Journal of the American Heart Association

ORIGINAL RESEARCH

Machine Learning Approach to Classify Cardiovascular Disease in Patients With Nonalcoholic Fatty Liver Disease in the UK Biobank Cohort

Divya Sharma, PhD;* Neta Gotlieb, MD;* Michael E. Farkouh, MD; Keyur Patel, MD; Wei Xu, PhD†
Mamatha Bhat, MD, PhD†

BACKGROUND: Nonalcoholic fatty liver disease (NAFLD) is the most prevalent liver disease worldwide. Cardiovascular disease (CVD) is the leading cause of mortality among patients with NAFLD. The aim of our study was to develop a machine learning algorithm integrating clinical, lifestyle, and genetic risk factors to identify CVD in patients with NAFLD.

METHODS AND RESULTS: We created a cohort of patients with NAFLD from the UK Biobank, diagnosed according to proton density fat fraction from magnetic resonance imaging data sets. A total of 400 patients with NAFLD with subclinical atherosclerosis or clinical CVD, defined by disease codes, constituted cases and 446 NAFLD cases with no CVD constituted controls. We evaluated 7 different supervised machine learning approaches on clinical, lifestyle, and genetic variables for identifying CVD in patients with NAFLD. The most significant clinical and lifestyle variables observed by the predictive modeling were age (59 years [54.00–63.00 years]), hypertension (145 mm Hg [134.0–156.0 mm Hg] and 85 mm Hg [79.00–93.00 mm Hg]), waist circumference (98 cm [95.00–105.00 cm]), and sedentary lifestyle, defined as time spent watching TV >4 h/d. In the genetic data, single-nucleotide polymorphisms in IL16 and ANKLE1 gene were most significant. Our proposed ensemble-based integrative machine learning model achieved an area under the curve of 0.849 using the random forest modeling for CVD prediction.

CONCLUSIONS: We propose a machine learning algorithm that identifies CVD in patients with NAFLD through integration of significant clinical, lifestyle, and genetic risk factors. These patients with NAFLD at higher risk of CVD should be flagged for screening and aggressive treatment of their cardiometabolic risk factors to prevent cardiovascular morbidity and mortality.

Key Words: cardiovascular disease ■ machine learning ■ nonalcoholic fatty liver disease

Nonalcoholic fatty liver disease (NAFLD) has become the most prevalent liver disease worldwide, affecting ≈25% of the population globally. It has become the main cause for liver cirrhosis and hepatocellular carcinoma, and is predicted to soon become the leading indication for liver transplantation, thereby representing a significant economic burden.[1–6]

Cardiovascular disease (CVD) is the most important cause of morbidity and mortality among patients with NAFLD. CVD dictates outcomes in patients with NAFLD to a greater extent than does the progression of liver disease, resulting in ≈40% to 45% of the total deaths in this population.[7,8] Furthermore, a meta-analysis[9] found that patients with NAFLD had a 64%

Correspondence to: Mamatha Bhat, MD, Department of Medicine, University of Toronto, 585 University Ave, Toronto M5G 2N2, Canada. E-mail: mamatha.bhat@uhn.caWei Xu, PhD, Dalla Lana School of Public Health, University of Toronto, 10-511, Princess Margaret Hospital, 610 University Ave, Toronto, ON, M5G 2M9, Canada. E-mail: wei.xu@uhnresearch.ca

Supplemental material for this article is available at https://www.ahajournals.org/doi/suppl/10.1161/JAHA.121.022576

*D. Sharma and N. Gotlieb contributed equally.

†W. Xu and M. Bhat contributed equally.

For Sources of Funding and Disclosures, see page 12.

JAHA is available at: www.ahajournals.org/journal/jaha

CLINICAL PERSPECTIVE

What Is New?

- An integrative machine learning model can identify patients with nonalcoholic fatty liver disease (NAFLD) at high risk for developing subclinical and clinical cardiovascular complications.
- Components of the "metabolic syndrome," sedentary lifestyle, and specific genetic single-nucleotide polymorphisms are among the most significant contributors for cardiovascular disease (CVD) complications in patients with NAFLD.
- Best model performance is when integrating clinical, lifestyle, and genetic data, reflecting the complexity of NAFLD as a risk factor for CVD.

What Are the Clinical Implications?

- A machine learning algorithm can be used in clinical practice to flag those patients with early NAFLD at high risk of CVD.
- CVD screening and treatment of metabolic risk factors as early as possible can potentially reduce the morbidity and mortality associated with CVD as the most common complication of NAFLD.

Nonstandard Abbreviations and Acronyms

CIMT	carotid intima-media thickness
ML	machine learning
MRI-PDFF	magnetic resonance imaging–derived proton density fat fraction
NAFLD	nonalcoholic fatty liver disease
RF	random forest

increased odds ratio for CVD during a median follow-up period of 7 years.[9]

