Prevalence of NAFLD Risk Factors Used for Prediction in Adults

Daniel Zhang and Victoria Caruso
University of Massachusetts Amherst
{danielzhang,vcaruso}@cs.umass.edu

1 BACKGROUND

Chronic liver disease, specifically non-alcoholic fatty liver disease is a high contributor of mortality in the United States Population. Because of difficulty in finding the presence of NAFLD in a timely, less expensive method than liver biopsy, research has been conducted to examine methods to predict NAFLD using associated factors [2].

In the article, “Prevalence and factors associated with NAFLD detected by vibration controlled transient elastography among US adults: Results from NHANES 2017–2018,” researchers examine the prevalence of NAFLD through identifying associated factors. This article describes the current gold standard algorithm used on NHANES 2017-2018 data. Some of the features found associated with NAFLD include: metabolic dysfunction (metabolic syndrome, obesity, diabetes) and inadequate physical activity. [1]

In the article, "Prevalence of Non-Alcoholic Fatty Liver Disease and Risk Factors for Advanced Fibrosis and Mortality in the United States,” researchers use the United States Fatty Liver Index (USFLI) to determine risk factors associated with advanced fibrosis within the U.S. population. They use a NHANES dataset from 1999–2012 and found that Mexican Americans were at higher risk for NAFLD [2]. Combining their findings they construct a diagnosis algorithm using factors of ethnicity along with weighting age, GGT, waist circumference, fasting glucose, and fasting insulin [2]. They construct a logistic regression algorithm using age, ethnicity, waist circumference, GGT, insulin and glucose [2].

The article, "Clinical advances in Liver, Pancreas, and Biliary Tract" researchers describe a study using questionaire and ultrasound results from Brooke Army Medical Center to identify factors associated with NAFLD. Some of the identified factors determined by calculating a p-value include: Alanine, aminotransferase, aspartate aminotransferase, BMI, and insulin.

2 THE GOLD STANDARD ALGORITHM

The Gold Standard Algorithm is an algorithm used to determine whether or not someone has NAFLD. The algorithm is specified as follows: Liver ultrasound has a median controlled attenuation parameter (CAP) 248 db/m, Liver median stiffness 7 kPA, they do not have hepatitis B or hepatitis C, and no high alcohol intake (2 or less drinks per day for men and 1 or less drinks per day for women). Those who fulfill all the above criteria are said to have NAFLD.

3 UNITED STATES FATTY LIVER INDEX

The United States Fatty Liver Index (USFLI) is a logistic regression model based off the Fatty Liver Index. USFLI takes into account age, sex, ethnicity, education level, smoking status, BMI, diabetes status, and metabolic syndrome. The algorithm provides a score between 0-100 and those with USFLI scores >= 30 are considered to have NAFLD. The algorithm details are below [2].

$$
\begin{aligned}
USFLI = e\ \hat{}\ (&-0.8073 * non - HispanicBlack + 0.3458 \\
&* MexicanAmerican + 0.0093 * Age + 0.6151 \\
&*\log\_e(GGT) + 0.0249 * WaistCircumference \\
&+ 1.1792 * \log\_e[\![(Insulin)]\!] \quad 0.8242 \\
&*\log\_e[\![(Glucose) - 14.7812]\!]) \\
/((1 + e\ \hat{}\ (&-0.8073 * non - HispanicBlack + 0.3458 \\
&*MexicanAmerican + 0.0093 * Age + 0.6151 * \log \\
&\_e(GGT) + 0.0249 \\
&*WaistCircumference + 1.1792 * \log\_e[\![(Insulin) + \\
&]\!] \ 0.8242 * \log\_e[\![(Glucose) \\
&-14.7812]\!]))) * 100
\end{aligned}
$$

Figure 1: from [2]

4 OBJECTIVE

Our goal is to use machine learning to create a model that can predict Nonalcoholic fatty liver disease (NAFLD) using features from the "Gold Standard Algorithm" and the United States Fatty Liver Index and additional associated features. We will compare our model to the existing commonly used logistic regression model used to predict NAFLD in the cited paper [2]. We will use the gold standard algorithm for detecting NAFLD [1] to create classifier labels for each patient in the data-set in order to see the accuracy of our predictive model.

