Risk prediction of fatty liver disease by machine learning using physical and blood biomarkers independent of liver biopsy

Weihong Zhou[1*], Yuan Hong Sun[2], Wendy Huang[3], Qijian Liu[2], Nathan Yee Lee[4], Yousef Yasin[2], Zhong Chen[1], Jing Wang[1], Pingqiang Cai[5], Zhong-ping Feng[2*], Kang Lee[2*]

1. Nanjing Drum Tower Hospital, The Affiliated Hospital of Nanjing University Medical School, Nanjing, China
2. University of Toronto, Canada
3. University of Limerick School of Medicine, Republic of Ireland
4. Western University, Canada
5. Nanjing University, Nanjing, China.

*Corresponding authors

Email list:

Weihong Zhou: njzhouwh@126.com

Yuan Hong Sun: billyuanhong.sun@mail.utoronto.ca

Wendy Huang: huangwy3@gmail.com

Qijian Liu: qijian.liu@mail.utoronto.ca

Nathan Yee Lee: nlee293@uwo.ca

Yousef Yasin: YousefYasin@nuralogix.ai

Zhong Chen: zdongy@sohu.com

Jing Wang: glyywj@163.com

Pingqiang Cai: pqcai@nju.edu.cn

Zhong-ping Feng: zp.feng@utoronto.ca

Kang Lee: kang.lee@utoronto.ca

Summary

Background: Non-alcoholic fatty liver disease (NAFLD) is becoming a major global burden of disease, but its diagnosis based on liver biopsy is invasive and expensive. This study aimed to investigate the feasibility of using patient physical and blood biomarkers to build computational models to predict NAFLD risks accurately.

Methods: We obtained a large dataset (N=81,552) containing patients' physical and blood measurements, and their NAFLD status based on the invasive liver biopsy. Using this dataset, the most important patient physical and blood biomarkers were extracted using feature selection techniques. Then, using the most important features, six different machine learning algorithms trained and tested models that predicted the NAFLD status. Finally, these models were validated with a pristine set of patient data never used in model training and testing, and SHAP (SHapley Additive exPlanations) analysis was performed on the final model.

Results: We identified the top 20 features for predicting NAFLD risk including Weight, BMI, Alanine aminotransferase, HDL Cholesterol, and Fasting blood sugar level. Our best machine learning model using the Extreme Gradient Boosting Decision Tree (XGB) technique achieved 84.77% area under curve (ROC AUC), 83.10% accuracy (87.88% sensitivity and 81.66% specificity) for predicting NAFLD.

Interpretation: Our findings suggest that machine learning models based on patient physical and blood biomarkers provide a viable, non-invasive, and inexpensive method to assess patients' risks for non-alcoholic fatty liver disease accurately.

Funding: Canadian Institute for Health Research and the Natural Science and Engineering Research Council of Canada.

Introduction

Non-alcoholic fatty liver disease (NAFLD) is becoming a global burden of disease with a worldwide prevalence of approximately 25%, and on the rise.[1,2] However, its diagnosis based on the gold standard liver biopsy is invasive and expensive. This study aimed to investigate the feasibility of using patient physical and blood biomarkers to build machine learning models to predict NAFLD risks accurately.

NAFLD is an umbrella term that includes both simple steatosis and non-alcoholic steatohepatitis (NASH) where the liver has undergone inflammation. It is characterized by the presence of fat cells exceeding 5% of the weight of the liver in the absence of significant alcohol consumption.[3] Typically, NAFLD progresses through the stages of simple steatosis, NASH, fibrosis, cirrhosis, and finally hepatocellular carcinoma.[3] Patients with simple steatosis have a low risk of progression and thus early diagnosis of NAFLD is beneficial to prevent disease progression and even reversal of the disease.[4]

Early screening of NAFLD has been challenging due to the lack of symptoms in patients. A histological evaluation from liver biopsy is the gold standard to identify NAFLD.[5] However, this invasive procedure is hampered by its cost and procedure-related complications.[5] The less invasive procedures such as measuring a single blood biomarker (e.g., specific cytokines as markers of cell apoptosis) and imaging tests (ultrasound, computerized tomography, and magnetic resonance imaging) are unreliable relative to liver biopsy.[5–7] Thus, recent attention has been directed towards the development of statistical indexes using multiple surrogate biomarkers to screen NAFLD, which delivered promising results.[8–11] However, the accuracy of these indexes relies heavily on the diagnosis of diabetes, which has a high level of comorbidity with NAFLD.[12] Thus, they have limited utility for determining NAFLD risks among patients without diabetes.

Recently, advanced machine learning techniques have been employed to develop models to predict disease risks[13,14] due to their ability to capture the complex multi-dimensional and non-linear relations between variables. Further, the availability of large clinical datasets offers adequate statistical power for developing robust and generalizable computational models of disease risks. Consequently, researchers have developed computational models of the risks for diseases such as diabetes and SARS-CoV-2 with high accuracy (e.g., over 80% Area under the Curve or AUC).[14,15,16] However, whether a similar approach can be used to predict patients' NAFLD risks with similarly high accuracy is unknown, which is the focus of this study.

Here we examined the feasibility of using data extracted from patient records along with an ensemble of advanced machine learning techniques to predict NAFLD risks. We capitalized on the availability of a large clinical dataset (N=81,552) of patients with or without NAFLD based on liver biopsy, selected from a clinical dataset of over 1.5 million patients across Canada. We also retrieved the patients' results of concurrent serum bloodwork and other relevant clinical information (e.g., BMI; henceforth referred to as patient biomarkers). Following the modern machine learning convention, we first applied multiple features selection techniques to select top patient biomarkers that are potentially predictive of NAFLD. Using the top features, we used such machine learning techniques as Logistic Regression, Random Forest (RF), Multilayer Perceptron (MLP), and Extreme gradient boosting decision trees (XGB) to train models to classify patients with or without NAFLD based on their liver biopsy results. We randomly divided the patient dataset into 80% for training, 10% for internal testing, and 10% never used in training and testing as an external validation set for model evaluation.

We hypothesized that compared to the existing indexes that rely heavily on diabetes diagnosis, without such information, our best machine learning models would still be able to

predict patients' NAFLD status with comparable, if not better, performance, and our model performance would generalize well to an external pristine validation dataset. This is because we use an extensive dataset, a comprehensive feature selection process to identify the most important patient biomarkers, multiple machine learning techniques, and the train-test-validate scheme.

We also aimed to provide explanations of model performance by using the SHAP (SHapley Additive exPlanations). SHAP is derived from the Shapley importance values, which was initially developed mathematically for fairly distributing benefits among collaborators.[17] It has become one of the important tools for explaining machine learning results.[18–20] Specifically, the SHAP values would allow us to identify the most important patient biomarkers contributing to our models' accurate prediction of NAFLD status. We hypothesized that the important biomarkers for predicting NAFLD found by our analysis would match with those found in the existing literature (e.g., BMI,[21–23] HDL cholesterol,[24] triglyceride level,[24,25] blood pressure, fasting blood sugar level[26-29]).

Methods

Dataset:

We derived our dataset from the electronic medical record (EMR) of over 1.5 million patients established for chronic disease surveillance by the Canadian Primary Care Sentinel Surveillance Network (CPCSSN).[30]

The derived dataset contains information from 120,376 patients. There are 37 patient features in the dataset containing various demographics, physical, and bloodwork measurements, and the predicted outcome is the diagnosis of NAFLD based on liver biopsy. After removing patients with missing values or data entry errors, the sample size was reduced to 81,552 (see

Figure 1 for the data selection process). There were 62,553 samples without NAFLD (0) and 18,999 samples with NAFLD (1). Out of these patients, 584 patients were diagnosed with cardiovascular diseases and 307 with diabetes. See Table 1 for detailed information. Note that these diagnoses were not used as predictors in the following machine learning process.