The strong association between NAFLD and CVD is the result of shared metabolic risk factors, such as hypertension, dyslipidemia, and insulin resistance. In addition, NAFLD is an independent risk factor for CVD,[7] suggesting that it should be considered as the hepatic component of the metabolic syndrome.[10–14] Through a bidirectional relationship between NAFLD and metabolic syndrome, NAFLD accelerates the progression of subclinical atherosclerosis and promotes premature CVD events and mortality. Furthermore, NAFLD may directly contribute to atherosclerosis and CVD via hepatic secretion of proinflammatory markers, atherogenic lipoproteins, and procoagulant factors, which results in arterial wall inflammation and secondary plaque vulnerability.[15,16] Consequently, NAFLD is strongly associated with several markers of subclinical atherosclerosis, including carotid intima-media thickening (CIMT), increased coronary artery calcification, impaired flow-mediated vasodilation, and arterial stiffness. Indeed, several large cross-sectional studies have shown that NAFLD is associated with clinical CVD independent of traditional risk factors and metabolic syndrome,[7] making CVD prediction in patients with NAFLD an important research topic.[17–20] In the recent times, researchers have also explored the capability of machine learning (ML) algorithms to improve accuracy of cardiovascular risk prediction.[21–25]

The aim of our study was to develop a novel integrative ML algorithm that could classify CVD in patients with NAFLD using the richly annotated clinical, demographic, and laboratory data from the UK Biobank. Identifying those patients with NAFLD at higher risk of CVD could guide appropriate preventive and therapeutic interventions, thereby preventing the most important reason for morbidity and mortality in patients with NAFLD.

METHODS

Data Availability

Publicly available data from the UK Biobank study were analyzed in this study. The data sets are available to researchers through an open application via https://www.ukbiobank.ac.uk/register-apply/. Code for our integrative ML modeling is available at the link: https://github.com/divya031090/ML_NAFLD_CVD.

Setting

The UK Biobank is a large, prospective study of >500 000 individuals aged 40 to 69 years,[26] recruited between 2006 and 2010. For the UK Biobank, ethical procedures are controlled by a dedicated Ethics and Guidance Council (http://www.ukbiobank.ac.uk/ethics), with institutional review board approval obtained from the North-West Multi-Center Research Ethics Committee. All participants provided written informed consent before enrollment in the UK Biobank. Access to the data was granted for this work under UK Biobank application number 53976.

The study collected extensive phenotypic and genotypic details about its participants, including data from questionnaires, physical measures, accelerometery, multimodal imaging, genome-wide genotyping, and longitudinal follow-up for a wide range of health-related outcomes.[26,27] Detailed cohort protocol, scientific rationale, and study design are available online.[28]

Definition and Diagnosis of NAFLD

The definition of NAFLD requires evidence of hepatic steatosis by either histology or imaging, with exclusion of other causes for liver diseases. Magnetic resonance imaging (MRI) and magnetic resonance spectroscopy are now considered "gold-standard" methods for quantitative hepatic fat measurement.[29] MRI-derived proton density fat fraction (MRI-PDFF) is a method that quantifies hepatic steatosis with a high degree of accuracy and is considered a well-validated diagnostic tool that is not significantly impacted by demographics, histologic activity, or coexisting hepatic conditions.[30,31] To create a cohort of subjects with NAFLD, we selected subjects with a PDFF >5% (MRI-PDFF ≥5), which is the threshold for hepatic steatosis, with high sensitivity and specificity according to previous validated studies.[32,33] From the cohort of patients with MRI-PDFF ≥5, we excluded all subjects who were diagnosed with alcoholic liver disease (defined as alcohol consumption >30 g for men and 20 g daily for women), alcoholic cirrhosis, obstruction/ascending cholangitis/sclerosing cholangitis, α-1 antitrypsin deficiency, Wilson disease, hemochromatosis, primary biliary cholangitis, and viral hepatitis.

CVD Diagnosis

For the diagnosis of CVD, we used parameters indicating both subclinical atherosclerosis and clinical CVD. CIMT is a noninvasive measurement of the arterial wall thickness secondary to atherosclerotic plaques, using ultrasound imaging. CIMT indicates subclinical atherosclerosis and is a validated and well-described predictive marker of major cardiovascular events. Previous studies showed that maximum CIMT >900 mm is associated with increased risk for coronary artery disease.[34–38] We calculated the mean of the maximum CIMT in 4 angles: 120, 150, 210, and 240 degrees. An individual whose mean maximal CIMT was >900 mm was considered to have subclinical CVD. We further included individuals with clinical CVD characterized by ischemic heart disease (history of myocardial infarction and angina pectoris) and/or heart failure. Subjects with prior CVD, defined as self-reported prior myocardial infarction, stroke, and transient ischemic attack, as well as family history of CVD and prior diagnoses identified using *International Classification of Diseases* (*ICD-10*) codes were excluded from the study. Clinical CVD was defined as the subject's first year of hospital admission attributable to CVD after recruitment or death from CVD based on *ICD-10* and I20 to I25 codes identified from linkages to the national death index and Hospital Episode Statistics. CIMT measurements were recorded for the subjects at an imaging visit in the year 2014.

Clinical Data

We analyzed clinical and demographic variables that are known as risk factors for NAFLD and CVD.[39,40] To expand the data available on metabolic risk factors, we also looked at the medication intake of the subjects (presented in Table S1). In addition, we analyzed laboratory parameters, including aspartate aminotransferase, alanine aminotransferase, γ-glutamyl transferase, alkaline phosphatase, total bilirubin, creatinine, hemoglobin, platelet count, urate, and levels of total cholesterol, low-density lipoproteins, triglycerides, glucose, and hemoglobin A1c. The summary of the clinically important variables for cardiovascular outcome used in the analysis is presented in Table 1. Novel risk factors for CVD, such as markers for inflammation (eg, hs-CRP [high-sensitivity C-reactive protein], high-density lipoproteins, and albumin), were also included in the analysis.

