Contents

EDITORIAL

1203  Transition of an acronym from nonalcoholic fatty liver disease to metabolic dysfunction-associated fatty liver disease
Alam S, Fahim SM

OPINION REVIEW

1208  Non-invasive real-time assessment of hepatic macrovesicular steatosis in liver donors: Hypothesis, design and proof-of-concept study
Rajamani AS, Rammohan A, Sai VR, Rela M

REVIEW

1215  Impact of COVID-19 pandemic on liver, liver diseases, and liver transplantation programs in intensive care units
Omar AS, Kaddoura R, Orabi B, Hanoura S

1234  In the era of rapid mRNA-based vaccines: Why is there no effective hepatitis C virus vaccine yet?
Echeverría N, Comas V, Aldunate F, Perbolianachis P, Moreno P, Cristina J

1269  Pediatric non-cirrhotic portal hypertension: Endoscopic outcome and perspectives from developing nations
Sarma MS, Seetharaman J

MINIREVIEWS

1289  Acute-on-chronic liver failure in children
Islek A, Tunçgor G

1299  Coronavirus disease 2019 in liver transplant patients: Clinical and therapeutic aspects
Loinaz-Segurola C, Marcacuzco-Quinto A, Fernández-Ruíz M

1316  Pediatric vascular tumors of the liver: Review from the pathologist’s point of view
Cordier F, Hooeens A, Van Dorpe J, Creytens D

1328  Autoimmune hepatitis in genetic syndromes: A literature review
Capra AP, Chiara E, Briguia S

1341  Assessing the prognosis of cirrhotic patients in the intensive care unit: What we know and what we need to know better
da Silveira F, Soares PHR, Marchesan LQ, da Fonseca RSA, Nedel WL

1351  Liver transplantation for pediatric inherited metabolic liver diseases
Vimalesvaran S, Dhawan A
Contents

Liver and COVID-19: From care of patients with liver diseases to liver injury
Gaspar R, Castelo Branco C, Macedo G

ORIGINAL ARTICLE

Basic Study
Direct modulation of hepatocyte hepcidin signaling by iron
Yu LN, Wang SJ, Chen C, Rausch V, Elshaarawy O, Mueller S

Serum zonulin levels in patients with liver cirrhosis: Prognostic implications
Voulgaris TA, Karagiannakis D, Hadziyannis E, Manolakopoulos S, Karamanolis GP, Papatheodoridis G, Vlachogiannakos J

Retrospective Cohort Study
Impact of biliary complications on quality of life in live-donor liver transplant recipients

Retrospective Study
Machine learning models for predicting non-alcoholic fatty liver disease in the general United States population: NHANES database
Atsawarungruangkit A, Laoveeravat P, Promrat K

Acute liver failure with hemolytic anemia in children with Wilson’s disease: Genotype-phenotype correlations?
Pop TL, Grama A, Stefanescu AC, Willheim C, Ferenci P

Observational Study
Clinical outcomes of patients with two small hepatocellular carcinomas

CASE REPORT
Focal nodular hyperplasia associated with a giant hepatocellular adenoma: A case report and review of literature
Gaspar-Figueiredo S, Kefleyesus A, Sempoux C, Uldry E, Halkic N
ABOUT COVER
Editorial Board Member of *World Journal of Hepatology*, Julio Coelho, MD, PhD, Professor, Department of Surgery, Federal University of Parana, Curitiba 80240-110, Parana, Brazil. coelhojcu@yahoo.com.br

AIMS AND SCOPE
The primary aim of *World Journal of Hepatology* (WJH, *World J Hepatol*) is to provide scholars and readers from various fields of hepatology with a platform to publish high-quality basic and clinical research articles and communicate their research findings online.

WJH mainly publishes articles reporting research results and findings obtained in the field of hepatology and covering a wide range of topics including chronic cholestatic liver diseases, cirrhosis and its complications, clinical alcoholic liver disease, drug induced liver disease autoimmune, fatty liver disease, genetic and pediatric liver diseases, hepatocellular carcinoma, hepatic stellate cells and fibrosis, liver immunology, liver regeneration, hepatic surgery, liver transplantation, biliary tract pathophysiology, non-invasive markers of liver fibrosis, viral hepatitis.

INDEXING/ABSTRACTING
The WJH is now abstracted and indexed in PubMed, PubMed Central, Emerging Sources Citation Index (Web of Science), Scopus, China National Knowledge Infrastructure (CNKI), China Science and Technology Journal Database (CSTJ), and Superstar Journals Database. The 2021 edition of Journal Citation Reports® cites the 2020 Journal Citation Indicator (JCI) for WJH as 0.61. The WJH’s CiteScore for 2020 is 5.6 and Scopus CiteScore rank 2020: Hepatology is 24/62.

RESPONSIBLE EDITORS FOR THIS ISSUE
Production Editor: Xu Guo; Production Department Director: Xiang Li; Editorial Office Director: Xiang Li.