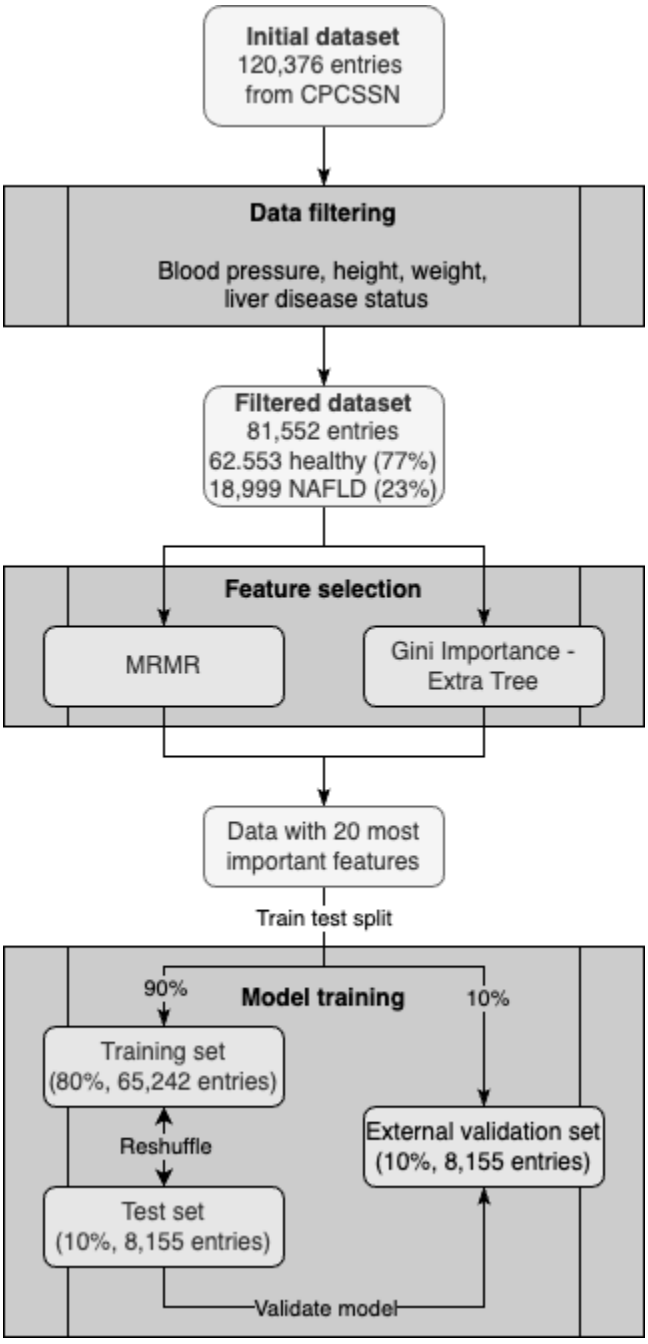

Figure 1: Flowchart describing the process of this study

Table 1: Participants demographics and the mean and standard deviations of their physical and blood biomarkers.

Patient Feature	Value
Total Sample	81,552
Sex, n	
Male	42,049
Female	39,503
Physical Measurements, mean ± SD	
Age at time of measurement (years)	40.0 ± 14.7
Height (cm)	166.6 ± 8.6
Weight (kg)	65.8 ± 12.9
BMI (kg/m^2)	23.6 ± 3.5
Waist Circumference (cm)	81.0 ± 16.4
Waist to Height Ratio	0.5 ± 0.1
Blood Biomarkers, mean ± SD	
Triglycerides	1.3 ± 1.1
Platelet Count	231.8 ± 53.9
High Density Lipoprotein Cholesterol (HDL)	1.3 ± 0.4

Hemoglobin	146.2 ±15.8
Average Hemoglobin Concentration	330.3 ± 8.8
Red Blood Cell	4.8 ± 0.5
Alanine Aminotransferase	24.2 ± 21.6
Fasting Blood Sugar	0.1 ± 0.3
White Blood Cell Count	6.2 ± 1.5
Neutrophil Count	3.6 ± 1.2
Lymphocyte Count	2.0 ± 0.6
Aspartate Aminotransferase	21.1 ± 12.1
Eosinophil Count	0.1 ± 0.1
Systolic Blood Pressure	123.5 ± 16.3
Diastolic Blood Pressure	75.7 ± 10.0
Cholesterol	4.6 ± 0.9
Low Density Lipoprotein Cholesterol (LDL)	2.7 ± 0.7
Alkaline Phosphatase	63.4 ± 19.4
Total Bilirubin	11.5 ± 5.1
Glycated Hemoglobin	5.8 ± 0.8
Hemoglobin A1c	0.1 ± 0.3

Albumin	45.5 ± 2.3
Globulin	29.6 ± 3.6
Total Protein	75.1 ± 4.1
Lactate Dehydrogenase	187.3 ± 37.5
Cholinesterase	9.1 ± 1.9
Total Bile Acid	3.4 ± 3.5
Leucine Aminopeptidase	51.9 ± 10.9

Feature Selection:

To find the most important features (i.e., patient biomarkers including demographics, physical and blood measurements) that can accurately predict a patient's NAFLD status, we performed feature selection on the whole patient dataset using two feature selection techniques: (1) Minimum Redundancy Maximum Relevance (MRMR) and (2) Gini importance in the Extra Tree Classifier (ETC). See Appendix A for details.

We specifically excluded the diagnosis of diabetes and cardiovascular disease from the feature selection process such that our trained computational models would not depend on such diagnosis and therefore can be used with healthy individuals.

Machine Learning:

We used the six machine learning algorithms to train models to predict a patient's NAFLD status: Gaussian Naïve Bayes (GNB), Logistic Regression (LR), Random Forest,

Extreme Gradient Boosting decision trees (XGB), Support Vector Machine (SVM), and Multilayer Perceptron (MLP). See Appendix B for details.

For all the model trainings, we split the whole patient dataset into training, internal testing, and external validation datasets, which was 80% (n=65,242), 10% (n=8,155), and 10% (n=8,155) of the entire dataset respectively. The machine learning training was done on the training dataset and was tested on the internal test set, and evaluated on the external validation set. The training, test, and validation sets were kept the same across all different algorithms for fair comparison. The predicted variable was patients' NAFLD status based on liver biopsy.

Once a model was trained and tested, the training and internal test sets were recombined, and randomly split into new training and test set in the same way as before, with training being 80% of the entire dataset, and internal testing being 10% of the entire dataset. The recombination was performed 50 times to obtain 50 different models. The 50 different recombination of the datasets were also kept the same across the different algorithms.

These 50 models were then evaluated against the external validation dataset that had never been used in model training and testing. Thus, the validation results provide an unbiased evaluation of model performance, which in essence, is an assessment of the generalizability of the trained and tested models' performance to a pristine patient dataset.

We used the accuracy, Area Under Curve (AUC) score of Receiver Operating Characteristic (ROC) curve, sensitivity, and specificity scores for model performance evaluation (see Appendix C for details). Note that all the scores are weighted to reflect the class imbalance between the healthy and NAFLD-positive samples.

Model explanation

We used SHapley Additive exPlanations (SHAP) values to provide an explanation of the best models. It is derived from Shapley's Value, originally developed from cooperative game theory where each cooperative game player is assigned a unique distribution (among all players) of a total surplus generated by the coalition of all players.[19] The Shapley's Value measures the relative importance of each player's contributions to a game in a principled and mathematical manner.

The formula of Shapley's equation is as follows:

$$\varphi_i(v) = \sum_{S \in N \setminus i} \frac{|S|!\,(N - |S| - 1)!}{N!} \big(v(S \cup i) - v(S)\big)$$

Where $\frac{|S|!(N-|S|-1)!}{N!}$ is the weight and $\big(v(S \cup i) - v(S)\big)$ is the marginal contribution.

We used our best models to compute the SHAP values for the patient biomarkers used in the models.

Results

Feature Selection:

From the original dataset, we extracted 37 patient physical and blood biomarkers. Out of the 37 biomarkers, we selected 20 to be used as the most important features for predicting NAFLD status based on patient liver biopsy. The feature selection process was done on the processed dataset using Minimum Redundancy Maximum Relevance (MRMR) and Gini importance of an Extra Tree Classifier. The top 20 features selected by each of the two methods can be found in Appendix A.