Lifestyle Data

Lifestyle variables related to CVD that were analyzed include alcohol consumption, salt intake, and status of cigarette smoking. Individuals who consume alcohol >30 g for men and >20 g daily for women, as well as all patients with any kind of alcoholic disorder or alcoholic liver disease, defined by *ICD* codes, were excluded. We included in the analysis those patients who reported "yes" for alcohol consumption status. Because of a lot of missing data for other diet variables in the subjects with NAFLD, their inclusion in the analysis was not feasible.

We also analyzed variables for physical activity, including time spent watching TV and time spent using a computer, which are markers of sedentary lifestyle and are considered as risk factors for CVD. Moderate physical activity was binarized into 2 categories: subjects who take part in at least 150 minutes weekly of moderate exercise compared with those who do <150 minutes weekly exercise. Time spent watching TV and using computer were categorized into 3 categories, as follows: usage <1, 1 to 4, and >4 h/d.

Genetic Data

From the cohort of 846 subjects with NAFLD, we procured genetic information for chromosome 1-22 from the UK Biobank, for 831 samples and 363 381 single-nucleotide polymorphisms (SNPs) after genetic quality control. On division of the cohort based on 70% training and 30% testing, we obtained 585 samples (322 men, 263 women) in the training set and 246 samples (131 men, 115 women) in the test set. We carried a genome-wide association study using 3 principal components (1, 2, and 3), age, sex, body mass index, and systolic blood pressure as covariates and CVD status as outcome in the training data. Top 100 SNPs with smallest *P* values from genome-wide association study results were selected, and these genetic data were used for our genetic domain-based ML models. SNPs were further identified on the basis of their importance to the CVD prediction.

Table 1. Baseline Characteristics of Variables That Significantly Contribute to CVD

Variables	Cases (N = 400)	Controls (N = 446)	*P* value
Sex			
Women	174 (43.5)	213 (47.7)	0.20
Men	226 (56.5)	233 (52.3)	
Diabetes			
Yes	132 (33)	80 (17.9)	1.83E-06
No	268 (67)	366 (82.1)	
Race			
White	392 (98)	433 (97)	0.978
East Asian	2 (0.5)	4 (0.9)	
Southeast Asian	3 (0.75)	6 (1.3)	
Black	1 (0.25)	0 (0)	
Other*	3 (0.75%)	3 (0.7%)	
Age, y	59.00 (54.00–63.00)	55.00 (48.00–60.00)	1.74E-15
Weight, kg	86.55 (75.90–95.78)	83.20 (73.30–94.00)	7.91E-3
BMI, kg/m^2	29.23 (26.80–31.94)	28.46 (25.99–31.08)	1.55E-3
Diastolic blood pressure, mm Hg	85.00 (79.00–93.00)	82.00 (76.00–89.00)	6.92E-7
Systolic blood pressure, mm Hg	145.0 (134.0–156.0)	135.00 (125.00–146.5)	4.22E-16
Waist circumference, cm	98.00 (91.00–105.00)	95.00 (88.00–101.00)	1.37E-5
Low-density lipoprotein, mmol/L	3.53 (2.95–4.133)	3.74 (3.18–4.34)	1.63E-4
Glucose, mmol/L	5.04 (4.63–5.61)	4.97 (4.62–5.36)	0.04
Alanine aminotransferase, U/L	27.24 (20.18–36.80)	25.60 (19.09–34.06)	0.04
Aspartate aminotransferase, U/L	26.60 (22.60–31.73)	25.60 (22.12–30.70)	0.05
Alkaline phosphatase, U/L	82.10 (68.42–97.47)	78.90 (67.70–94.67)	0.04
γ-Glutamyl transferase, U/L	34.10 (23.82–50.08)	33.20 (22.60–51.95)	0.2
Bilirubin, μmol/L	8.29 (6.65–10.37)	8.31 (6.67–10.52)	0.90
Albumin, g/L	45.52 (43.84–47.24)	45.41 (43.70–47.15)	0.40
White blood cell count, 10^9 cells/L	6.89 (6.00–7.98)	6.70 (5.70–7.77)	0.02
High-density lipoprotein, mmol/L	1.25 (1.08–1.41)	1.24 (1.06–1.47)	0.8
Triglycerides, mmol/L	1.97 (1.39–2.80)	1.94 (1.40–2.71)	0.99
Creatinine, μmol/L	73.95 (64.20–83.55)	72.70 (62.98–82.30)	0.1
Moderate physical activity			
Yes	137 (34.2)	147 (32.95)	0.81
No	261 (65.2)	294 (65.91)	
Alcohol consumption status			
Yes	373 (93.2)	405 (90.6)	0.48
No	26 (6.5)	40 (9.1)	
Smoking status			
Yes	190 (47.5)	174 (39)	0.02
No	207 (51.75)	265 (59.41)	

The summary of the categorical variables is represented using frequency and percentage of each category of the variables, whereas continuous variables are summarized using median values and interquartile range. BMI indicates body mass index; and CVD, cardiovascular disease..

*The Race, subjects categorized as "Other" are those whose race was not categorized as either White, Mixed, Asian or Black as per the UK Biobank documentation.

Proposed Framework

The novel integrative framework is described in Figure 1, and each step in the learning process is detailed in the flowchart provided in Figure 2. The proposed framework consists of 2 levels of assessment: (1) ML model assessment for each individual domain and (2) integration/ensemble of the best model from each domain into a naive Bayes classifier for final prediction of CVD outcome.