NAME OF JOURNAL
*World Journal of Hepatology*

ISSN
ISSN 1948-5182 (online)

LAUNCH DATE
October 31, 2009

FREQUENCY
Monthly

EDITORS-IN-CHIEF
Nikolaos Pyrsopoulos, Ke-Qin Hu, Koo Jeong Kang

EDITORIAL BOARD MEMBERS

PUBLICATION DATE
October 27, 2021

COPYRIGHT
© 2021 Baishideng Publishing Group Inc

INSTRUCTIONS TO AUTHORS
https://www.wjgnet.com/bpg/gerinfo/204

GUIDELINES FOR ETHICS DOCUMENTS
https://www.wjgnet.com/bpg/GerInfo/287

GUIDELINES FOR NON-NATIVE SPEAKERS OF ENGLISH
https://www.wjgnet.com/bpg/gerinfo/240

PUBLICATION ETHICS
https://www.wjgnet.com/bpg/gerinfo/288

PUBLICATION MISCONDUCT
https://www.wjgnet.com/bpg/gerinfo/208

ARTICLE PROCESSING CHARGE
https://www.wjgnet.com/bpg/gerinfo/242

STEPS FOR SUBMITTING MANUSCRIPTS
https://www.wjgnet.com/bpg/gerinfo/239

ONLINE SUBMISSION
https://www.f6publishing.com
Retrospective Study

Machine learning models for predicting non-alcoholic fatty liver disease in the general United States population: NHANES database

Amporn Atsawarungruangkit, Passisd Laoveeravat, Kittichai Promrat

ORCID number: Amporn Atsawarungruangkit 0000-0003-0622-6839; Passisd Laoveeravat 0000-0001-6855-0437; Kittichai Promrat 0000-0002-4003-2598.

Author contributions: Atsawarungruangkit A and Laoveeravat P contributed equally to this work including study design, data analysis, result interpretation, and manuscript writing. Promrat K critically revised the manuscript and provided supervision.

Institutional review board statement: The National Health and Nutrition Examination Survey protocol was approved by the National Center for Health Statistics Research Ethics Review Board (Hyattsville, MD, United States).

Informed consent statement: In NHANES III, the consent form was signed by participants in the survey.

Conflict-of-interest statement: No conflict of interest exists.

Data sharing statement: The dataset used in this manuscript is NHANES III, which is publicly available dataset.

Open-Access: This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

Abstract

BACKGROUND
Non-alcoholic fatty liver disease (NAFLD) is the most common chronic liver disease, affecting over 30% of the United States population. Early patient identification using a simple method is highly desirable.

AIM
To create machine learning models for predicting NAFLD in the general United States population.

METHODS
Using the NHANES 1988-1994. Thirty NAFLD-related factors were included. The dataset was divided into the training (70%) and testing (30%) datasets. Twenty-four machine learning algorithms were applied to the training dataset. The best-performing models and another interpretable model (i.e., coarse trees) were tested using the testing dataset.

RESULTS
There were 3235 participants (n = 3235) that met the inclusion criteria. In the training phase, the ensemble of random undersampling (RUS) boosted trees had the highest F1 (0.53). In the testing phase, we compared selective machine learning models and NAFLD indices. Based on F1, the ensemble of RUS boosted trees remained the top performer (accuracy 71.1% and F1 0.56) followed by the fatty liver index (accuracy 68.8% and F1 0.52). A simple model (coarse trees) had...
open-access article that was selected by an in-house editor and fully peer-reviewed by external reviewers. It is distributed in accordance with the Creative Commons Attribution NonCommercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/Licenses/by-nc/4.0/

Manuscript source: Invited manuscript

Specialty type: Gastroenterology and hepatology

Country/Territory of origin: United States

Peer-review report’s scientific quality classification

Grade A (Excellent): 0
Grade B (Very good): B
Grade C (Good): 0
Grade D (Fair): 0
Grade E (Poor): 0

Received: March 7, 2021
Peer-review started: March 7, 2021
First decision: May 2, 2021
Revised: May 11, 2021
Accepted: September 19, 2021
Article in press: September 19, 2021
Published online: October 27, 2021

P-Reviewer: Wu SZ
S-Editor: Gao CC
L-Editor: A
P-Editor: Liu JH

an accuracy of 74.9% and an F1 of 0.33.

CONCLUSION

Not every machine learning model is complex. Using a simpler model such as coarse trees, we can create an interpretable model for predicting NAFLD with only two predictors: fasting C-peptide and waist circumference. Although the simpler model does not have the best performance, its simplicity is useful in clinical practice.

Key Words: Artificial intelligence; Machine learning; Non-alcoholic fatty liver disease; Fatty liver; United States population; NHANES

©The Author(s) 2021. Published by Baishideng Publishing Group Inc. All rights reserved.

Core Tip: A simple method with a good accuracy for identifying patients with non-alcoholic fatty liver disease is highly desirable. Among 24 machine learning models, the ensemble of random undersampling boosted trees was the top performer (accuracy 71.1% and F1 0.56). A simple model (coarse trees) with only two predictors (fasting C-peptide and waist circumference) had an accuracy of 74.9% and an F1 of 0.33. Not every machine learning model is complex. Using a simple model such as coarse trees, physicians can easily integrate machine learning model into their practice without any software implementation.