After combining the two lists, the top 20 features, in order, became: Weight, Diastolic blood pressure, Triglycerides, Platelet count, Systolic blood pressure, Height, High-density

lipoprotein (HDL) cholesterol, Hemoglobin, BMI, Red blood cell, Alanine aminotransferase, Fasting blood sugar, White blood cell count, Age, Neutrophil count, Lymphocyte count, Aspartate aminotransferase, average hemoglobin concentration, Gender, Eosinophil count (for the means and standard deviations of these features, see Table 1).

Machine Learning:

Table 2 shows the performance of the models tested against the external validation dataset using all top 20 features (for results of testing against the internal testing sets, see Appendix B). All machine learning algorithms performed significantly above chance to predict the NAFLD status and their performance generalized well to the external validation dataset that was not used in the model training and testing. In fact, all algorithms produced accuracies to be all significantly above 80%, with Random Forest and MLP to be top performing. However, in terms of AUC, XGBoost, SVM, and Logistic Regression were ranked on the top and significantly above 80%. In terms of sensitivity and specificity that should ideally be balanced, XGB and Logistic Regression were ranked on the top. When all these performance measures are considered together, XGB was the best performing algorithm for the data at hand. Figure 2 shows the Receiver Operating Characteristic Curves of the 50 XGB models evaluated on the external validation set.

Table 2: Model results evaluated on the external validation set, showing the mean, standard deviation, and 95% confidence interval for our various performance measures across the 50 ensembles, for each machine learning technique trained on the top 20 features for predicting NAFLD risk. (The best model overall is highlighted in bold).

	Accuracy Score	95% CI	AUC ROC	95% CI	Specificity	95% CI	Sensitivity	95% CI
GNB	80.40% ± 0.19%	80.35%-80.46%	80.08% ± 0.01%	80.06%-80.10%	80.68% ± 0.29%	80.60%-80.77%	79.48% ± 0.18%	79.42%-79.53%
LR	83.60% ± 0.04%	83.59%-83.61%	84.25% ± 0.04%	84.24%-84.26%	80.03% ± 0.05%	83.01%-83.04%	85.47% ± 0.10%	85.45%-85.50%
RF	86.41% ± 0.13%	86.38%-86.45%	77.59% ± 0.22%	77.53%-77.66%	94.11% ± 0.11%	94.08%-94.14%	61.08% ± 0.44%	60.95%-61.20%
SVM	82.82% ± 0.11%	82.79%-82.85%	84.54% ± 0.10%	84.51%-84.69%	81.31% ± 0.16%	81.27%-81.36%	87.77% ± 0.18%	87.72%-87.82%
MLP	86.37% ± 0.21%	86.31%-86.43%	79.56% ± 0.83%	79.32%-79.79%	92.32% ± 0.72%	92.11%-92.52%	66.79% ± 2.30%	66.14%-67.45%
XGB	**83.10% ± 0.12%**	**83.07%-83.14%**	**84.77% ± 0.14%**	**84.73%-84.81%**	**81.66% ± 0.13%**	**81.62%-81.69%**	**87.88% ± 0.26%**	**87.80%-87.95%**

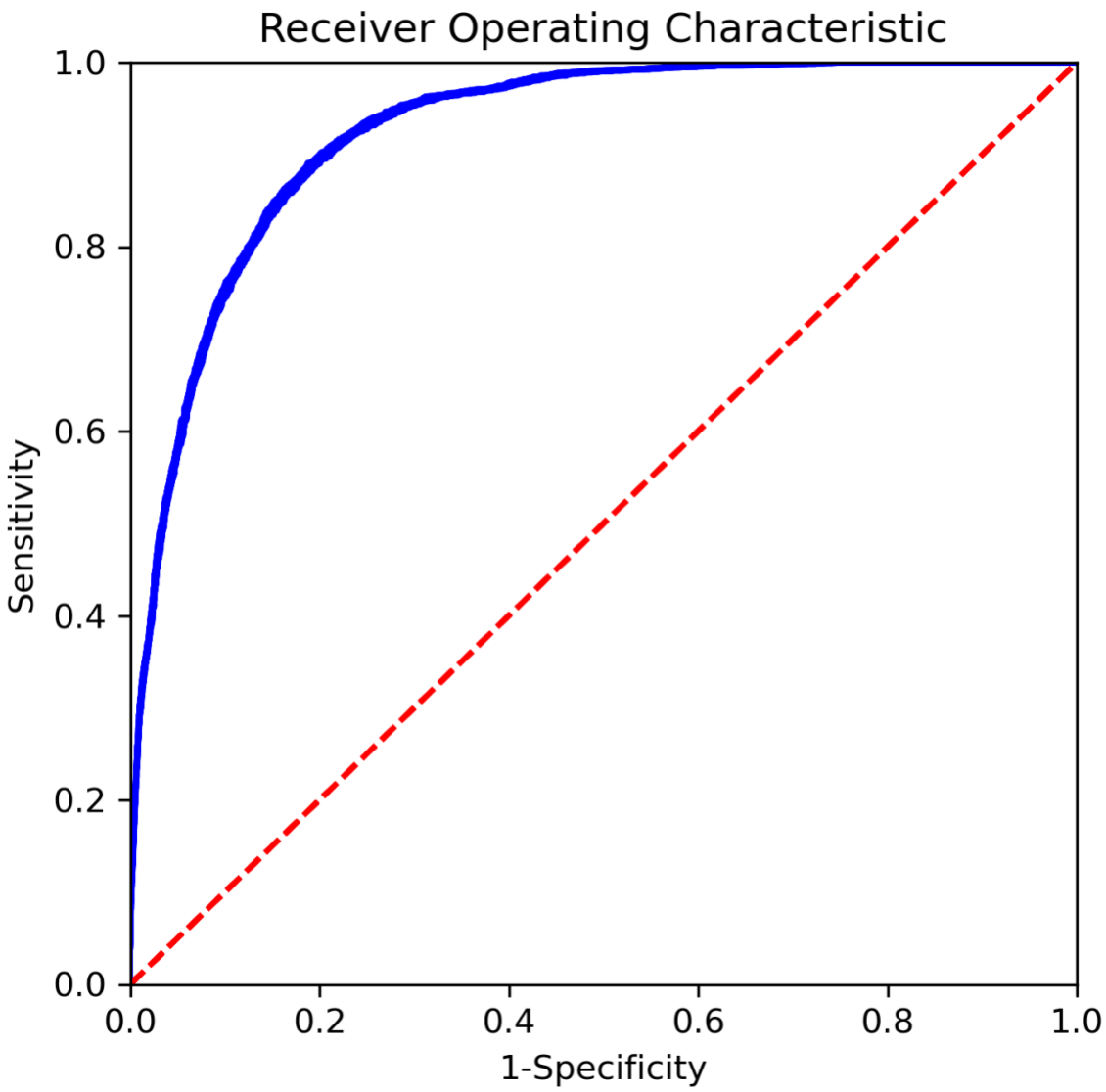

Figure 2. Receiver Operating Characteristic Curves of 50 XGB models evaluated on the external validation set (blue). The dashed red line represents an AUC ROC of 50% or a random classifier.