5 INITIAL DATASET

For our dataset, we will use the 2017-2018 NHANES dataset. We limited the data set to those who were ages 18 years or older to ensure that all participants we examine are eligible for the data collection of the variables. We choose to use the older dataset of 2017-2018 because datasets after this date, the datasets did not have enough data on all of the variables we wanted to consider.

We will be using variables to determine alcohol use and participants who engage in harmful drinking will not be classified as having NAFLD ($\geq$ more than 2 drinks per day for women and $\geq$ and more than 3 drinks per day for men) using ALQ130 (Avg alcohol drinks/day) and RIAGENDR (male/female). This is because Liver Disease due to high alcohol intake is a different disease, so we want to eliminate the subset of people that may have Liver Disease due to alcohol instead of NAFLD. We will use the gold standard metrics of LUXCAPM for median controlled attenuation parameter(CAP) $\geq$ 248 dB/m based on previous analysis [1] and and LUXSMED for stiffness of the liver. We will be using HEQ010 and HEQ030 for Hepatitis B and C. We will also be using the variables from competing algorithm in [2] as a base for our model as they have been proven to show association including: ethnicity (RIDRETH3), GGT (LBXSGTSI), waist circumference (BMXWAIST), fasting glucose (LBXGLU), and insulin (LBXIN).

6 PREPROCESSING DATA

For our preprocessing of the NAFLD data we first limited to adults by using participants with an age of 18 years or older. We do this because participants that are not adults to assure that all of the participants we examine are eligible for the data collection of the variables we select. Also, the gold standard and competing algorithms followed this limitation which allows us to compare our algorithms under similar circumstances. We then proceeded to remove participants with incomplete or empty data. Those who did not have values for one or more of the features in the Gold Standard, USFLI, or our machine learning model where not included in our dataset. After all of our preprocessing we have a total number of 1312 participants. We identified 83 of these participants as having NAFLD by using the Gold Standard algoirthm. This means that our of our total testable subset, 7% of the participants have NAFLD.

7 NEW RISK FACTORS

For gathering new risk factors, we examined the works in the article for our competing algorithm which included the p-value for many other variables that were not included in their algorithm despite showing statistical significance under a 5% significance level [2]. These variables include: BMI, Diabetes, Albumin, HbA1c, Fasting insulin, Fasting glucose and more [2]. Some of the variables found from the article, "Clinical advances in Liver, Pancreas, and Biliary Tract" include BMI and insulin. They also claim that, "NAFLD patients also ate at fast-food restaurants more frequently and exercised less," which influenced are decision to include physical activity and nutritional data as variables. The variables we plan to add after we research their potential association with NAFLD include: HDL cholesterol (mg/dL)(LBDHDD), Albumin (LBDSALSI), HbA1c levels ($\geq$ 8%) indicating uncontrolled diabetes (LBXGH), BMI (BMXBMI), income (INDFMPIR), Physical Activity (P_PAQ), and nutritional data (P_DBQ) [2, 3]. For considering Physical Activity, we examine physical activity at work which is described as doing 10 minutes of intense physical activity at work for at least 10 minutes at least once a week. We also considered recreational physical activity which is also for for at least 10 minutes at least once a week. Our nutritional data is slit between 6 possible values. 1 means extremely good diet while 5 means bad diet. For our data, we considered 1-3: extremely good - good for good diet and 4-5: fair - bad as bad diet.

8 CALCULATING ASSOCIATION OF VARIABLES

To calculate the association of numerical variables, we calculated the p-value using a t-test with a 5% significance level. In order to calculate the assocication of categorical variables, we calculated the comparison percentages between those with and without NAFLD. We also computed the odds ratio. We also visually depicted the comparisons and spread of the data across these variables using charts and figures that are presented at end end of the article.

