Machine Learning Algorithms for Identifying Undiagnosed Nonalcoholic Steatohepatitis: A Veterans Affairs Health System Study

Onur Baser, MA, MS, PhD[1]; Katarzyna Rodchenko, MA, MPH[2]; Munira Mohamed, MPH[2]; Alexandra Passarelli, MPH[2]; Shuangrui Chen, MS[2]; Nehir Yapar[2]; Erdem Baser, MS, PhD[3]
[1]Graduate School of Public Health, City University of New York (CUNY), New York, NY; [2]Columbia Data Analytics, New York, NY; [3]Mergen Medical Research, Bilkent Cyberpark, Ankara, Turkey

BACKGROUND

Non-alcoholic steatohepatitis (NASH) can lead to the development of end-stage liver disease, cirrhosis, or hepatocellular carcinoma and is a leading risk factor for cardiometabolic diseases. Its prevalence is rapidly increasing alongside the global epidemics of obesity and diabetes.[1] Despite its substantial burden on patients, providers, and payers, it often goes undiagnosed in clinical practice because of its nonspecific symptoms and the need for direct imaging assessments.[1] Since it is challenging to diagnose by routine laboratory tests or clinical examination, liver biopsy is currently the reference standard for diagnosing and staging NASH.[2] Liver biopsy carries certain risks, including postsurgical bleeding, pain, and additional expenses.[2]

Machine learning (ML), an area of artificial intelligence, has increased in tandem with electronic health records and the advancement of data analytics to address issues affecting patients, payers, and providers.[1,3,4] It is defined as the use of computer algorithms that can learn complex patterns from data. These algorithms can help acquire, interpret, and synthesize healthcare data from diverse sources.[4] A practical application of ML is the prediction of disease presence among individual patients within large databases.

OBJECTIVES

This analysis aimed to identify patients in the Veterans Health Administration (VHA) population who likely had undiagnosed NASH using ML algorithms.

METHODS

A retrospective analysis was conducted utilizing the VHA dataset of 25 million adult enrollees. The study population was categorized as NASH-positive, non-NASH, and at-risk cohorts.

Machine learning models, including logistic regression, naïve Bayes, gradient boosting, and random forest, were developed and compared using receiver operator characteristics, area under the curve (AUC), and accuracy metrics.

Analysis: Logistic regression was used to assess the odds of NASH and controlled for age, sex, geographic region, and comorbidities.

RESULTS

A retrospective analysis of the VHA database population identified 4,223,443 patients who met study inclusion criteria (**Table 1**).

Table 1. Positive NASH, non-NASH, and at-risk NASH cohorts

Cohort Characteristics	n
Positive NASH cohort	
Inclusion	
1 diagnosis of NASH during outcome windows	8,180
Age 18-85 years	8,144
Exclusion	
1 NASH diagnosis in the 24 months pre index date (baseline period)	4,903
Non-NASH cohort	
Inclusion	
1 diagnosis of nonalcoholic liver fibrosis or cirrhosis during outcome window	40,753
Age 18-85 years	40,241
Exclusion	
1 NASH diagnosis during 24 months pre index date (baseline period)	35,528
Unknown NASH cohort	
Inclusion	
Age 18-85 years	4,674,355
Exclusion	
1 diagnosis of alcoholic liver disease, liver disorders in pregnancy, primary sclerosing cholangitis, hepatorenal syndrome, portal hypertension, primary biliary cirrhosis, nonalcoholic liver disease, toxic liver disease, spontaneous bacterial peritonitis, ascites, esophageal varices, alcohol abuse, alpha-1 antitrypsin deficiency, hemochromatosis, Wilson disease, Gaucher disease, liver cancer, hepatitis, other alcohol-related conditions, and/or encephalopathy	4,183,012

NASH: nonalcoholic steatohepatitis

RESULTS (cont'd)

Compared with the NASH-positive diagnosis group, the non-NASH group was more likely to be aged >65 years, male, Black, and less likely to have any comorbidities or procedures. The at-risk group was more likely to be >65 years of age, White, and less likely to have any comorbidities or procedures than the NASH cohort (**Table 2**). The NASH cohort was more likely to be older, male, White, and have comorbidities, most commonly type 2 diabetes and obesity.