Statistical Analysis

We considered 7 algorithms covering different classes of ML modeling approaches for the first level of assessment: support vector machines,[41] random forest (RF),[42] neural networks,[43] logistic regression,[44] Lasso regression,[45] ridge regression,[46] and naive Bayes classification.[47] To tune the ML models and select the models with highest accuracy, hyperparameters were determined via grid search.

We trained our networks on an NVIDIA Tesla P100 GPU with 16GB of RAM in R version 3.5.3. In RF training, a maximum of 500 trees and 3 node-wise predictors sampled for splitting were set. The support vector machine was trained with a linear kernel and regularization term of 10. The Lasso and ridge regression models were trained using iterative fitting of L1 penalty and λ. In the neural network model, tuning of learning rate was ensured to achieve lowest loss and highest accuracy. Missing data imputation through chained equations, followed by standardization and normalization of variables, was done. A total of 70% of the subjects were part of the training set, and the remaining 30% were part of the test set. The 10-times, 10-fold cross-validation was performed on the training set to tune parameters. The performance was evaluated through a mean area under the curve (AUC),

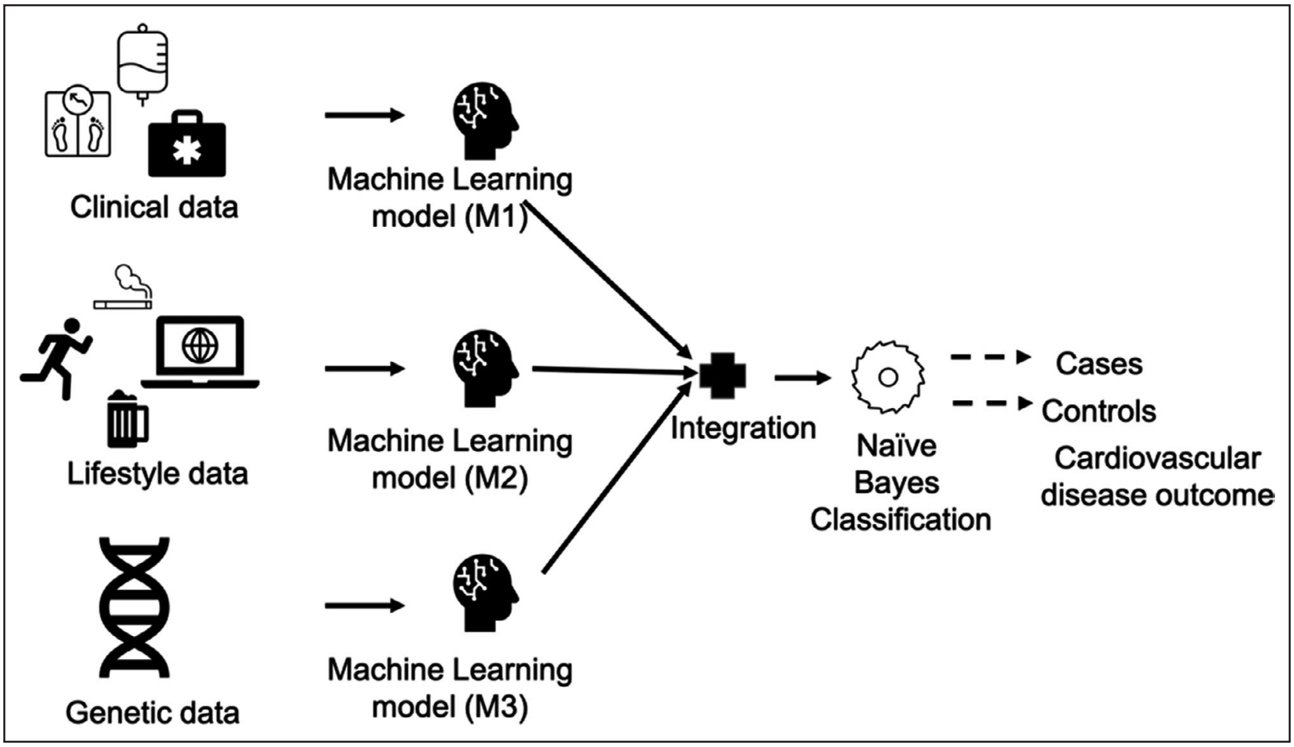

Clinical data
Machine Learning model (M1)
Lifestyle data
Machine Learning model (M2)
Genetic data
Machine Learning model (M3)
Integration
Naïve Bayes Classification
Cases
Controls
Cardiovascular disease outcome

Figure 1. Multimodal integrative framework for cardiovascular disease prediction from clinical, genetic, and lifestyle data domains among subjects with nonalcoholic fatty liver disease.

calculated through receiver operating characteristics (ROC) curve. Bootstrapping was performed on the test set to calculate 95% CIs of the AUC values.

RESULTS

Characteristics of the Study Population

The study flowchart is shown in Figure 3. PDFF and patient meta-data were obtained through UK Biobank access application number 53976. PDFF was successfully calculated from 4617 MRI samples, among whom 1011 individuals had an MRI-PDFF ≥5. After the exclusion of other liver diseases and alcohol consumption above the threshold, as mentioned before, a total of 846 were considered to have NAFLD. Subjects were further classified according to the presence of CVD. Cases were composed of patients with NAFLD and CVD, and controls were composed of patients with NAFLD and without CVD. A total of 400 cases were diagnosed with CVD compared with 446 controls, defined as patients with NAFLD with no CVD. A total of 194 cases had subclinical CVD detected through CIMT, and 285 cases had CVD detected through disease codes specified in the UK Biobank, with 79 subjects common between both. Patient characteristics and significant variables are presented in Table 1. The full distribution of CVD among NAFLD cases is presented in Table S2. The complete summary of variables is presented in Table S3.