8.1 Variables with Association

The numerical variables with visible association were HDL Cholesterol, Blood HbA1c level ((≥ 8%) indicating uncontrolled diabetes) and BMI which all had a p-value of approximately 0.000. Other categorical variables that show association were Recreational Physical Activity with an odds ratio of 0.5625 which shows that physical activity is protective over NAFLD. Diet had a reverse result from what we anticipated by having an odds ratio of 1.7145 suggesting that the odds of NAFLD are higher for those with a healthier diet. This is possibly due to our small dataset.

8.2 Variables with No Association

The variables with no visible association were Physical Activity at Work with an odds ratio of 1.0538 showing that there is little to no association. Income was not shown to have high association because of a p-value equal to 0.185. Albumin also had a non significant p-value of 0.665.

9 METHODS

First we will split our data-set into two parts the training set and the testing set each with 50% of the data. Next we run the NAFLD gold standard algorithm to create classifier labels for our data-sets. We then implement the algorithm computing the United States Fatty Liver Index (USFLI) which is an existing well known algorithm made from a logistic regression model. This algorithm gives patients a score from 0-100 with scores greater than or equal to 30 indicating patients having NAFLD. Lastly, we implement classifier two machine learning models. The first model we will use will be k-nearest neighbors and we will optimize parameter n for the number of neighbors to query using the k-cross fold validation technique. The next model is the Decision Tree model, which takes no parameters. Lastly, we analyze our results using four evaluation metrics: f1-score, precision, recall and auroc.

10 METRICS

Here we will talk about the metrics we used to evaluate our machine learning model

10.1 Precision

In our prediction we have the set of patients we predicted to have NAFLD. Out of those patients, those who actually have NAFLD are considered to be True Positive (TP), while those that do not are considered to be False Positive (FP). Thus precision = $\frac{TP}{TP+FP}$.

10.2 Recall

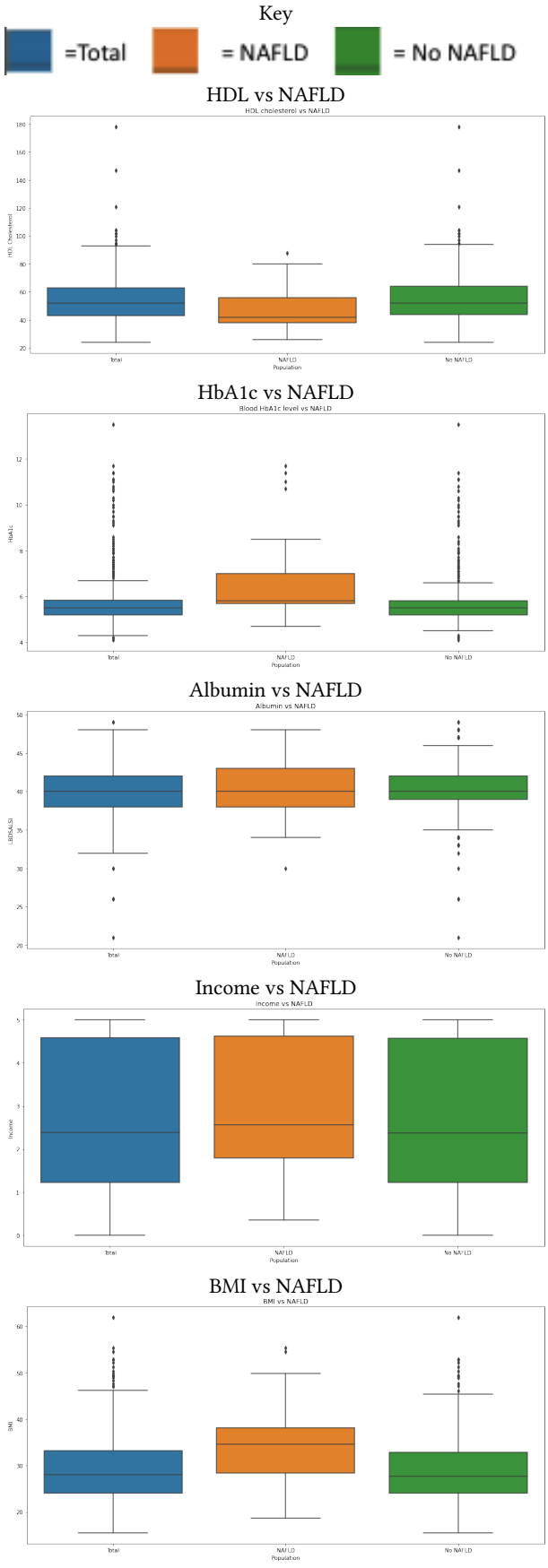
In our data-set we have the set of patients whose condition is they have NAFLD. Out of those patients, those who we predicted to have NAFLD are considered True Positive (TP), while those we predicted to not have NAFLD are considered False Negative (FN). Recall = $\frac{TP}{TP+FN}$