Citation: Atsawarungruangkit A, Laoveeravat P, Promrat K. Machine learning models for predicting non-alcoholic fatty liver disease in the general United States population: NHANES database. World J Hepatol; 2021; 13(10): 1417-1427
URL: https://www.wjgnet.com/1948-5182/full/v13/i10/1417.htm
DOI: https://dx.doi.org/10.4254/wjh.v13.i10.1417

INTRODUCTION

Non-alcoholic fatty liver disease (NAFLD) is a common chronic metabolic disease found in 25.5% of the United States population, and it is more common in patients with diabetes (55.5%), leading to a health and economic burden[1-3]. Non-alcoholic steatohepatitis (NASH) can lead to liver-related consequences, such as cirrhosis, hepatocellular carcinoma, and mortality. NASH is the second most common indication for liver transplantation in the United States and is likely to replace hepatitis C infection as the leading cause of liver transplantation in the future[4]. NAFLD is diagnosed primarily with imaging studies, transient elastography, magnetic resonance elastography, or liver biopsy[5]. Some of these diagnostic modalities are not available in every health care facility, require expert interpretation, and are invasive in case of biopsy[5,6]. To prevent adverse outcomes in these patients, early screening and detection based on risk factors are warranted. Healthcare providers and patients are aware of the risk factors of NAFLD, which include diabetes, obesity, dyslipidemia, and metabolic syndrome[5,7,8]. However, there is no well-performing tool for the early prediction of NAFLD; for example, liver enzyme levels can be normal in patients with NAFLD[9,10]. There are existing studies on the risk factors and prediction risk scores; however, their results are controversial[11-13]. Machine learning is a potential approach for the identification of the best predictive model[14,15].

Machine learning can be used to construct a predictive model by teaching computer algorithms to learn from data without being explicitly programmed. Applications of machine learning in gastroenterology field are steadily increasing[17]. However, there is no machine learning model for predicting NAFLD in the United States. The published models in China, Germany, and Canada focus on NAFLD prediction scores using laboratory parameters and demographic data[11,13-15]. Therefore, we aimed to evaluate the applications of machine learning in NAFLD diagnosis for easy use at clinical setting.
MATERIALS AND METHODS

Study population and study design
The Third National Health and Nutrition Examination Survey (NHANES III) was a nationwide probability sample of 39,695 persons aged 2 mo and older, conducted from 1988-1994 by the National Center for Health Statistics (NCHS). It aimed to evaluate the health and nutritional status of the general United States population.[18] Multiple datasets were collected in this survey, including demographics, interviews, physical examinations, and laboratory testing of biologic samples. The NHANES protocol was approved by the NCHS Research Ethics Review Board.

Definitions
Participants aged 20 years or older in NHANES III with gradable ultrasound results were included in this study. The exclusion criteria included: (1) Excessive alcohol consumption; (2) Hepatitis B or C infection; (3) Fasting period outside of 8-24 h; and (4) Incomplete or missing data on physical examination and laboratory testing. The participants were divided into two groups: The NAFLD participants and non-NAFLD participants. Since participants aged above 74 years were not eligible for ultrasonography in NHANES III, participants aged above 74 years were excluded from this study. ‘NAFLD participants’ was defined based on: (1) Moderate to severe hepatic steatosis on ultrasound; (2) No history of alcohol drinking more than 2 drinks per day for men or 1 drink per day for women in the last 12 mo; and (3) No history of hepatitis B or C infection.

Thirty factors associated with NAFLD were included in this study: demographic (i.e., age, gender, and race/ethnicity), body measurement [i.e., body mass index (BMI) and waist circumference], general biochemistry tests [i.e., iron, total iron-binding capacity, transferrin saturation, ferritin, cholesterol, triglyceride, high-density lipoprotein (HDL) cholesterol, C-reactive protein, and uric acid], liver chemistry (aspartate aminotransferase, alanine aminotransferase, gamma glutamyl transferase, alkaline phosphatase, total bilirubin, total protein, albumin, and serum globulin), diabetes testing profile [i.e., glycated hemoglobin, fasting plasma glucose, fasting C-peptide, and fasting insulin], and the use of diabetes medication.

Statistical analysis
Categorical and ordinal factors are presented as frequencies (%). Continuous factors are presented as medians (interquartile ranges). The dataset was divided into the training (70%) and testing (30%) datasets using stratified sampling. Differences between the two datasets were tested using the Mann-Whitney U test. Twenty-four machine learning algorithms were applied to the training dataset. Then, we selected the best performing models determined by accuracy and the F1 score and compared the out-of-sample performance with another interpretable model (coarse trees, a decision tree model with a maximum of four splits) and three NAFLD indices on the testing dataset. The selected NAFLD indices included fatty liver index (FLI), hepatic steatosis index (HSI), and triglyceride and glucose index (TyG)[19-21]. The cut-off levels for NAFLD were ≥ 60 for FLI, > 36 for HSI, and ≥ 8.5 for TyG. The performance metrics include accuracy, sensitivity or recall, specificity, precision, area under the receiver operating characteristic curve (AUC), and the F1 score. It is worth noting that the F1 score is the harmonic mean of precision and recall. All statistical analyses were performed using MATLAB R2020a (MathWorks, MA, United States).

RESULTS
The study had 3235 participants (n = 3235). The participant selection process is shown in Figure 1. Based on ultrasound findings, 817 (25.26%) participants had NAFLD. The data of 2265 (70%) and 970 (30%) participants made up the training and testing groups, respectively. The baseline characteristics of participants in the training and testing groups are summarized in Table 1. There were no significant differences between the datasets for all factors.