Comparison of Predictive Performance for CVD

To test the robustness and generalizability of the ML models, 10 times, 10-fold cross-validation analysis was performed on the clinical variables by partitioning the training set into 1-fold of test set and 9-folds of training sets to evaluate the model, as illustrated in Figure S1. For our test set of the cohort, the ROC curves obtained are presented in Figure 4, wherein, the orange plot line depicts the ROC curve for RF with an AUC of 0.799, followed in performance by Lasso regression (AUC = 0.753). DeLong test using the pROC package in R,[48] to compare significance of AUC difference between best-performing RF model and rest 6 comparative approaches, gave significant *P* values of 0.05 in comparison with Lasso, 0.02 in comparison with ridge, 0.02 in comparison with support vector machine, 0.02 in comparison with naive Bayes, 0.01 in comparison with logistic regression, and 0.005 in comparison with neural network model.

In the genetic data, we conducted ML modeling using the 7 ML models and observed AUC values

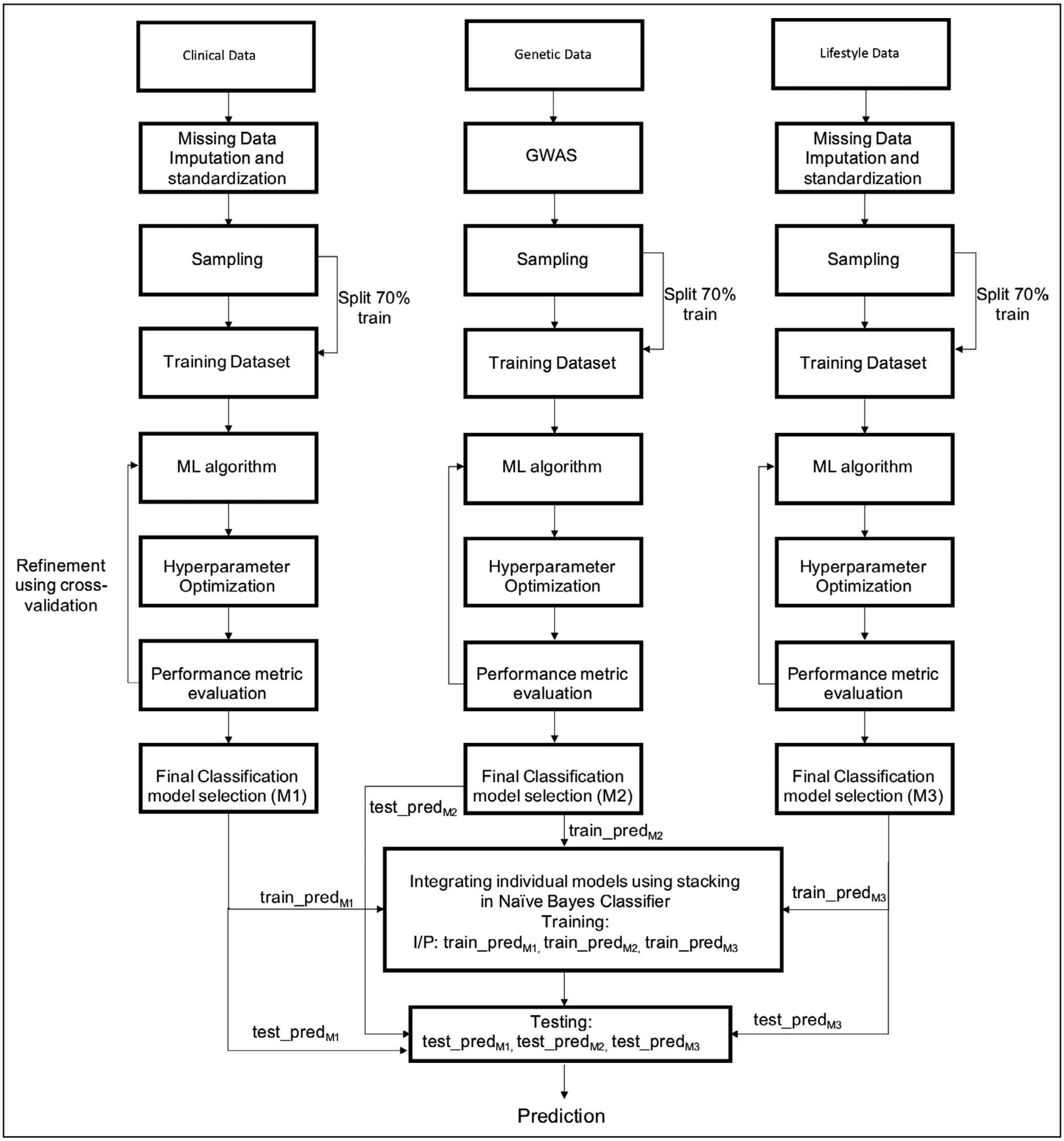

Clinical Data
Genetic Data
Lifestyle Data
Missing Data Imputation and standardization
GWAS
Missing Data Imputation and standardization
Sampling
Sampling
Sampling
Split 70% train
Split 70% train
Split 70% train
Training Dataset
Training Dataset
Training Dataset
ML algorithm
ML algorithm
ML algorithm
Refinement using cross-validation
Hyperparameter Optimization
Hyperparameter Optimization
Hyperparameter Optimization
Performance metric evaluation
Performance metric evaluation
Performance metric evaluation
Final Classification model selection (M1)
Final Classification model selection (M2)
Final Classification model selection (M3)
test_pred$_{M2}$
train_pred$_{M2}$
train_pred$_{M1}$
Integrating individual models using stacking in Naïve Bayes Classifier
Training:
I/P: train_pred$_{M1}$, train_pred$_{M2}$, train_pred$_{M3}$
train_pred$_{M3}$
test_pred$_{M1}$
Testing:
test_pred$_{M1}$, test_pred$_{M2}$, test_pred$_{M3}$
test_pred$_{M3}$
Prediction