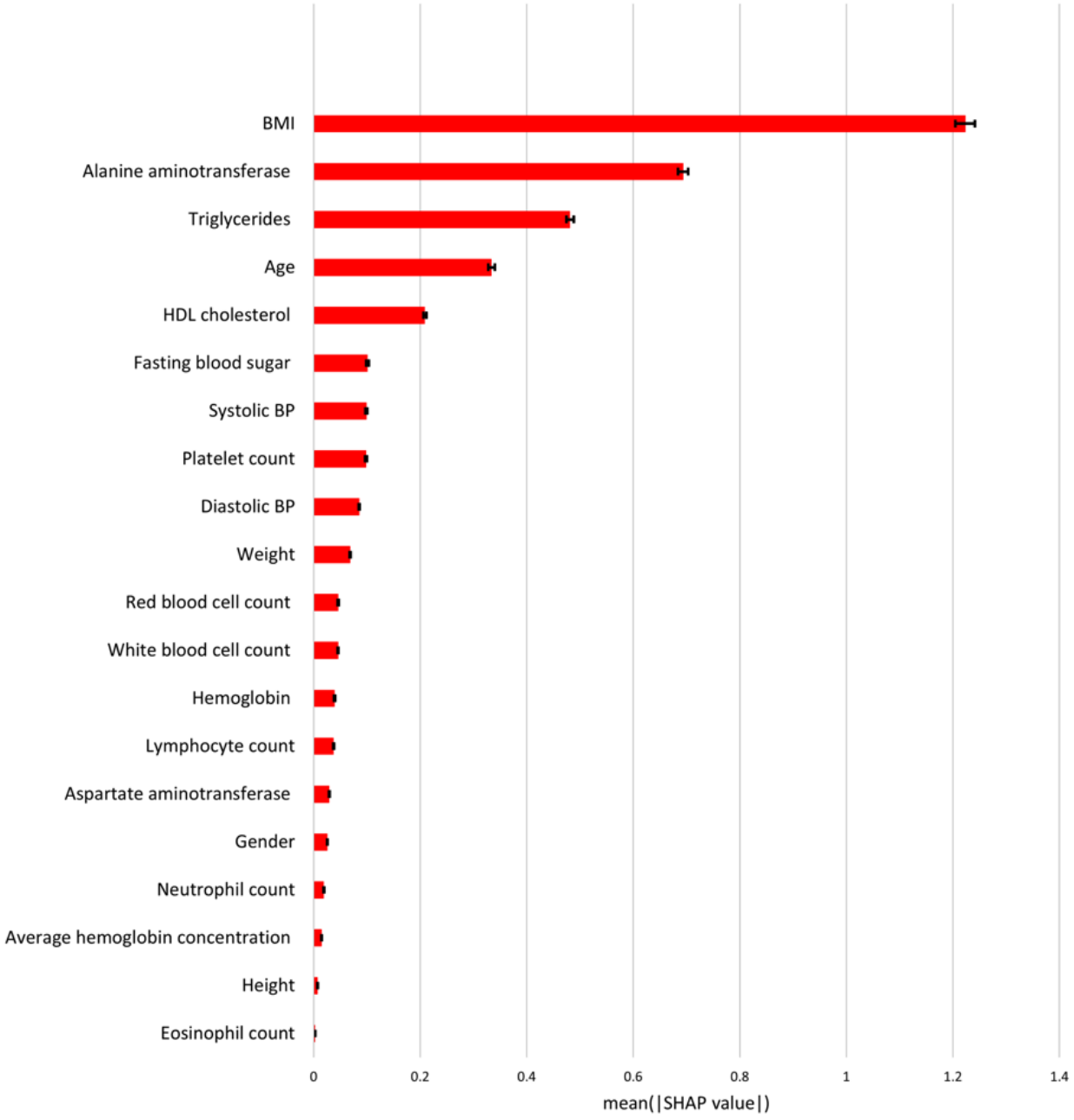

Figure 3: Mean SHAP (Shapley Additive Explanation) value and 95% confidence interval for each of the top 20 features of XGB Model. The higher SHAP values represent higher feature importance.

Figure 3 shows the mean SHAP value and the 95% confidence intervals for each of the top 20 features for the XGB Models. BMI was the most important factor, followed by alanine aminotransferase, and triglycerides (Also see Appendix D for numeric values).

Model implementation:

To showcase the functionality of the NAFLD risk prediction model in digital health, a web implementation was developed (https://kangleelab-surveys.herokuapp.com/nafld). Users can enter their physical measurements and blood biomarkers into the form and submit. These values are sent to our best XBG machine learning model to compute a likelihood of NAFLD, which is shown as a percentage on the results page.