The performances of 24 machine learning algorithms that were applied to the training dataset are illustrated in Table 2. The ensemble of subspace discriminant and ensemble of random undersampling (RUS) boosted trees had the highest accuracy (78.3%) and highest F1 score (0.53), respectively; both models had an AUC of 0.76. The coarse trees, decision trees with a few leaves, had an accuracy of 76%, AUC of 0.68,
Table 1 Baseline characteristics of participants in training and testing data

<table>
<thead>
<tr>
<th></th>
<th>Training data (n = 2265)</th>
<th>Testing data (n = 970)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographic</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (yr)</td>
<td>43 (29)</td>
<td>43.5 (28)</td>
<td>0.328</td>
</tr>
<tr>
<td>Gender (male) (%)</td>
<td>944 (41.68)</td>
<td>428 (44.12)</td>
<td>0.197</td>
</tr>
<tr>
<td><strong>Race/ethnicity</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White (non-Hispanic) (%)</td>
<td>959 (42.34)</td>
<td>392 (40.41)</td>
<td>0.308</td>
</tr>
<tr>
<td>Black (non-Hispanic) (%)</td>
<td>627 (27.68)</td>
<td>271 (27.94)</td>
<td>0.682</td>
</tr>
<tr>
<td>Mexican American (%)</td>
<td>576 (25.43)</td>
<td>254 (26.19)</td>
<td>0.652</td>
</tr>
<tr>
<td>Others (%)</td>
<td>103 (4.55)</td>
<td>53 (5.46)</td>
<td>0.265</td>
</tr>
<tr>
<td><strong>Body measurement</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Body mass index (kg/m²)</td>
<td>26.4 (7.2)</td>
<td>26.7 (7.4)</td>
<td>0.120</td>
</tr>
<tr>
<td>Waist circumference (cm)</td>
<td>93 (20.5)</td>
<td>93.5 (20.8)</td>
<td>0.182</td>
</tr>
<tr>
<td><strong>Biochemistry tests</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iron (ug/dL)</td>
<td>73 (39)</td>
<td>74 (39)</td>
<td>0.098</td>
</tr>
<tr>
<td>Total iron-binding capacity (ug/dL)</td>
<td>355 (72)</td>
<td>356 (72)</td>
<td>0.450</td>
</tr>
<tr>
<td>Transferrin saturation (%)</td>
<td>20.5 (11.1)</td>
<td>20.8 (11.8)</td>
<td>0.329</td>
</tr>
<tr>
<td>Ferritin (ng/mL)</td>
<td>87 (125)</td>
<td>84.5 (124)</td>
<td>0.508</td>
</tr>
<tr>
<td>Cholesterol (mg/dL)</td>
<td>201 (57)</td>
<td>204 (59)</td>
<td>0.155</td>
</tr>
<tr>
<td>Triglyceride (mg/dL)</td>
<td>120 (100.25)</td>
<td>122.5 (102)</td>
<td>0.562</td>
</tr>
<tr>
<td>HDL cholesterol (mg/dL)</td>
<td>48 (18)</td>
<td>48.5 (18)</td>
<td>0.585</td>
</tr>
<tr>
<td>C-reactive protein (mg/dL)</td>
<td>0.21 (0.29)</td>
<td>0.21 (0.23)</td>
<td>0.686</td>
</tr>
<tr>
<td>Uric acid (mg/dL)</td>
<td>5 (1.9)</td>
<td>5.1 (2)</td>
<td>0.427</td>
</tr>
<tr>
<td><strong>Liver chemistry</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aspartate aminotransferase (U/L)</td>
<td>19 (8)</td>
<td>19 (7)</td>
<td>0.908</td>
</tr>
<tr>
<td>Alanine aminotransferase (U/L)</td>
<td>14 (10)</td>
<td>14 (10)</td>
<td>0.581</td>
</tr>
<tr>
<td>Gamma glutamyl transferase (U/L)</td>
<td>21 (18)</td>
<td>21 (18)</td>
<td>0.787</td>
</tr>
<tr>
<td>Alkaline phosphatase (U/L)</td>
<td>83 (33)</td>
<td>81 (32)</td>
<td>0.524</td>
</tr>
<tr>
<td>Total bilirubin (mg/dL)</td>
<td>0.5 (0.2)</td>
<td>0.5 (0.2)</td>
<td>0.835</td>
</tr>
<tr>
<td>Total protein (g/dL)</td>
<td>7.4 (0.6)</td>
<td>7.4 (0.6)</td>
<td>0.559</td>
</tr>
<tr>
<td>Albumin (g/dL)</td>
<td>4.1 (0.5)</td>
<td>4.1 (0.4)</td>
<td>0.543</td>
</tr>
<tr>
<td>Serum globulin (g/dL)</td>
<td>3.3 (0.6)</td>
<td>3.3 (0.7)</td>
<td>0.941</td>
</tr>
<tr>
<td><strong>Diabetes testing profile</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glycated hemoglobin (%)</td>
<td>5.4 (0.8)</td>
<td>5.4 (0.7)</td>
<td>0.075</td>
</tr>
<tr>
<td>Fasting plasma glucose (mg/dL)</td>
<td>91.6 (12.52)</td>
<td>92.05 (12.2)</td>
<td>0.726</td>
</tr>
<tr>
<td>Fasting C-peptide (pmol/mL)</td>
<td>0.65 (0.68)</td>
<td>0.66 (0.69)</td>
<td>0.746</td>
</tr>
<tr>
<td>Fasting insulin (uU/mL)</td>
<td>9.36 (9.51)</td>
<td>9.73 (10.04)</td>
<td>0.378</td>
</tr>
<tr>
<td>Diabetes medication</td>
<td>165 (7.28%)</td>
<td>68 (7.01%)</td>
<td>0.782</td>
</tr>
</tbody>
</table>

and F1 score of 0.36.

