. Williams, J. Stengel, M.I. Asike, D.M. Torres, J. Shaw, M. Contreras, C.L. Landt, S.A. Harrison, Prevalence of nonalcoholic fatty liver disease and nonalcoholic steatohepatitis among a largely middle-aged population utilizing ultrasound and liver biopsy: a prospective study, Gastroenterology 140 (2011), 124-131.
[4] S.G. Sheth, S. Chopra, Epidemiology, clinical features, and diagnosis of nonalcoholic fatty liver disease in adults, Waltham (MA): UpToDate 2017.
[5] Q.M. Anstee, G. Targher, C.P. Day, Progression of NAFLD to diabetes mellitus, cardiovascular disease or cirrhosis, Nature Reviews Gastroenterology and Hepatology 10 (2013), 330-344.
[6] W. Raghupathi, V. Raghupathi, Big data analytics in healthcare: promise and potential, Health information science and systems (2014), 2-3.
[7] P. Groves, B. Kayyali, D. Knott, S.V. Kuiken, The big data revolution in healthcare: Accelerating value and innovation, 2016.
[8] J. Tang, S. Alelyani, H. Liu, Feature selection for classification: A review, Data Classification: Algorithms and Applications (2014), 37.
[9] R. Kohavi, F. Provost, Glossary of terms. Machine Learning 30 (1998), 271-274.
[10] J. Wu, J. Roy, W.F. Stewart, Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches, Medical care 48 (2010), S106-S113.
[11] J. Kang, T. Lee, I. Yap, K. Lun, Analysis of cost-effectiveness of different strategies for hepatocellular carcinoma screening in hepatitis B virus carriers. Journal of gastroenterology and hepatology **7** (1992), 463-468.
[12] T. Condie, P. Mineiro, N. Polyzotis, M. Weimer, Machine learning on big data. *Data Engineering (ICDE)*, 2013 IEEE 29th International Conference on. IEEE (2013), 1242-1244.
[13] T.B. Murdoch, A.S. Detsky, The inevitable application of big data to health care, *Jama* **309** (2013), 1351-1352.
[14] G.K. Savova, P.V. Ogren, P.H. Duffy, J.D. Buntrock, C.G. Chute, Mayo clinic NLP system for patient smoking status identification, *Journal of the American Medical Informatics Association* **15** (2008), 25-28.