Figure 2. Flowchart describing the details of each step of the integrative machine learning (ML) modeling.
GWAS indicates genome-wide association study.

for RF model to be higher than the other comparative methods (refer Table 2, column 5). A total of 26 SNPs were identified in gene IL16 from chromosome 15, and 6 SNPs were identified in gene ANKLE1 from chromosome 19 to be important to the CVD outcome. Furthermore, in IL16 gene, SNP *rs4531696* with a *P* value of 0.012, and in ANKLE1 gene, SNP *rs891017* with a *P* value of 0.009 were selected and adjusted for in the clinical data. Integrating these covariates in the clinical data improved the performance of prediction, increasing AUC from 0.799 to 0.820 on the test set, as shown in the orange plot line in Figure 5. The DeLong test, comparing AUC difference between the model with only clinical data versus the model with both clinical and genetic data, gave a *P* value of 0.03. As tabulated in Table 2, in the lifestyle variables, RF method performed the best with an AUC of 0.652, followed by ridge (AUC = 0.633), Lasso (AUC = 0.632), support

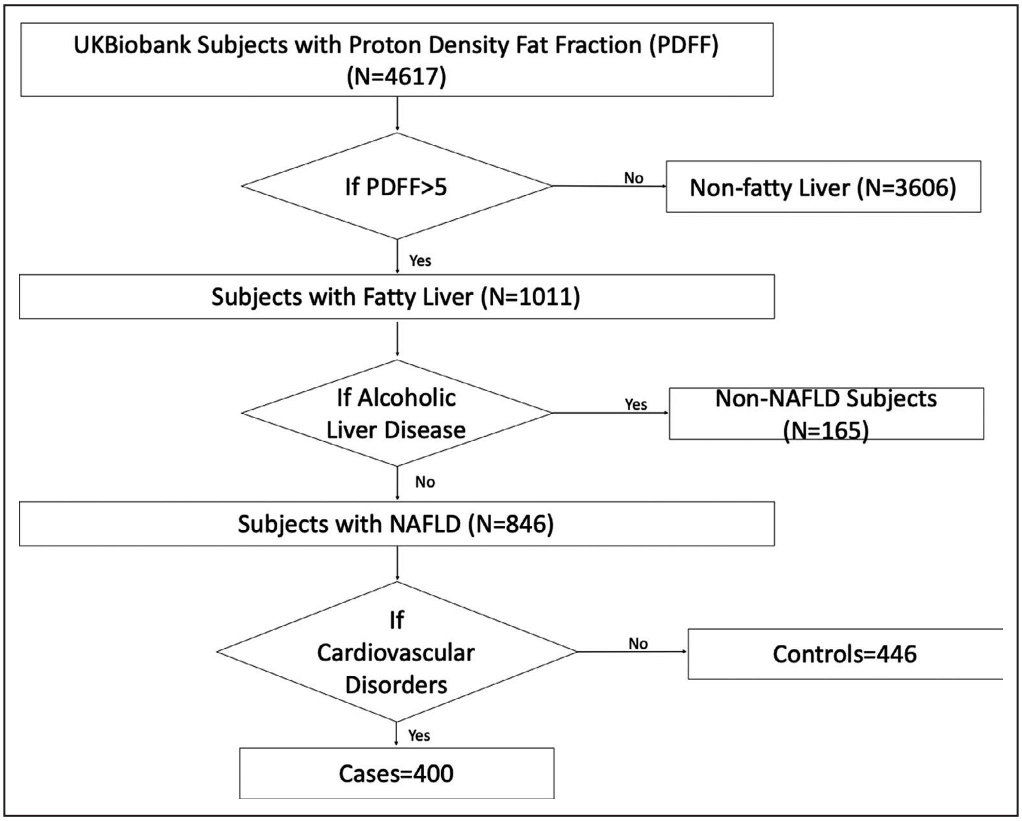

UKBiobank Subjects with Proton Density Fat Fraction (PDFF) (N=4617)
If PDFF>5
No
Non-fatty Liver (N=3606)
Yes
Subjects with Fatty Liver (N=1011)
If Alcoholic Liver Disease
Yes
Non-NAFLD Subjects (N=165)
No
Subjects with NAFLD (N=846)
If Cardiovascular Disorders
No
Controls=446
Yes
Cases=400

Figure 3. Flowchart illustrating stepwise study design to categorize subjects with nonalcoholic fatty liver disease (NAFLD) who develop cardiovascular disease.
PDFF indicates proton density fat fraction.

vector machine (AUC = 0.612), logistic regression (AUC = 0.610), naive Bayes (AUC = 0.591), and neural network (AUC = 0.585). However, herein, we observed a dip in AUC values, attributed to the smaller number of variables feasible to include in the lifestyle data. Table 2 illustrates mean AUC and 95% CIs across all individual domains for the 7 ML approaches.