As shown in the first half of Table 3, the ensemble of subspace discriminant, coarse trees, and ensemble of RUS-boosted trees models were selected for testing the process on the testing data. When tested on the testing data, ensemble of subspace discriminant and ensemble of RUS-boosted trees still had a high accuracy (77.7%) and high F1 (0.56), respectively. The coarse tree had an accuracy of 74.9% and an F1 of 0.33. All the machine learning models and datasets are available for public access in the File
## Table 2 The performance comparison of machine learning models on training data

<table>
<thead>
<tr>
<th>No.</th>
<th>Description</th>
<th>Accuracy (%)</th>
<th>AUC</th>
<th>PPV/precision (%)</th>
<th>NPV (%)</th>
<th>Sensitivity/recall (%)</th>
<th>Specificity (%)</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Fine tree</td>
<td>71.6</td>
<td>0.64</td>
<td>42.9</td>
<td>79.8</td>
<td>37.8</td>
<td>83.0</td>
<td>0.40</td>
</tr>
<tr>
<td>2</td>
<td>Medium tree</td>
<td>74.4</td>
<td>0.70</td>
<td>48.9</td>
<td>79.1</td>
<td>30.1</td>
<td>89.4</td>
<td>0.37</td>
</tr>
<tr>
<td>3</td>
<td>Coarse tree</td>
<td>76.0</td>
<td>0.68</td>
<td>55.1</td>
<td>78.9</td>
<td>26.4</td>
<td>92.7</td>
<td>0.36</td>
</tr>
<tr>
<td>4</td>
<td>Linear discriminant</td>
<td>78.0</td>
<td>0.75</td>
<td>61.1</td>
<td>80.9</td>
<td>35.5</td>
<td>92.4</td>
<td>0.45</td>
</tr>
<tr>
<td>5</td>
<td>Logistic regression</td>
<td>78.1</td>
<td>0.75</td>
<td>62.2</td>
<td>80.6</td>
<td>33.9</td>
<td>93.0</td>
<td>0.44</td>
</tr>
<tr>
<td>6</td>
<td>Gaussian naïve Bayes</td>
<td>75.1</td>
<td>0.74</td>
<td>50.8</td>
<td>81.1</td>
<td>40.2</td>
<td>86.8</td>
<td>0.45</td>
</tr>
<tr>
<td>7</td>
<td>Kernel naïve Bayes</td>
<td>72.7</td>
<td>0.73</td>
<td>46.8</td>
<td>85.1</td>
<td>60.1</td>
<td>76.9</td>
<td>0.53</td>
</tr>
<tr>
<td>8</td>
<td>Linear SVM</td>
<td>77.0</td>
<td>0.74</td>
<td>64.4</td>
<td>78.1</td>
<td>19.9</td>
<td>96.3</td>
<td>0.30</td>
</tr>
<tr>
<td>9</td>
<td>Quadratic SVM</td>
<td>77.4</td>
<td>0.70</td>
<td>59.9</td>
<td>80.1</td>
<td>31.8</td>
<td>92.8</td>
<td>0.42</td>
</tr>
<tr>
<td>10</td>
<td>Cubic SVM</td>
<td>72.8</td>
<td>0.64</td>
<td>45.1</td>
<td>79.6</td>
<td>35.3</td>
<td>85.5</td>
<td>0.40</td>
</tr>
<tr>
<td>11</td>
<td>Fine Gaussian SVM</td>
<td>74.7</td>
<td>0.67</td>
<td></td>
<td></td>
<td></td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Medium Gaussian SVM</td>
<td>77.5</td>
<td>0.74</td>
<td>63.9</td>
<td>79.0</td>
<td>25.3</td>
<td>95.2</td>
<td>0.36</td>
</tr>
<tr>
<td>13</td>
<td>Coarse Gaussian SVM</td>
<td>75.7</td>
<td>0.74</td>
<td>66.2</td>
<td>76.0</td>
<td>7.9</td>
<td>98.6</td>
<td>0.14</td>
</tr>
<tr>
<td>14</td>
<td>Fine KNN</td>
<td>68.9</td>
<td>0.58</td>
<td>38.0</td>
<td>78.9</td>
<td>36.9</td>
<td>79.7</td>
<td>0.37</td>
</tr>
<tr>
<td>15</td>
<td>Medium KNN</td>
<td>76.5</td>
<td>0.71</td>
<td>59.7</td>
<td>78.1</td>
<td>21.0</td>
<td>95.2</td>
<td>0.31</td>
</tr>
<tr>
<td>16</td>
<td>Coarse KNN</td>
<td>76.6</td>
<td>0.75</td>
<td>78.1</td>
<td>76.5</td>
<td>10.0</td>
<td>99.1</td>
<td>0.18</td>
</tr>
<tr>
<td>17</td>
<td>Cosine KNN</td>
<td>76.6</td>
<td>0.72</td>
<td>57.9</td>
<td>79.2</td>
<td>27.6</td>
<td>93.2</td>
<td>0.37</td>
</tr>
<tr>
<td>18</td>
<td>Cubic KNN</td>
<td>77.0</td>
<td>0.72</td>
<td>62.0</td>
<td>78.5</td>
<td>22.6</td>
<td>95.3</td>
<td>0.33</td>
</tr>
<tr>
<td>19</td>
<td>Weighted KNN</td>
<td>76.5</td>
<td>0.71</td>
<td>56.7</td>
<td>79.4</td>
<td>28.8</td>
<td>92.6</td>
<td>0.38</td>
</tr>
<tr>
<td>20</td>
<td>Ensemble of boosted trees</td>
<td>76.9</td>
<td>0.74</td>
<td>57.3</td>
<td>80.3</td>
<td>33.6</td>
<td>91.6</td>
<td>0.42</td>
</tr>
<tr>
<td>21</td>
<td>Ensemble of bagged trees</td>
<td>77.2</td>
<td>0.74</td>
<td>58.9</td>
<td>80.2</td>
<td>32.5</td>
<td>92.3</td>
<td>0.42</td>
</tr>
<tr>
<td>22</td>
<td>Ensemble of subspace discriminant</td>
<td>78.3</td>
<td>0.76</td>
<td>66.7</td>
<td>79.7</td>
<td>28.3</td>
<td>95.2</td>
<td>0.40</td>
</tr>
<tr>
<td>23</td>
<td>Ensemble of subspace KNN</td>
<td>75.5</td>
<td>0.69</td>
<td>54.7</td>
<td>77.2</td>
<td>16.4</td>
<td>95.4</td>
<td>0.25</td>
</tr>
<tr>
<td>24</td>
<td>Ensemble of RUS boosted trees</td>
<td>70.4</td>
<td>0.76</td>
<td>44.2</td>
<td>86.3</td>
<td>66.4</td>
<td>71.7</td>
<td>0.53</td>
</tr>
</tbody>
</table>