Table 2. Demographics and Clinical Profiles of Each NASH Cohort

Characteristics	NASH (n = 4903)	Non-NASH (n = 35,528)	At Risk of NASH (n = 4,183,012)	*P* Value	
				NASH vs Non-NASH	NASH vs At Risk of NASH
Age, years (n, %)					
Mean (SD[a]/SE[b])	58.54 (12.82[a])	64.31 (7.23[a])	61.12 (15.24[b])	<0.0001	<0.0001
18-34	284 (5.79)	82 (0.23)	362,065 (8.66)	<0.0001	<0.0001
35-44	503 (10.26)	325 (0.91)	333,367 (7.97)	<0.0001	<0.0001
45-54	842 (17.17)	1959 (5.51)	517,268 (12.37)	<0.0001	<0.0001
55-64	1,213 (24.74)	15,358 (43.23)	773,434 (18.49)	<0.0001	<0.0001
³65	2,061 (42.04)	17,804 (50.11)	2,196,878 (52.52)	<0.0001	<0.0001
Sex, n (%)					
Male	4550 (92.80)	34,474 (97.03)	3,839,218 (91.78)	<0.0001	0.0094
Race, n (%)					
White	3822 (77.95)	23,596 (66.42)	2,928,774 (70.02)	<0.0001	<0.0001
Black	503 (10.26)	8466 (23.83)	665,163 (15.90)	<0.0001	<0.0001
Unknown	370 (7.55)	2428 (6.83)	457,766 (10.94)	0.0654	<0.0001
Other	208 (4.24)	1038 (2.92)	131,309 (3.14)	<0.0001	<0.0001
Comorbidities, n (%)					
Obesity	2177 (44.40)	7143 (20.11)	760,739 (18.19)	<0.0001	<0.0001
Type 2 diabetes	2480 (50.58)	14,474 (40.74)	1,051,137 (25.13)	<0.0001	<0.0001
Metabolic disorder	2 (0.04)	16 (0.05)	715 (0.02)	0.895	0.205
NAFL	751 (15.32)	1392 (3.92)	10,479 (0.25)	<0.0001	<0.0001
Hypertension	3358 (68.49)	24,975 (70.30)	2,220,271 (53.08)	<0.0001	<0.0001
Procedures, n (%)					
Liver biopsy	194 (3.96)	796 (2.24)	910 (0.02)	<0.0001	<0.0001
Liver panel	4364 (89.01)	28,352 (79.80)	3,044,204 (72.78)	<0.0001	<0.0001
Abnormal liver function test	1546 (31.53)	4344 (12.23)	74,539 (1.78)	<0.0001	<0.0001
Abnormal levels of other serum enzymes	316 (6.45)	1050 (2.96)	20,412 (0.49)	<0.0001	<0.0001
Abdominal ultrasound	2524 (51.48)	20,930 (58.91)	141,171 (3.37)	<0.0001	<0.0001
Comprehensive metabolic panel	3271 (66.71)	23,537 (66.25)	2,165,517 (51.77)	0.5184	<0.0001

NAFL: nonalcoholic fatty liver; NASH: nonalcoholic steatohepatitis
[a]Standard deviation (SD)
[b]Standard error

The random forest model performed best, achieving an AUC of 83% and accuracy of 90% (**Table 3**). Since true positives and true negatives were underrepresented, the data were considered imbalanced. Additionally, of the 4 models applied, random forest yielded the highest scores.