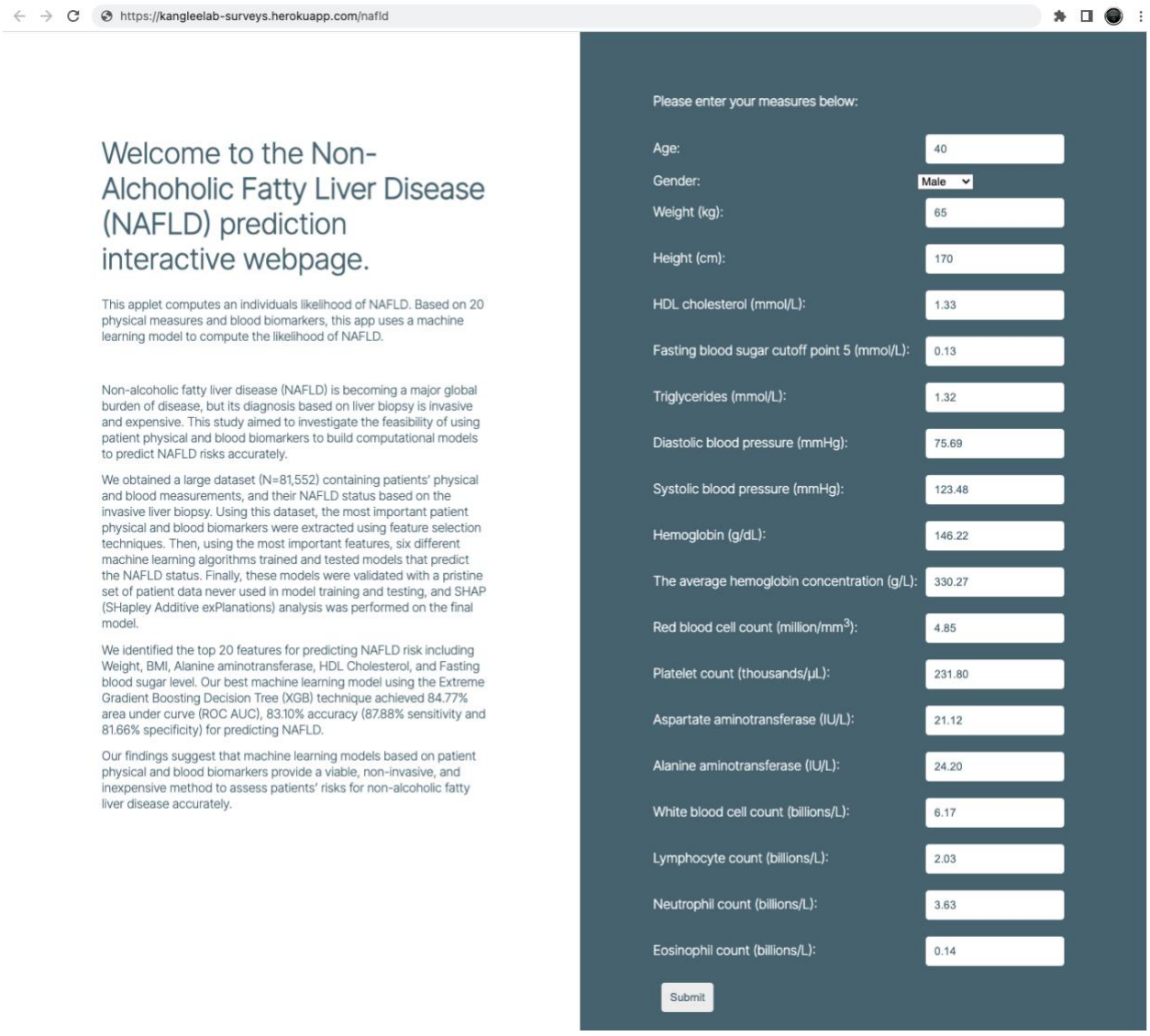

Figure 4: Input page for the NAFLD risk online assessment tool

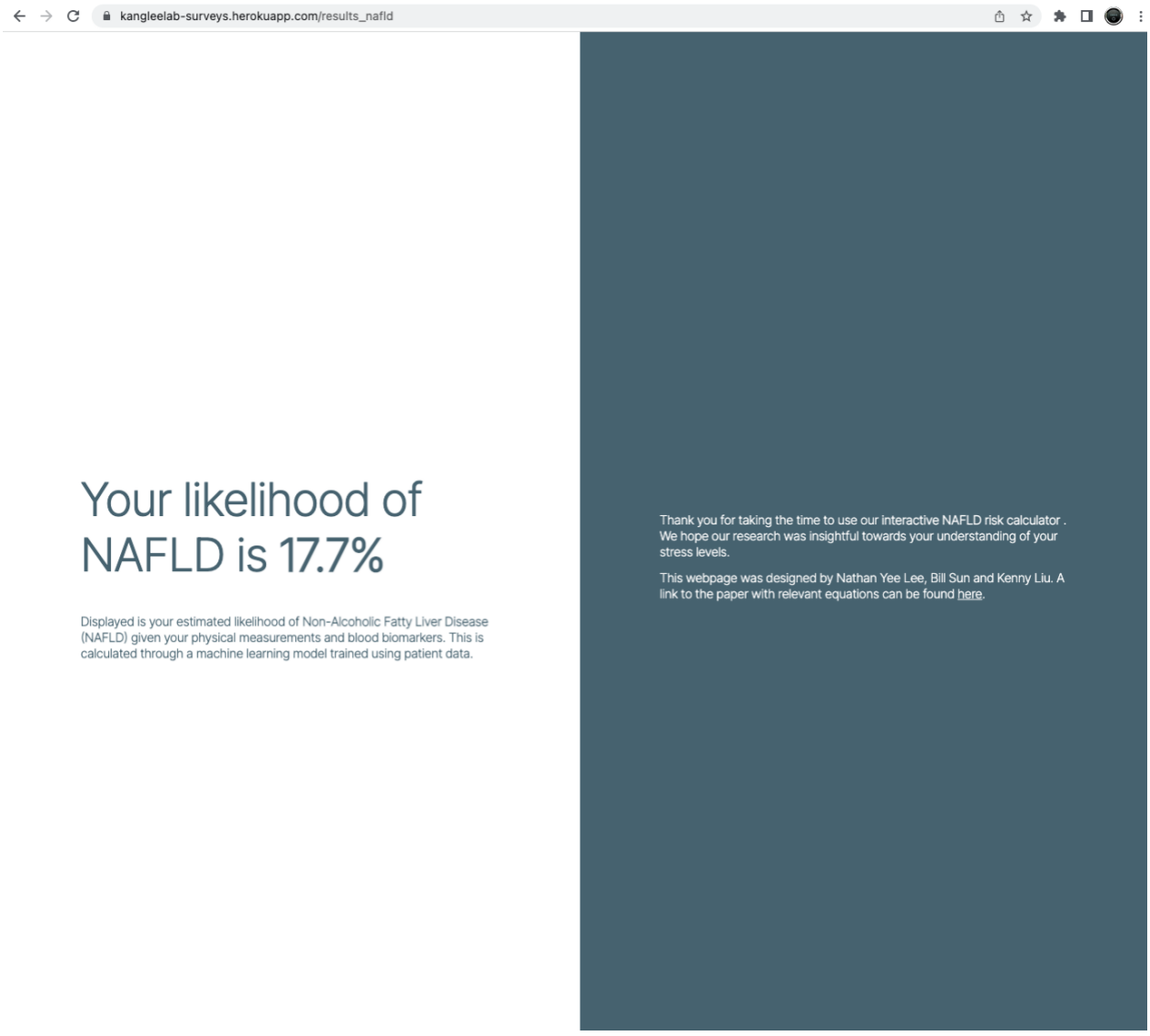

Figure 5: Results page for the NAFLD risk online assessment tool

Discussion

We investigated the feasibility of developing a machine learning based model to predict patients' NAFLD using their physical and blood biomarkers with high accuracy. To that end, we first identified the top 20 most important features that included weight, BMI, Alanine aminotransferase, HDL Cholesterol, fasting blood sugar, triglycerides, diastolic blood pressure, hemoglobin, systolic blood pressure and red blood cell count. Then, we used six advanced

machine learning techniques on the 20 most important features to predict patients' NAFLD status. Our results indicated that all machine learning algorithms were able to predict the NAFLD status with significantly above 80% accuracy, and most with significantly above 80% AUC with high specificity and sensitivity.

To the best of our knowledge, this study was the first of its kind to use multiple advanced machine learning algorithms along with the readily available biomarkers to produce NAFLD prediction models that are highly accurate, specific, and sensitive. Such models also generalize well to a pristine dataset that has never been used in model training and testing, suggesting our models' robustness. The success of our models benefited from not only the use of multiple machine learning algorithms that allow us to identify the best algorithm for the data at hand but also the large patient dataset and the feature selection procedures as well as the training-testing-validation scheme. Our approach thus allows us to overcome various shortcomings of the existing computational methods to predict NAFLD risks that either use rather small datasets, analyze the data using a single analytic approach (e.g., logistic regression), fail to use external validation data to test model generalizability, or rely on too heavily on the diagnosis of diabetes.

In addition, our models are also highly implementable. Our models are small (3412 kilobytes for the XGB model) and can be readily stored in a desktop computer, a smart device (e.g., a smartphone) or a website with a simple user interface. Because the biomarkers needed for our models are readily available from patients who take a routine blood serum test, they can enter the values of these biomarkers to our models to obtain the likelihood of having NAFLD from 0% to 100%. They could seek medical attention as soon as they show signs of being at high risk of NALFD. Physicians can also use the tool to identify patients that could be at high risk and require attention and further assessment. The tool, when further validated in the field, can also be

used to monitor medically diagnosed NAFLD patients after they undergo medical treatment and lifestyle related changes. The availability of such a NAFLD prediction tool will revolutionize how we monitor patient NAFLD risks and can play a significant role in the global effort to combat NAFLD which is quickly becoming an epidemic worldwide.[42]

The present findings also attest to the tremendous potentials of using the machine learning approach to predict the risks of such disease as NAFLD. In particular, we found that when all performance measures were considered together, the XGB technique produced the best performance with accuracy, AUC, specificity and sensitivity all to be significantly above 80%. The strength of XGB lies in the gradient tree boosting method, which uses decision trees as the base estimators and keeps adding them to form a stronger estimator through the greedy algorithm.[36] The usage of weaker base estimators to form a stronger estimator is also referred to as ensemble learning, which has the advantage of being more robust than having only one estimator. Perhaps for these reasons, the XGB algorithms produced, on balance, the best NAFLD prediction models in terms of accuracy, AUC, specificity, and sensitivity.

We hypothesized that our machine learning approach would be able to produce models with performance comparable or exceeding the existing NAFLD indexes. Indeed, our results confirmed this hypothesis. Compared to the hepatic steatosis index (HSI) developed by Lee et al., which is a formula for detecting NAFLD derived from a logistic regression model,[9] our XGB model (0.85) outperformed the HSI model (0.81) substantially when it comes to AUC. However, their HSI model was able to achieve higher sensitivity and specificity, which were 93.1% and 92.4% respectively, whereas our XGB model achieved 87.88% in sensitivity and 81.66% in specificity. In contrast, another study by Kotronen et al. used multivariate logistic regression analysis to build a NAFLD liver fat score to predict NAFLD status.[10] Its AUC was 0.86 in

validation, which only outperformed our XGB model slightly, but its sensitivity (84%) and specificity (69%) were significantly lower than our XGB model.

We also hypothesized that the most important features determined by SHAP would match those in existing literature, such as BMI,[21] HDL cholesterol,[24] triglyceride level,[24,25] blood pressure and fasting blood sugar,[26–29] which is also largely supported by our results. For example, our finding of weight and BMI to be the top two most important features in predicting risk for NAFLD aligns well with established association between obesity and NAFLD.[21] One large population study in the UK found a strong near linear relation between BMI and future risk of NAFLD/NASH with the risk being approximately 5-to-9-fold higher at BMIs of 30.0-32.5 kg/m^2 and increases to 10-to-14-fold at BMIs of 37.5-40.0 kg/m^2.[21] This association is likely the result of increased exogenous consumption of high fat diet over time leading to increased adipose tissue and free fatty acids being circulated to the liver.

The existing research has also established that HDL cholesterol,[24] triglyceride level,[24,25] blood pressure and fasting blood sugar are major components of Metabolic Syndrome.[26–29] There are strong associations between Metabolic Syndrome and NAFLD.[43–46] Approximately 90% of patients diagnosed with NAFLD have more than one component of Metabolic Syndrome and about 36% have three or more components.[43] Note that three out of the five components must be present to be considered as Metabolic Syndrome[45] with the fifth component being waist circumference which our study did not include due to lack of data. It remains unclear if Metabolic Syndrome precedes NAFLD or vice versa. Interestingly, NAFLD is still found in individuals without Metabolic Syndrome. Our results align closely with the existing findings by showing that the four components of Metabolic Syndrome are among the top ten most important biomarkers to predict NAFLD, confirming our second hypothesis.