AUC: Area under the curve; KNN: K-nearest neighbors; NPV: Negative predictive value; PPV: Positive predictive value; RUS: Random undersampling; SVM: Support vector machine.

Exchange portal of the MATLAB Central File Exchange[22]. The performance of three NAFLD on the testing data are also displayed in the second half of Table 3. FLI was the best performer among the NAFLD indices with the accuracy of 68.6% and F1 score of 0.52. However, the ensemble of RUS boosted trees was superior to FLI in all metrics.

## DISCUSSION

Our study compared 24 different machine learning techniques to determine the optimal clinical predictive model for NAFLD. The accuracy of these models on the training data did not show much variation (range 9.4%), with an average of 75.5% (Table 2). The top two models were ensemble of subspace discriminant and ensemble of RUS boosted trees. The ensemble of subspace discriminant model had a higher accuracy while the ensemble of RUS boosted trees model had a better performance in classifying positive NAFLD, as indicated by the F1 score. Both models were ensemble type, which use multiple diverse models in combination to produce an optimal prediction. They are more complex machine learning models that apparently yield better predictions. Compared to accuracy, the F1 score is regarded as a superior performance metric for a class imbalance problem (often a large number of actual
Table 3 The performance of machine learning models and other non-alcoholic fatty liver disease indices on testing data

<table>
<thead>
<tr>
<th>No.</th>
<th>Description</th>
<th>Accuracy (%)</th>
<th>AUC</th>
<th>PPV/precision (%)</th>
<th>NPV (%)</th>
<th>Sensitivity/recall (%)</th>
<th>Specificity (%)</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ensemble of subspace discriminant</td>
<td>77.7</td>
<td>0.78</td>
<td>66.7</td>
<td>78.8</td>
<td>23.7</td>
<td>96</td>
<td>0.35</td>
</tr>
<tr>
<td>2</td>
<td>Coarse trees</td>
<td>74.9</td>
<td>0.72</td>
<td>50.8</td>
<td>78.3</td>
<td>24.5</td>
<td>92</td>
<td>0.33</td>
</tr>
<tr>
<td>3</td>
<td>Ensemble of RUS boosted trees</td>
<td>71.1</td>
<td>0.79</td>
<td>45.5</td>
<td>88.4</td>
<td>72.7</td>
<td>70.6</td>
<td>0.56</td>
</tr>
<tr>
<td>4</td>
<td>Fatty liver index</td>
<td>68.6</td>
<td>0.74</td>
<td>42.4</td>
<td>86.6</td>
<td>68.6</td>
<td>68.6</td>
<td>0.52</td>
</tr>
<tr>
<td>5</td>
<td>Hepatic steatosis index</td>
<td>65.1</td>
<td>0.70</td>
<td>37.9</td>
<td>83.3</td>
<td>60.4</td>
<td>66.6</td>
<td>0.47</td>
</tr>
<tr>
<td>6</td>
<td>Triglyceride and glucose index</td>
<td>56.9</td>
<td>0.69</td>
<td>34.8</td>
<td>88.3</td>
<td>80.8</td>
<td>48.8</td>
<td>0.49</td>
</tr>
</tbody>
</table>

AUC: Area under the curve; NAFLD: Non-alcoholic fatty liver disease; NPV: Negative predictive value; PPV: Positive predictive value; RUS: Random undersampling.