In the final integrative modeling, we first selected best performing models in individual domains (clinical, lifestyle, and genetic). Furthermore, we experimented with a few common ensemble methods, such as bagging (bootstrap aggregating), RF, AdaBoost, and naïve Bayes classifier for integrating the domains[49–51]; and as illustrated in Table S4, the performance of naïve Bayes was better than other ensemble methods. Therefore, we determined the performance of the final predictive modeling that combines the 3 domains through integration using the naive Bayes ensemble. The ROC plot for the comparison of performance of the integrative modeling is illustrated in Figure 5. The black plot line with an AUC of 0.849 shows the performance edge that the integrative modeling has compared with the prediction through individual domains. We observed that the classification improved considerably from AUCs of 0.799 (95% CI, 0.779–0.817), 0.652 (95% CI, 0.639–0.665), and 0.617 (95% CI, 0.599–0.637) using clinical, lifestyle, and genetic data domains individually, respectively, to 0.849 (95% CI, 0.840–0.855) using the integrated model. The sensitivity and the specificity of the integrated model on the test data set were 71.4% and 84.2%, respectively, using the Youden index,[52] with a positive and negative predictive value of 80.3% and 76.7%, respectively, showing our model's efficiency during classification. The DeLong test for comparison of AUC differences along with Bonferroni correction[53] between the ML model on the clinical domain versus the ML model on the clinical and genetic domains gave a *P* value of 0.03, and a significant *P* value of 0.009, for the integrative modeling on clinical, genetic, and lifestyle data domains compared with AUC obtained using only the clinical domain. We also did a subgroup analysis for both clinical and subclinical CVD (determined by CIMT) and observed higher AUCs (≈11% increase) in the clinical CVD group compared with the subgroup determined by CIMT threshold, as tabulated in Table S5.

Variable Importance

Figure 6 illustrates an importance plot for the clinical data variables ranked according to their contribution to the predictions through the RF model. Age followed

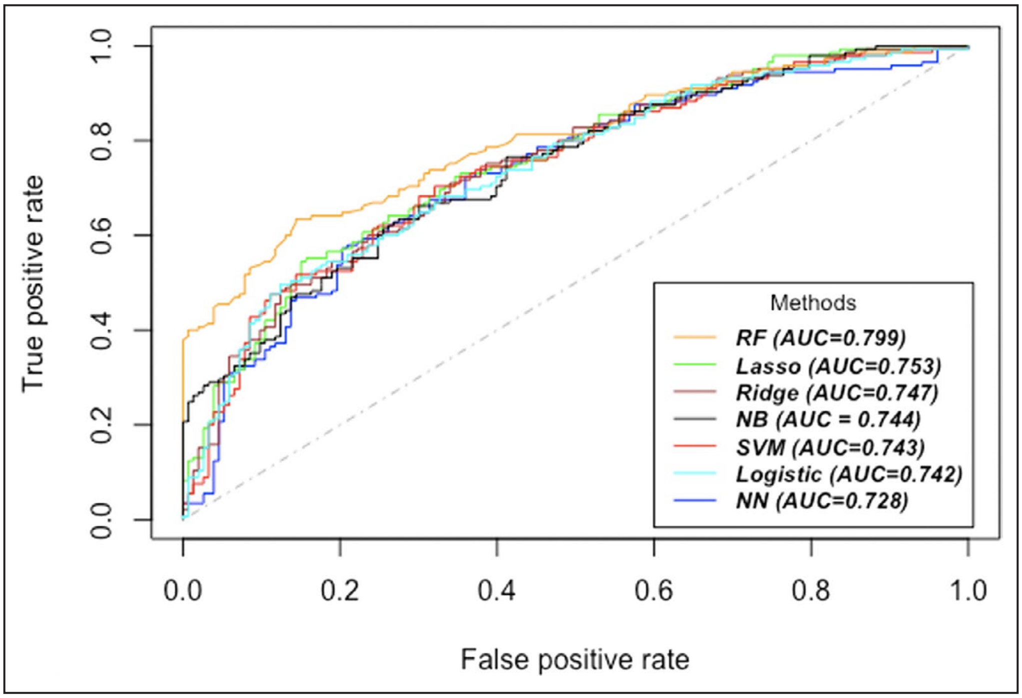


Figure 4. Receiver operating characteristics curve obtained on the test set of the cardiovascular disease cohort using the clinical variables.

The test set was composed of 153 controls and 145 cases. The gray dotted line corresponds to area under the curve (AUC) equal to 0.5, indicating a random classification model. NB indicates naïve Bayes; NN, neural network; RF, random forest; and SVM, support vector machine.