10.3 F1-Score

The F1-score combines precision and recall into a single metric by taking their harmonic mean. F1 = $2^* \frac{precision*recall}{precision+recall}$

10.4 Auroc

Area Under the Receiver Operating Characteristic Curve (Auroc) is a plotted graph with The False Positive Rate on the x-axis and True Positive Rate on the Y-axis. We then take the area under the curve to determine how well our model is performing. Values around



0.5 indicates guessing or no logic, while values closer to 1 indicate better performance, with 1 being the best possible score.

11 RESULTS

After running our model we found that the K-nearest neighbor model performed worse than the USFLI algorithm across all metrics, while the decision tree model performed better than the USFLI algorithm across the board.

11.1 Machine Learning Model Results

	Decision Tree	K-nearest neighbors	USFLI
F1-score	0.810	0.214	0.263
Precision	0.780	0.196	0.263
Recall	0.842	0.237	0.263
Auroc	0.914	0.588	0.609

11.2 Interpretation

In our findings we found that the k-nearest neighbors performed quite poorly. This is likely do to the fact that those with NAFLD do not necessarily share similar data clusters. While those with NAFLD likley have similar characteristics, not all patients may share each characteristics. On the other hand the decision tree performed extremely well and much better than the USFLI algorithm or the k-nearest neighbors model. This is likely due to the ability of decision trees to filter out less useful features when predicting for NAFLD combined with the smaller data-set. Given a larger data-set, we believe it is unlikely for the algorithm to continue performing this well.

12 LIMITATIONS

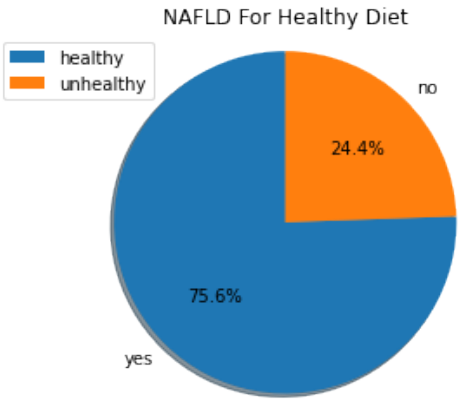
Some of the limitations of our work is our small dataset. Because our goal was to better predict NAFLD, we included additional factors to our algorithm which in turn decreases the amount of data available because when there is a null value for at least one column of each individual, we remove that row from the dataset. This may have been the cause for the occurrence of some of the variables giving a high p-value such as Albumin even through they have been shown to have a low p-value in other works such as [2].