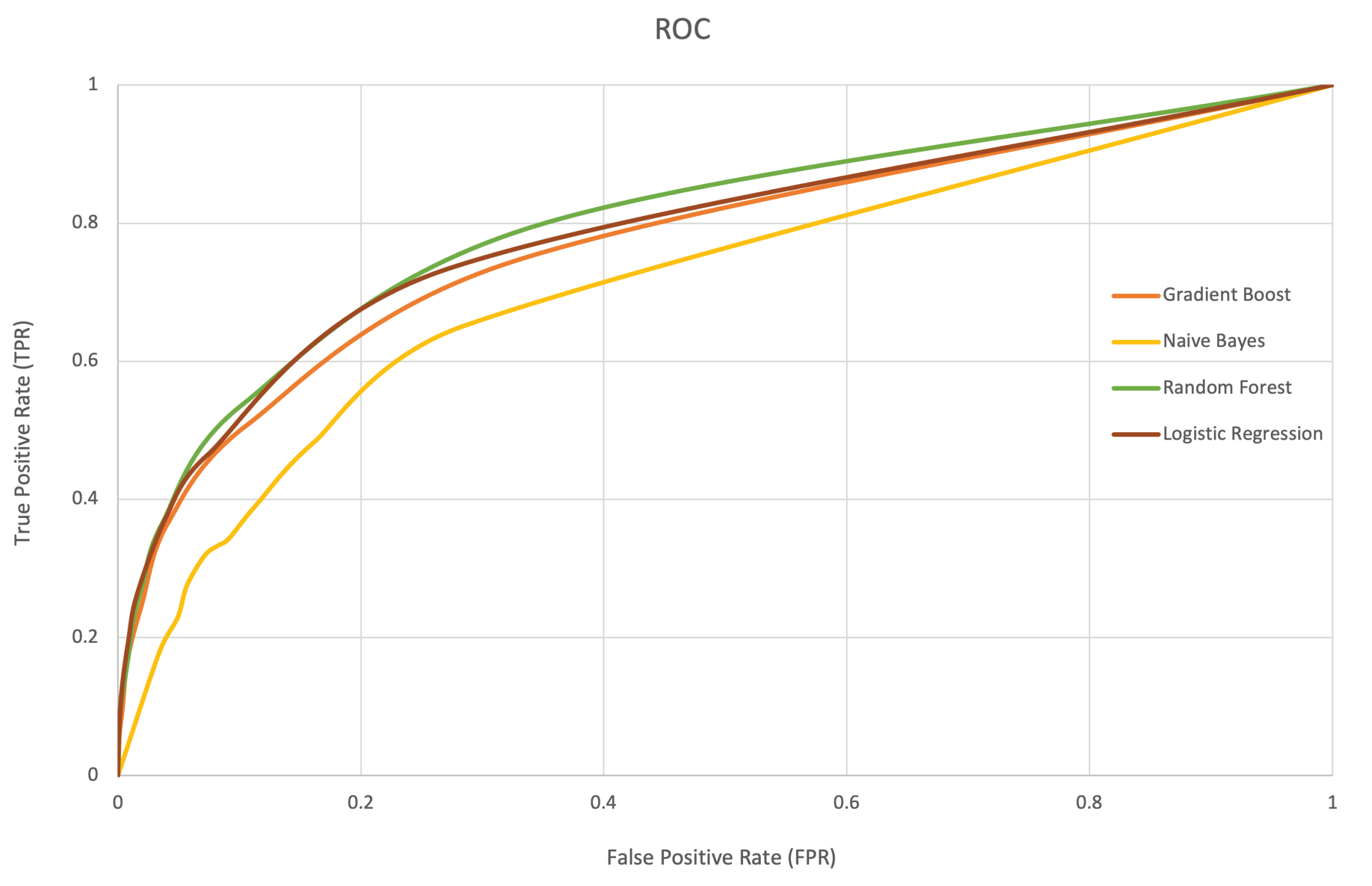
Table 3. Models Used for Analyses and Their Scores

	Logistic Regression	Naïve Bayes	Gradient Boost	Random Forest
AUC	82%	63%	81%	83%
Accuracy	89%	84%	89%	90%

AUC: area under the curve

The ROC curve shown in **Figure 1** denotes the performance of the classification models at all classification thresholds. This curve plots 2 parameters: true-positive rate and false-positive rate. The green curve (random forest model) covers slightly more area; thus, the AUC was highest with that model and obtained better results and the worst results were obtained by the naïve Bayes model.

Figure 1. ROC Curves for the Models



ROC, receiver operator characteristics

RESULTS (cont'd)

In the error analysis of the random forest model, 9% of negative cases (type II errors) were incorrectly predicted vs 26% of positive cases (type I errors). This indicates a higher rate of false positives than false negatives. The prevalence of type II errors is particularly concerning as it signifies the model's failure to identify existing cases, which may result in missed diagnoses and inadequate treatment. The model demonstrates better performance in minimizing type II errors than type I errors.

Table 4 shows the confusion matrix. This model identified 514,997 patients (12%) from the at-risk cohort as likely to have undiagnosed NASH, approximately 125 times higher than the number of patients initially identified as NASH-positive.