Figure 1 Study design and data partitioning flow chart. NAFLD: Non-alcoholic fatty liver disease.

negatives). In our opinion, the ensemble of RUS boosted trees model was the best performing machine learning model in this study.

Technically, the final prediction of the ensemble method was derived from a combination of multiple predictions from different algorithms. In our case, the predicted outcome of the ensemble of RUS boosted trees model was derived from a weighted average outcome of 30 RUS boosted trees; the sample visualization of these RUS boosted trees can be found in the file uploaded to the MATLAB Central File Exchange[22].
On the other hand, we compared the performance of the previous model with the
cosy trees model, simple decision trees with several leaves and splits (Figure 2). The
decision logic of the coarse trees model consisted of only two factors: Waist circum-
ference and serum C-peptide. In terms of testing performance, it had a reasonable
accuracy (AUC, 0.72; accuracy, 74.9%; and F1 score, 0.33). Since it is simple-to-use and
easily interpretable, the coarse trees model can be more practically used in clinical
practice.
Waist circumference is directly associated with obesity and metabolic syndrome[23,
24]. They are also the established risk factors of NAFLD. The cut-off of 109.35 cm
seems to be slightly higher than the general cut off value for metabolic syndrome
(men, 102 cm and women, 80 cm)[25]. It is used to calculate the visceral adiposity
index, which provides a good predictive capability[26]. The advantage of incor-
porating waist circumference into the model is its retrieval ability.
Our results are similar to those of a previous study identifying the risk factors of
NAFLD[27]. C-peptide is an indicator of insulin resistance[28,29]. Serum C-peptide
is associated with NAFLD, NASH, and fibrosis progression[28-30]. Additionally, serum
C-peptide levels increase with NAFLD severity[29,31,32]. In our study, serum C-
peptide is more significantly associated with NAFLD prediction than liver function
test. This can be explained by the fact that liver enzymes are possibly not specific to
NAFLD. They can also be elevated in other liver diseases. On the contrary, serum C-
peptide is related to metabolic alterations, which play a direct role in NAFLD
development.
We compared the performance of three NAFLD indices (FLI, HSI, and TyG) on the
testing data. Among these three NAFLD indices, FLI had the highest performance in
terms of accuracy (68.6%) and F1 (0.52). However, performance-wise, the ensemble of
RUS boosted trees was superior to FLI in all aspects. In terms of simplicity, FLI is not
complex, but it might be impossible for physicians to use it without spreadsheets or
computers because it involves many mathematical operations, such as multiplication,
logarithm function, and exponential function. Therefore, coarse trees remained the
simplest model.
Previously developed machine learning models for NAFLD prediction have used
more complex parameters, including laboratory and noninvasive scores. A population-
based study in Italy developed a score for NAFLD diagnosis with a moderate accuracy
of 68% in the model development phase, but extremely high performance in the
testing (prediction) phase using the small sample size of 50. The predictors used in the
model were of abdominal volume index, glucose, gamma glutamyl transferase, age,
and sex[33]. A Chinese study incorporated three demographic factors and 15
Laboratory tests as predictors for Bayesian network model[34]. The inclusion of simple
costituents, liver enzymes, lipid panels, and complete blood count resulting in an
accuracy of up to 80% in a 10-fold cross validation; there was no separate data set for
external validation or testing. A Taiwanese study revealed that waist circumference
was the most influential factor in the model resulting in a high performance with an
AUC of 0.925[13]. Similarly, such performance was based on a 10-fold cross validation,
not on a separate data set for external validation or testing. In addition, the ethnic
Chinese population generally has a lower alcohol consumption; it might not be
generalized to other ethnic groups[12,15]. A Canadian study revealed that HDL, BMI,
sex, plasma glucose, blood pressure, and age were factors used in the decision criteria
of decision trees with an AUC of 0.73[14]. These reports showed different significant
factors in their models. This might be explained by the different populations in terms
of ethnicity, alcohol consumption, and obesity prevalence. Compared to prior reports,
our study involved a general population of the United States, which has less selection
bias and contains diverse races. Therefore, the derived models in this study can be
applied to diverse ethnic and racial backgrounds. A detailed comparison of the
proposed machine learning models in prior reports is summarized in Table 4.
The application of machine learning in regarding NAFLD has evolved from the
diagnosis with the noninvasive screening methods to liver biopsy. The new score
achieves the reasonable performance with AUC of 0.70, in terms of differentiating
between NAFL and NASH[11]. Deep learning model was evaluated for diagnosis
NAFLD based on ultrasound images and had a good predictive ability (AUC > 0.7)
[34]. Given the advancement in this field, it can also be used to quantify steatosis,
inflammation, ballooning, and fibrosis in biopsy histology of patients with NAFLD
having excellent results[35].
This study had strengths. First, this is the first United States population-based study
with more than 3000 individuals from NHANES III. Secondly, we aimed to propose
the simple model with a reasonable predictive power for NAFLD. This model will be
potentially applied in clinical practice, especially by primary care providers, prior to
Table 4 The performance comparison of published machine learning models on non-alcoholic fatty liver disease prediction