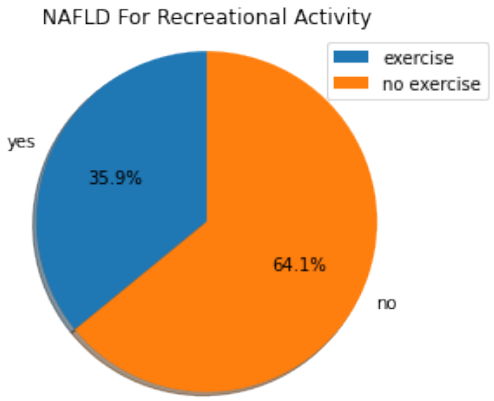
REFERENCES

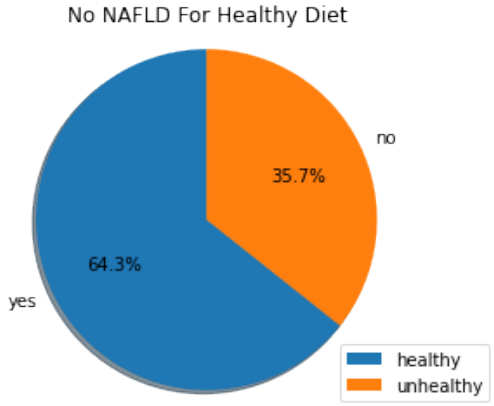
[1] Zhang, X., Heredia, N. I., Balakrishnan, M., Thrift, A. P. (2021). Prevalence and factors associated with NAFLD detected by vibration controlled transient elastography among US adults: Results from NHANES 2017–2018.

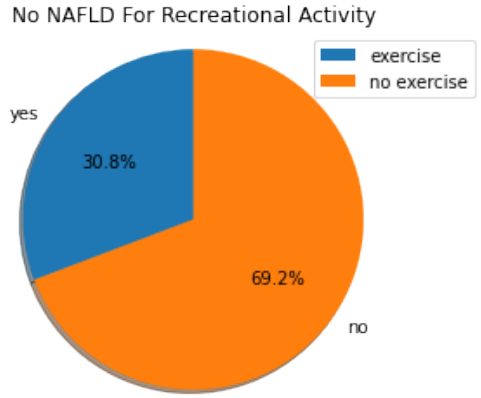
[2] Le, M.H., Devaki, P., Ha, N.B., Jun, D.W., Te, H.S., Cheung, R.C. and Nguyen, M.H., 2017. Prevalence of non-alcoholic fatty liver disease and risk factors for advanced fibrosis and mortality in the United States. PloS one, 12(3), p.e0173499.

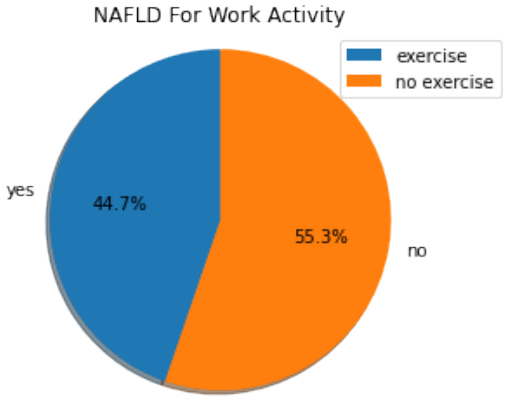
[3] Williams, C.D., Stengel, J., Asike, M.I., Torres, D.M., Shaw, J., Contreras, M., Landt, C.L. and Harrison, S.A., 2011. Prevalence of nonalcoholic fatty liver disease and nonalcoholic steatohepatitis among a largely middle-aged population utilizing ultrasound and liver biopsy: a prospective study. Gastroenterology, 140(1), pp.124-131.

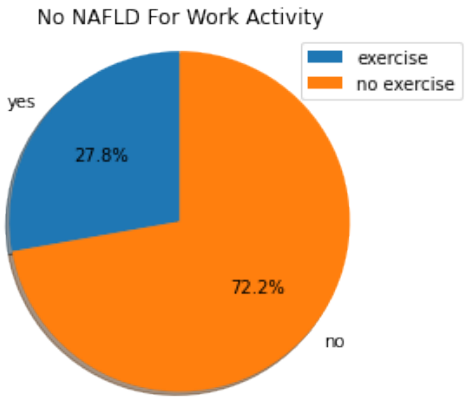

NAFLD For Healthy Diet
healthy
unhealthy
no
24.4%
75.6%
yes


NAFLD For Recreational Activity
exercise
no exercise
yes
35.9%
64.1%
no


No NAFLD For Healthy Diet
no
35.7%
64.3%
yes
healthy
unhealthy


No NAFLD For Recreational Activity
exercise
no exercise
yes
30.8%
69.2%
no


NAFLD For Work Activity
exercise
no exercise
yes
44.7%
55.3%
no


No NAFLD For Work Activity
exercise
no exercise
yes
27.8%
72.2%
no