The present study has several limitations. First, the current machine learning models only identify the presence of risk in NAFLD. However, NAFLD has several phases of severity, which include simple steatosis, steatohepatitis, fibrosis, cirrhosis, and hepatocellular carcinoma.[47] For greater accuracy and practicality, we need to develop models that can assess the risks in each phase using more detailed data in the future. Second, the present models were tested against a pristine dataset sampled from the same large dataset based on Canadian patients. Thus, the validation of our models is retrospective in nature and limited to Canadians. An important next step is to obtain prospective validation datasets to assess whether the accurate models based on the current dataset could be generalized to a new set of patients of different racial and ethnic groups from different countries.

Conclusion

We used multiple machine learning algorithms along with a large dataset of physical and biomarkers from patients with vs. without non-alcoholic fatty liver disease based on liver biopsy. We identified the top 20 features for predicting NAFLD risk including Weight, BMI, Alanine aminotransferase, HDL Cholesterol, and Fasting blood sugar level. Our best machine learning model, which uses the Extreme Gradient Boosting Decision Tree (XGB) technique, achieved 84.77% area under curve (ROC AUC), 83.10% accuracy (87.88% sensitivity and 81.66% specificity) for predicting NAFLD. Our findings demonstrated that machine learning models based on patient biomarkers provide a viable, non-invasive, and inexpensive method to detect patients' risks for non-alcoholic fatty liver disease. Such models when made widely available will play an important role in our effort to combat a health problem that is quickly becoming a major global burden of disease.

Acknowledgement

The authors like to acknowledge the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) for collecting and providing the data for this study. We also acknowledge the funders of this research, which include grants from the Canada Research Chairs Program to Kang Lee, and the Canadian Institute for Health Research to Zhong-ping Feng, and Nanjing Health Science Major Research Programs to Weihong Zhou (Grant number: ZDX21001).

Role of Funding Source

The research was funded in part by grants from the Canada Research Chairs Program to Kang Lee, and the Canadian Institute for Health Research to Zhong-ping Feng, and Nanjing Health Science Major Research Programs to Weihong Zhou (Grant number: ZDX21001).
The funding sources did not contribute to the research.

Contributions

- Weihong Zhou*: conceptualisation, project administration, investigation, data collection, writing – original draft, writing – review & editing, resources
- Yuan Hong Sun: methodology, formal analysis, software, writing – original draft, writing – review & editing, validation
- Wendy Huang: data curation, methodology, writing – original draft
- Qijian Liu: formal analysis, software, visualization, writing – original draft
- Nathan Yee Lee: software, validation
- Yousef Yasin: formal analysis, validation

- Zhong Chen: writing – review & editing, validation
- Jing Wang: writing – review & editing, validation
- Pingqiang Cai: writing – review & editing, validation
- Zhong-ping Feng*: conceptualisation, project administration, investigation, writing – review & editing, resources
- Kang Lee*: conceptualisation, methodology, project administration, supervision, writing – review & editing, resources

*Corresponding authors

References

[1] Zeigerer A. (2021). NAFLD - A rising metabolic disease. Molecular metabolism, 50, 101274. https://doi.org/10.1016/j.molmet.2021.101274

[2] Cotter, T. G., & Rinella, M. (2020). Nonalcoholic Fatty Liver Disease 2020: The State of the Disease. Gastroenterology, 158(7), 1851–1864. https://doi.org/10.1053/j.gastro.2020.01.052.

[3] J.-F. Dufour, R. Scherer, M.-M. Balp, S.J. McKenna, N. Janssens, P. Lopez, M. Pedrosa, The global epidemiology of nonalcoholic steatohepatitis (NASH) and associated risk factors–A targeted literature review, Endocr. Metab. Sci. 3 (2021) 1–19. doi:10.1016/j.endmts.2021.100089.

[4] G. Mazzolini, J.-P. Sowa, C. Atorrasagasti, Ö. Kücükoglu, W.-K. Syn, A. Canbay, Significance of Simple Steatosis: An Update on the Clinical and Molecular Evidence, Cells. 9 (2020) 2458. doi:10.3390/cells9112458.

[5] N. Chalasani, Z. Younossi, J.E. Lavine, M. Charlton, K. Cusi, M. Rinella, S.A. Harrison, E.M. Brunt, A.J. Sanyal, The Diagnosis and Management of Nonalcoholic Fatty Liver Disease: Practice Guidance From the American Association for the Study of Liver Diseases, Hepatology. 67 (2018) 328–357. doi:10.1002/hep.29367/suppinfo.

[6] A. Wieckowska, A.E. Feldstein, Diagnosis of nonalcoholic fatty liver disease: Invasive versus noninvasive, Semin. Liver Dis. 28 (2008) 386–395. doi:10.1055/s-0028-1091983.

[7] E. Cleveland, A. Bandy, L.B. VanWagner, Diagnostic challenges of nonalcoholic fatty liver disease/nonalcoholic steatohepatitis, Clin. Liver Dis. 11 (2018) 98–104. doi:10.1002/cld.716.

[8] G. Bedogni, S. Bellentani, L. Miglioli, F. Masutti, M. Passalacqua, A. Castiglione, C.

Tiribelli, The Fatty Liver Index: a simple and accurate predictor of hepatic steatosis in the general population, (2006). doi:10.1186/1471-230X-6-33.

[9] J.H. Lee, D. Kim, H.J. Kim, C.H. Lee, J.I. Yang, W. Kim, Y.J. Kim, J.H. Yoon, S.H. Cho, M.W. Sung, H.S. Lee, Hepatic steatosis index: A simple screening tool reflecting nonalcoholic fatty liver disease, Dig. Liver Dis. 42 (2010) 503–508. doi:10.1016/j.dld.2009.08.002.

[10] A. Kotronen, M. Peltonen, A. Hakkarainen, K. Sevastianova, R. Bergholm, L.M. Johansson, N. Lundbom, A. Rissanen, M. Ridderstråle, L. Groop, M. Orho-Melander, H. Yki-Järvinen, Prediction of Non-Alcoholic Fatty Liver Disease and Liver Fat Using Metabolic and Genetic Factors, Gastroenterology. 137 (2009) 865–872. doi:10.1053/j.gastro.2009.06.005.

[11] J. Wang, C. Xu, Y. Xun, Z. Lu, J. Shi, C. Yu, Y. Li, ZJU index: A novel model for predicting nonalcoholic fatty liver disease in a Chinese population, Sci. Rep. 5 (2015) 1–10. doi:10.1038/srep16494.

[12] Glass LM, Hunt CM, Fuchs M, Su GL. Comorbidities and Nonalcoholic Fatty Liver Disease: The Chicken, the Egg, or Both? Fed Pract. 2019 Feb;36(2):64-71. PMID: 30867626; PMCID: PMC6411365.

[13] Sun, Y.H., Luo, H. & Lee, K. A Novel Approach for Developing Efficient and Convenient Short Assessments to Approximate a Long Assessment. Behavior Research Methods (2022). https://doi.org/10.3758/s13428-021-01771-7

[14] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in Genetics, 9. https://doi.org/10.3389/fgene.2018.00515

[15] Xie, J., Shao, H., Shan, T., Jing, S., Shi, Y., Wang, J., … Liu, Y. (2022). Validation of type 2 diabetes subgroups by simple clinical parameters: a retrospective cohort study of NHANES data from 1999 to 2014. BMJ Open, 12(3). doi:10.1136/bmjopen-2021-055647

[16] Zoabi, Y., Deri-Rozov, S., & Shomron, N. (2021). Machine learning-based prediction of COVID-19 diagnosis based on symptoms. Npj Digital Medicine, 4(1). https://doi.org/10.1038/s41746-020-00372-6

[17] Shapley, L. S. A value for *n*-person games. Contributions to the Theory of Games, 2, 307.

[18] Ghorbani, A., & Zou, J. (2019). Data shapley: Equitable Valuation of Data for Machine Learning. In *International Conference on Machine Learning* (pp. 2242–2251). PMLR.