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Type of data/country or territory of data</th>
<th>Number of train/external testing data</th>
<th>Model</th>
<th>Accuracy (%)</th>
<th>AUC</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sorino et al</td>
<td>Population/Italy</td>
<td>2920/50</td>
<td>Support vector machine</td>
<td>68</td>
<td>N/A</td>
<td>98.5</td>
<td>100</td>
<td>N/A</td>
</tr>
<tr>
<td>Wu et al</td>
<td>Hospital/Taiwan</td>
<td>577/NA</td>
<td>Random forest</td>
<td>86.5</td>
<td>0.925</td>
<td>87.2</td>
<td>85.9</td>
<td>N/A</td>
</tr>
<tr>
<td>Islam et al</td>
<td>Hospital/Taiwan</td>
<td>994/NA</td>
<td>Logistic regression</td>
<td>70</td>
<td>0.763</td>
<td>74.1</td>
<td>64.9</td>
<td>N/A</td>
</tr>
<tr>
<td>Ma et al</td>
<td>Hospital/China</td>
<td>1050/NA</td>
<td>Bayesian network</td>
<td>82.92</td>
<td>N/A</td>
<td>67.5</td>
<td>87.8</td>
<td>0.655</td>
</tr>
<tr>
<td>Perveen et al</td>
<td>Primary care network/Canada</td>
<td>64%/34% of 40637</td>
<td>Decision trees</td>
<td>N/A</td>
<td>0.73</td>
<td>73</td>
<td>N/A</td>
<td>0.67</td>
</tr>
<tr>
<td>Yip et al</td>
<td>Hospital/Hong Kong</td>
<td>500/442</td>
<td>Ridge regression</td>
<td>87</td>
<td>0.87</td>
<td>92</td>
<td>90</td>
<td>N/A</td>
</tr>
<tr>
<td>Birjandi et al</td>
<td>Hospital/Iran</td>
<td>359/1241</td>
<td>Decision trees</td>
<td>75</td>
<td>0.75</td>
<td>73</td>
<td>77</td>
<td>N/A</td>
</tr>
<tr>
<td>Our study</td>
<td>Population based/United States</td>
<td>2265/970</td>
<td>Ensemble of RUS boosted trees</td>
<td>71.1</td>
<td>0.79</td>
<td>72.7</td>
<td>70.6</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Coarse trees</td>
<td>74.9%</td>
<td>0.72</td>
<td>24.5%</td>
<td>92%</td>
<td>0.33</td>
</tr>
</tbody>
</table>

Cross-validation performance (no separate dataset designated for testing the performance). RUS: Random undersampling; AUC: area under the receiver operating characteristic curve; N/A: Not applicable; NA: Not available.

Figure 2 The decision logic of coarse trees. NAFLD: Non-alcoholic fatty liver disease.

referring patients to hepatologists. This study had some limitations. (1) Missing data were inherited from the nature of population dataset from NHANES III; (2) NAFLD was diagnosed with ultrasonography, which is not the gold standard; however, it is the primary imaging modality for NAFLD diagnosis in population-based studies and available in primary care medical facilities; (3) At the time of writing this article, there was no external dataset available that like that of NHANES III for validating the models; and (4) It may be impossible to completely reproduce the machine learning algorithms in this study since randomization was used in the modeling process, such as data partitioning, cross validation, and creation of some machine learning models. This explains why we made the trained models available to the public so that anyone can use the models directly and/or validate our results.

CONCLUSION

Machine learning algorithms can summarize a large dataset into predictive models. The best performing model measured by the F1 score from our study is the ensemble of RUS boosted trees, which is a complex model that uses all 30 factors and behaves more like a black box to physicians. In contrast, the coarse trees model, which is composed of serum C-peptide and waist circumference, can generate a reasonable predictive performance, and most importantly is the simplest to use. To facilitate
Atsawarungruangkit A et al. Machine learning models in NAFLD

clinical decision-making, complex models should be incorporated into the electronic medical record system. This will lead to proper investigation and treatment selection for specific individuals at risk, helping to maximize healthcare resource utilization. If software deployment is not achievable, a simple model be used directly by physicians. Therefore, the model choice depends on the user objectives and resources. Therefore, the more complex model required more resources and was likely to outperform. The less complex model may not be the most accurate model but can be easily implemented and interpreted in clinical practice.

ARTICLE HIGHLIGHTS

Research background
Nonalcoholic fatty liver disease (NAFLD) is the most common chronic liver disease that can progress to more severe liver disease.

Research motivation
Early patient identification using a simple method is highly desirable for preventing the progression of NAFLD.

Research objectives
To create machine learning models for predicting NAFLD in the general United States population.

Research methods
This study was designed as a retrospective cohort by using the NHANES 1988-1994. Adults (20 years and above in age) with gradable ultrasound results were included in this study.

Research results
Based on F1, the ensemble of ensemble of random undersampling boosted trees was the top performer (accuracy 71.1% and F1 0.56) while a simple model (coarse trees) had an accuracy of 74.9% and an F1 of 0.33.

Research conclusions
Although a simpler model such as coarse trees was not the top performer, it consisted of only two predictors: fasting C-peptide and waist circumference. Its simplicity is useful in clinical practice.

Research perspectives
The findings from this study can facilitate clinical decision-making for clinicians and also allow researchers to investigate the developed machine learning models. This will lead to proper investigation and treatment selection for specific individuals at risk, helping to maximize healthcare resource utilization.

REFERENCES

Atsawarungruangkit A et al. Machine learning models in NAFLD.