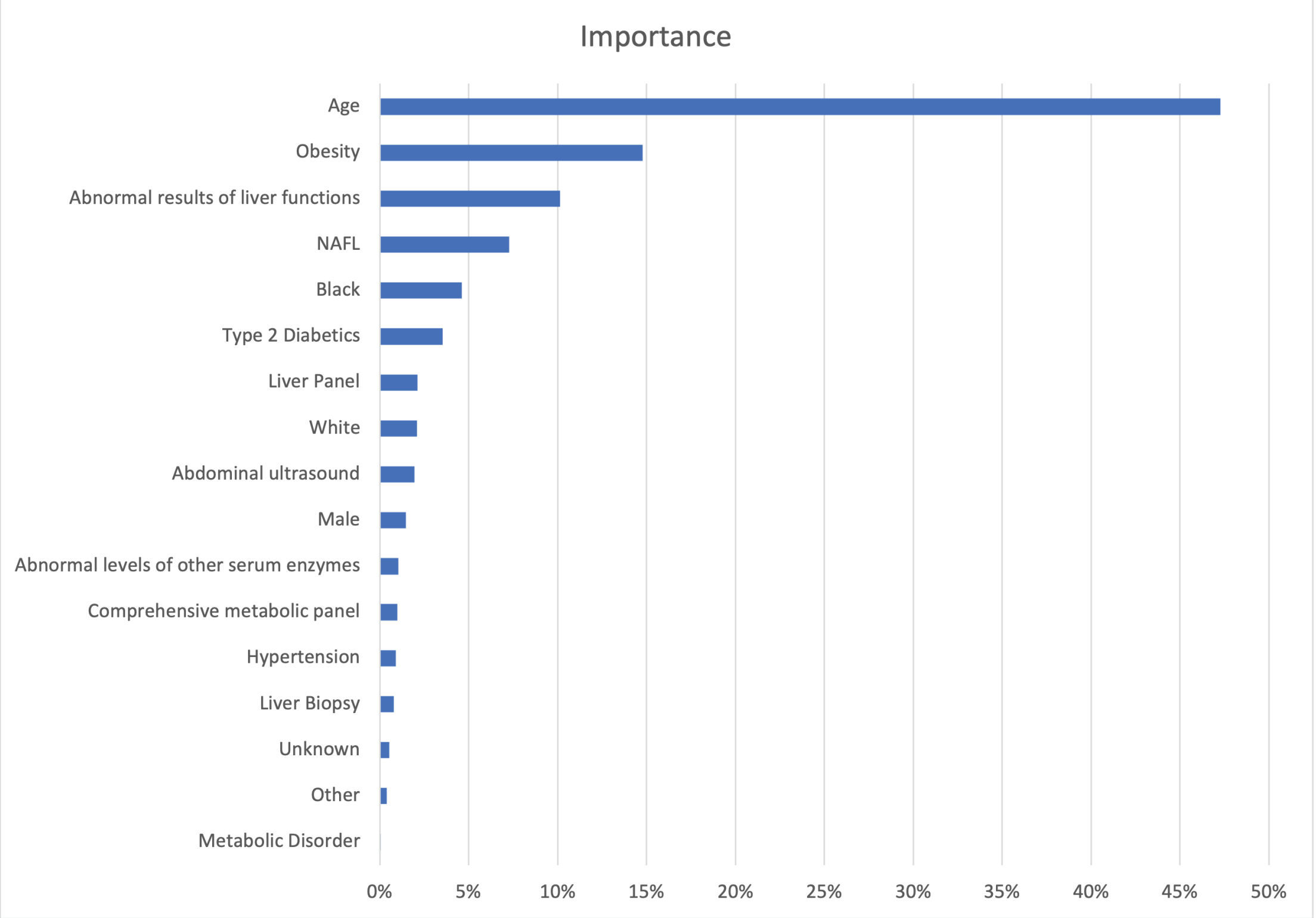
Table 4. Confusion Matrix and Risk Group with Potentials

Predicted Condition	Actual Condition		
	NASH	Non-NASH	At Risk
NASH	194 (TP)	70 (FP)	514,997 (12%)
Non-NASH	769 (FN)	7024 (TN)	3,668,015 (88%)

FN: false-negative; NASH: nonalcoholic steatohepatitis; TP: true-positive

Age, obesity, and abnormal liver function test results were the top determinants in assigning NASH probability. Figure 2 summarizes the top 17 features that accounted for 90% of the model's decrease in Gini impurity.

Figure 2. Feature Importance Measured by Contribution to Decrease in Gini Impurity



NAFL: nonalcoholic fatty liver

CONCLUSION

Machine learning algorithms can effectively identify patients with potential undiagnosed NASH from large at-risk populations using medical claims data. This approach could serve as an initial screening tool to select patients for further diagnostic evaluation and clinical management, potentially improving early detection and treatment of NASH.

REFERENCES

[1] Haas ME, Pirruccello JP, Friedman SN, et al. Machine learning enables new insights into genetic contributions to liver fat accumulation. *Cell Genom.* 2021;1(3).

[2] Docherty M, Regnier SA, Capkun G, et al. Development of a novel machine learning model to predict presence of nonalcoholic steatohepatitis. *J Am Med Informatics Assoc.* 2021;28(6):1235-41.

[3] Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA.* 2018;319(13):1317-1318.

[4] Quer G, Arnaout R, Henne M, Arnaout R. Machine learning and the future of cardiovascular care: JACC state-of-the-art review. *J Am College Cardiol.* 2021;77(3):300-313.

145 Hudson St., Suite 205
New York, NY 10013 USA
cdanyc.com

Columbia Data Analytics

Presented at ISPOR 2025
May 12-16, 2025, Montreal, Quebec, Canada