[19] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. CoRR, abs/1705.07874. Ανακτήθηκε από http://arxiv.org/abs/1705.07874

[20] Smith, M. & Alvarez, F. (2021). Identifying mortality factors from machine learning using Shapley Values – a case of covid19. Expert Systems with Applications, 176, 114832. https://doi.org/10.1016/j.eswa.2021.114832

[21] E. Fabbrini, S. Sullivan, S. Klein, Obesity and Nonalcoholic Fatty Liver Disease: Biochemical, Metabolic and Clinical Implications, Hepatology. 51 (2010) 679. doi:10.1002/HEP.23280.

[22] A. Katrina Loomis, S. Kabadi, D. Preiss, C. Hyde, V. Bonato, M.S. Louis, J. Desai, J.M.R. Gill, P. Welsh, D. Waterworth, N. Sattar, Body mass index and risk of nonalcoholic fatty liver disease: Two electronic health record prospective studies, J. Clin. Endocrinol. Metab. 101 (2016) 945–952. doi:10.1210/JC.2015-3444/SUPPL_FILE/JC-15-3444.PDF.

[23] R. Fan, J. Wang, J. Du, Association between body mass index and fatty liver risk: A dose-

response analysis, Sci. Reports 2018 81. 8 (2018) 1–7. doi:10.1038/s41598-018-33419-6.

[24] K.T. Wu, P.L. Kuo, S. Bin Su, Y.Y. Chen, M.L. Yeh, C.I. Huang, J.F. Yang, C.I. Lin, M.H. Hsieh, M.Y. Hsieh, C.F. Huang, W.Y. Lin, M.L. Yu, C.Y. Dai, H.Y. Wang, Nonalcoholic fatty liver disease severity is associated with the ratios of total cholesterol and triglycerides to high-density lipoprotein cholesterol, J. Clin. Lipidol. 10 (2016) 420-425.e1. doi:10.1016/J.JACL.2015.12.026.

[25] A.J. Amor, V. Perea, Dyslipidemia in nonalcoholic fatty liver disease, Curr. Opin. Endocrinol. Diabetes Obes. 26 (2019) 103–108. doi:10.1097/MED.0000000000000464.

[26] A. López-Suárez, J.M.R. Guerrero, J. Elvira-González, M. Beltrán-Robles, F. Cañas-Hormigo, A. Bascuñana-Quirell, Nonalcoholic fatty liver disease is associated with blood pressure in hypertensive and nonhypertensive individuals from the general population with normal levels of alanine aminotransferase, Eur. J. Gastroenterol. Hepatol. 23 (2011) 1011–1017. doi:10.1097/MEG.0B013E32834B8D52.

[27] L.Y. Qian, J.F. Tu, Y.H. Ding, J. Pang, X. Da Che, H. Zou, D.S. Huang, Association of blood pressure level with nonalcoholic fatty liver disease in nonhypertensive population Normal is not the new normal, Med. (United States). 95 (2016). doi:10.1097/MD.0000000000004293.

[28] R. Lorbeer, C. Bayerl, S. Auweter, S. Rospleszcz, W. Lieb, C. Meisinger, M. Heier, A. Peters, F. Bamberg, H. Hetterich, Association betweenMRI-derived hepatic fat fraction and blood pressure in participants without history of cardiovascular disease, J. Hypertens. 35 (2017) 737–744. doi:10.1097/HJH.0000000000001245.

[29] J.C. Bae, E.J. Rhee, W.Y. Lee, S.E. Park, C.Y. Park, K.W. Oh, S.W. Park, S.W. Kim, Combined Effect of Nonalcoholic Fatty Liver Disease and Impaired Fasting Glucose on

the Development of Type 2 DiabetesA 4-year retrospective longitudinal study, Diabetes Care. 34 (2011) 727–729. doi:10.2337/DC10-1991.

[30] Garies, S., Birtwhistle, R., Drummond, N., Queenan, J., & Williamson, T. (2017). Data Resource Profile: National Electronic Medical Record Data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN). International Journal of Epidemiology, 46(4). https://doi.org/10.1093/ije/dyw248

[31] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 1226–1238. doi:10.1109/TPAMI.2005.159.

[32] sklearn.ensemble.ExtraTreesClassifier — scikit-learn 0.24.2 documentation, 2020. (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html (accessed September 23, 2021).

[33] B.H. Menze, B.M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, F.A. Hamprecht, A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data, BMC Bioinforma. 2009 101. 10 (2009) 1–16. doi:10.1186/1471-2105-10-213.

[34] 1.9. I Bayes — scikit-learn 0.24.2 documentation, (2020). https://scikit-learn.org/stable/modules/naive_bayes.html (accessed September 23, 2021).

[35] C.-Y.J. Peng, K.L. Lee, G.M. Ingersoll, An Introduction to Logistic Regression Analysis and Reporting, Https://Doi.Org/10.1080/00220670209598786. 96 (2010) 3–14. doi:10.1080/00220670209598786.

[36] A. Cutler, D.R. Cutler, J.R. Stevens, Random Forests, Ensemble Mach. Learn. (2012)

157–175. doi:10.1007/978-1-4419-9326-7_5.

[37] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, (2016). doi:10.1145/2939672.2939785.

[38] T. Evgeniou, M. Pontil, Support vector machines: Theory and applications, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 2049 LNAI (2001) 249–257. doi:10.1007/3-540-44673-7_12.

[39] E. Wilson, D.W. Tufts, Multilayer perceptron design algorithm, Neural Networks Signal Process. - Proc. IEEE Work. (1994) 61–68. doi:10.1109/NNSP.1994.366063.

[40] C.E. Metz, Basic principles of ROC analysis, Semin. Nucl. Med. 8 (1978) 283–298. doi:10.1016/S0001-2998(78)80014-2.

[41] A.P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognit. 30 (1997) 1145–1159. doi:10.1016/S0031-3203(96)00142-2.

[42] Riazi, K., Azhari, H., Charette, J. H., Underwood, F. E., King, J. A., Afshar, E. E., Swain, M. G., Congly, S. E., Kaplan, G. G., & Shaheen, A.-A. (2022). The prevalence and incidence of NAFLD WORLDWIDE: A systematic review and meta-analysis. The Lancet Gastroenterology & Hepatology, 7(9), 851–861. https://doi.org/10.1016/s2468-1253(22)00165-0

[43] G. Marchesini, E. Bugianesi, G. Forlani, F. Cerrelli, M. Lenzi, R. Manini, S. Natale, E. Vanni, N. Villanova, N. Melchionda, M. Rizzetto, Nonalcoholic Fatty Liver, Steatohepatitis, and the Metabolic Syndrome, (2003). doi:10.1053/jhep.2003.50161.

[44] Angelo Scuteri, Samer S. Najjar, Christopher H. Morrell, Edward G. Lakatta; The Metabolic Syndrome in Older Individuals: Prevalence and Prediction of Cardiovascular

Events: The Cardiovascular Health Study*. Diabetes Care 1 April 2005; 28 (4): 882–887. https://doi.org/10.2337/diacare.28.4.882

[45] Subhan, F. B., & Chan, C. B. (2016). Review of Dietary Practices of the 21st Century: Facts and Fallacies. Canadian Journal of Diabetes, 40(4), 348–354. doi:10.1016/j.jcjd.2016.05.005

[46] H. Salameh, M. Al Hanayneh, M. Masadeh, M. Naseemuddin, T. Matin, A. Erwin, A.K. Singal, PNPLA3 as a Genetic Determinant of Risk for and Severity of Non-alcoholic Fatty Liver Disease Spectrum, J. Clin. Transl. Hepatol. 4 (2016) 175. doi:10.14218/JCTH.2016.00009.

[47] Kudaravalli, P., & John, S. (2022, May 8). Nonalcoholic fatty liver. National Library of Medicine. Retrieved September 18, 2022, from https://www.ncbi.nlm.nih.gov/books/NBK541033/

Appendix A – Feature selection techniques & results

To find the most important features (i.e., patient biomarkers including demographics, physical and blood measurements) that can accurately predict a patient's NAFLD status, we performed feature selection on the whole patient dataset using two feature selection techniques: (1) Minimum Redundancy Maximum Relevance (MRMR) and (2) Gini importance in the Extra Tree Classifier (ETC). We specifically excluded the diagnosis of diabetes and cardiovascular disease from the feature selection process such that our trained computational models would not depend on such diagnoses and therefore can be used with healthy individuals.

MRMR is an unsupervised feature selection technique. MRMR selects the most relevant features to predict the NAFLD risk based on pairwise correlations, or mutual information of each pair of variables in the dataset, while minimizing the redundancy between variables.[30] The criteria used by MRMR is the Mutual Information Quotient (MIQ) score.

The ETC is a supervised feature selection technique. It fits several randomized decision trees (a.k.a. Extra Trees) on sub-samples of the dataset and averages the results. The feature importance is obtained by computing the normalized total reduction of the criterion brought by that feature, which is known as the Gini Importance.[31] Gini Importance, also known as Mean Decrease in Impurity (MDI), calculates each feature importance as the sum over the number of splits across all decision trees that include the feature, proportionally to the number of samples it splits.[32]

We first applied MRMR and ETC on the patient dataset to identify the top 20 features from each method. The top features from MRMR were ranked by MIQ score, whereas the top features from the ETC were ranked based on their Gini importance. Then, all the important features from the two methods were pooled. To select the top 20 features for use in the

subsequent machine learning, the features were ranked based on how many times they occurred on the list. The higher the number of occurrences of the feature, the greater its importance.

Note: All variables were normalized using the z-score method before computing feature importance values.

Feature Rankings and Importance:

	MRMR Features	**MRMR Score**	**Extra Tree Features**	**Gini Importance**	**Combined Features**
1	Height	0.290	BMI	0.120045	Weight
2	Fasting blood sugar cutoff point 5	5.865	Weight	0.084719	BMI
3	Hemoglobin	5.805	Alanine aminotransferase	0.066559	Alanine aminotransferase
4	Eosinophil count	3.359	Triglycerides	0.052934	HDL cholesterol
5	HDL cholesterol	4.127	HDL cholesterol	0.046787	Fasting blood sugar cutoff point 5
6	Weight	3.045	Diastolic blood pressure	0.035752	Triglycerides
7	Total bilirubin	2.761	Age	0.035185	Diastolic blood pressure

8	Red blood cell count	2.600	Hemoglobin	0.031025	Hemoglobin
9	Systolic blood pressure	2.182	Fasting blood sugar cutoff point 5	0.030329	Systolic blood pressure
10	Leucine aminopeptidase	2.211	Gender	0.030163	Red blood cell count
11	The average hemoglobin concentration	2.184	Systolic blood pressure	0.028173	Height
12	Alkaline phosphatase	1.999	Aspartate aminotransferase	0.027646	Platelet count
13	Alanine aminotransferase	1.994	Red blood cell count	0.026644	Age
14	Platelet count	1.687	White blood cell count	0.024455	Gender
15	Cholinesterase	1.404	Height	0.024314	Aspartate aminotransferase
16	BMI	1.374	Lymphocyte count	0.023973	White blood cell count
17	Diastolic blood pressure	1.259	Platelet count	0.022744	Lymphocyte count

18	Globulin	1.141	Neutrophil count	0.022380	The average hemoglobin concentration
19	Triglycerides	1.157	The average hemoglobin concentration	0.021675	Neutrophil count
20	Albumin	1.072	Cholesterol	0.021091	Eosinophil count

PyMRMR: https://pypi.org/project/pymrmr/

Extra Tree Classifier feature importance: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html#sklearn.ensemble.ExtraTreesClassifier.feature_importances_

Appendix B. Machine learning algorithms & results against internal test sets

We used the six machine learning algorithms to train models to predict a patient's NAFLD status: Gaussian Naïve Bayes (GNB), Logistic Regression (LR), Random Forest, Extreme Gradient Boosting decision trees (XGB), Support Vector Machine (SVM), and Multilayer Perceptron (MLP).

- Gaussian Naïve Bayes (GNB): Applies the Bayes' theorem under the naïve assumption of independence between the features. The Gaussian Naïve Bayes assumes the likelihood of the features to be Gaussian.[33]
- Logistic Regression (LR): A statistical model using a logistic function to model a categorical variable, commonly a binary dependent variable.[34]
- Random Forest (RF): An ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.[35]
- Extreme Gradient Boosting decision trees (XGB): A variation of the gradient boosting technique, designed to increase system performance. Combines decision trees with Stochastic Gradient Boosting with Regularized Gradient Boosting.[36]
- Support Vector Machine (SVM): A clustering technique that applies the statistics of support vectors to categorize data, by deciding sets of hyperplanes that separate different classifications.[37]
- Multilayer Perceptron (MLP): A type of feedforward artificial neural network (ANN) that is composed of multiple layers of nodes with biases and activation thresholds and edges with weights.[38]

Table B1. Model results evaluated on the internal testing set, showing the mean, standard deviation, and 95% confidence interval for our various performance measures across the 50 ensembles, for each machine learning technique trained on the top 20 features for predicting NAFLD risk (The best model overall is highlighted in bold).

	Accuracy	95% CI	AUC ROC	95% CI	Specificity	95% CI	Sensitivity	95% CI
GNB	80.36% ± 0.45%	80.23%-80.49%	80.02% ± 0.48%	79.88%-80.16%	80.65% ± 0.56%	80.49%-80.81%	79.39% ± 0.87%	79.14%-79.64%
LR	83.20% ± 0.34%	83.10%-83.30%	83.59% ± 0.38%	83.48%-83.70%	82.87% ± 0.43%	82.74%-82.99%	84.32% ± 0.74%	84.10%-84.53%
RF	86.36% ± 0.28%	86.28%-86.44%	77.16% ± 0.41%	77.05%-77.28%	94.28% ± 0.29%	94.20%-94.36%	60.05% ± 0.81%	59.82%-60.28%
SVM	82.52% ± 0.33%	82.42%-82.61%	84.06% ± 0.37%	83.96%-84.17%	81.18% ± 0.41%	81.07%-81.30%	86.94% ± 0.68%	86.75%-87.14%
MLP	86.29% ± 0.34%	86.20%-86.39%	79.14% ± 0.89%	78.89%-79.40%	92.46% ± 0.84%	92.22%-92.70%	65.83% ± 2.43%	65.13%-66.53%
XGB	**82.41% ± 0.33%**	**82.32%-82.51%**	**83.89% ± 0.37%**	**83.79%-84.00%**	**81.14% ± 0.43%**	**81.01%-81.26%**	**86.65% ± 0.76%**	**86.43%-86.87%**

Appendix C – Equations of evaluation metrics

- 1. $Sensitivity = \frac{TP}{TP+FN}$
- 2. $Specificity = \frac{TN}{TN+FP}$
- 3. $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
- Where TP = True Positive, TN = True Negative, FP = False Positive and FN = False Negative

  4. $F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} = \frac{TP}{TP+\frac{1}{2}(FP+FN)}$

Appendix D. SHAP values (XGB model)

	Feature Name	Mean SHAP Value	Standard Deviation	95% Confidence Interval
1	BMI	1.22	0.84	1.20-1.24
2	Alanine aminotransferase	0.69	0.44	0.68-0.70
3	Triglycerides	0.48	0.30	0.48-0.49
4	Age	0.33	0.28	0.33-0.34
5	HDL cholesterol	0.21	0.13	0.21-0.21
6	Fasting blood sugar	0.10	0.09	0.10-0.10
7	Systolic BP	0.10	0.07	0.10-0.10
8	Platelet count	0.10	0.08	0.10-0.10
9	Diastolic BP	0.08	0.05	0.08-0.09
10	Weight	0.07	0.05	0.07-0.07
11	Red blood cell count	0.05	0.04	0.04-0.05
12	White blood cell count	0.04	0.03	0.04-0.05

13	Hemoglobin	0.04	0.04	0.04-0.04
14	Lymphocyte count	0.04	0.04	0.04-0.04
15	Aspartate aminotransferase	0.03	0.03	0.03-0.03
16	Gender	0.02	0.01	0.02-0.02
17	Neutrophil count	0.02	0.02	0.02-0.02
18	Average hemoglobin concentration	0.02	0.01	0.01-0.02
19	Height	0.01	0.01	0.01-0.01
20	Eosinophil count	0.00	0.01	0.00-0.00