Table 2. Performance of the 7 Models in Each Individual Domain for Both the Training and Test Data Sets

	Clinical		Genetic		Lifestyle	
Methods	Training	Test	Training	Test	Training	Test
Random forest	0.810 (0.788–0.828)	0.799 (0.779–0.817)	0.624 (0.605–0.643)	0.617 (0.599–0.637)	0.673 (0.655–0.687)	0.652 (0.639–0.665)
Lasso	0.760 (0.742–0.779)	0.747 (0.727–0.765)	0.610 (0.591–0.625)	0.602 (0.585–0.620)	0.649 (0.631–0.665)	0.632 (0.619–0.645)
Ridge	0.762 (0.749–0.781)	0.753 (0.731–0.777)	0.611 (0.593–0.629)	0.605 (0.581–0.628)	0.645 (0.630–0.657)	0.633 (0.617–0.643)
Naïve Bayes	0.764 (0.748–0.788)	0.744 (0.729–0.761)	0.573 (0.552–0.591)	0.564 (0.549–0.581)	0.603 (0.586–0.618)	0.591 (0.578–0.604)
SVM	0.759 (0.747–0.776)	0.743 (0.731–0.759)	0.611 (0.593–0.634)	0.603 (0.585–0.620)	0.620 (0.608–0.633)	0.612 (0.599–0.625)
Logistic regression	0.743 (0.724–0.759)	0.740 (0.722–0.757)	0.579 (0.550–0.595)	0.571 (0.550–0.588)	0.619 (0.602–0.635)	0.610 (0.597–0.623)
Neural network	0.734 (0.718–0.751)	0.728 (0.708–0.745)	0.600 (0.582–0.619)	0.592 (0.575–0.610)	0.624 (0.600–0.639)	0.612 (0.595–0.628)

Data are given as area under the curve (95% CI). The top row shows that random forest method performed the best in each of the domains for predicting cardiovascular outcome in subjects with nonalcoholic fatty liver disease. SVM indicates support vector machine.

by systolic blood pressure, diastolic blood pressure, and waist circumference were the most important variables. Red blood cell size distribution and diabetes were also significant to CVD prediction, however, to a lesser extent. We also took into account the influence of medication on the CVD prediction by categorizing subjects into 3 groups, as per their medication consumption: cholesterol-lowering medication versus blood pressure medication versus others. As tabulated in Table S6, the prediction performance in terms of AUC was comparable and consistent with and without inclusion of medication information to our analysis.

The important variables, as observed through the RF modeling, were in concordance with the univariate

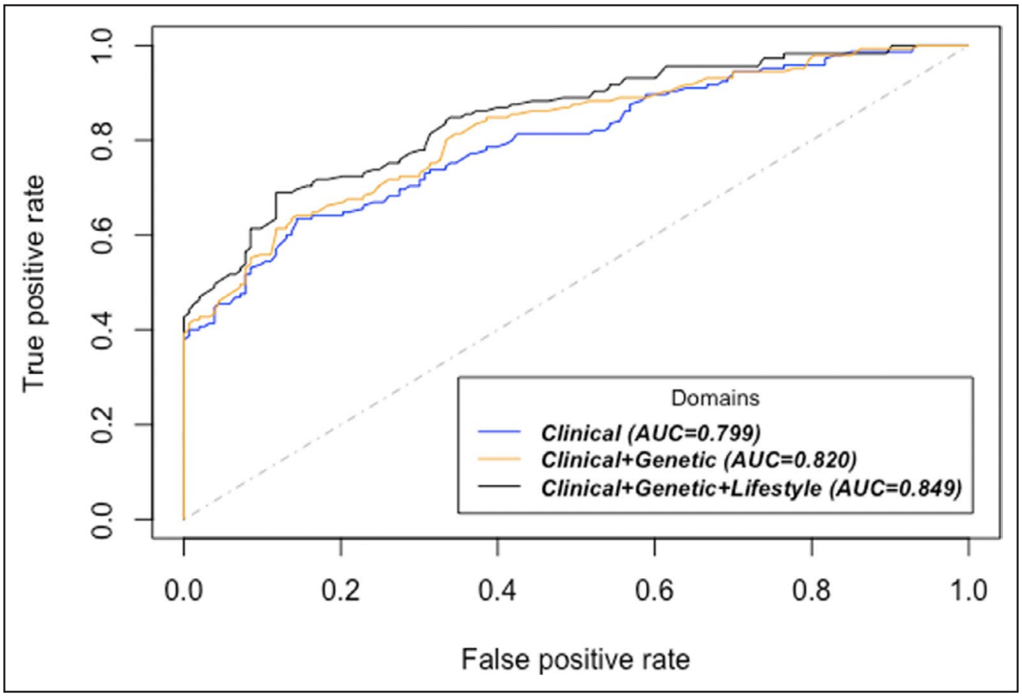


Figure 5. Receiver operating characteristics curves comparing the performance enhancement observed by integrating domains relevant to the cardiovascular disease outcome.
AUC indicates area under the curve.

analysis, as tabulated in Table S7, showing that the RF model can capture the accurate essential clinical data variables in CVD prediction. A closer look at the RF tree gave a set of tree-based rules and thresholds used to classify subjects into risk of CVD versus no risk. An example tree illustrating such rules is shown in Figure 7. The full tree obtained can be traversed by a computational tool to evaluate the risk of CVD in the subjects based on various clinical/lifestyle parameters and aid the clinician to screen subjects with high risk of CVD.

Similarly, to assess the important variables in the lifestyle data, we plotted a similar variable importance plot, presented in Figure S2. Time spent watching TV was the most significant variable, followed by salt intake and smoking status.

DISCUSSION

We have established an integrated ML model that accurately identifies individuals with CVD in the setting of NAFLD using the UK Biobank database. Our model integrated clinical, lifestyle, and genetic parameters of patients with NAFLD to identify those with CVD, the most common and fatal complication of NAFLD, with a high AUC of 0.849 (71.4% sensitivity and 84.2% specificity). This reflects the fact that NAFLD is a complex entity integrating environmental and genetic factors that influence each other in a reciprocal manner. This algorithm could be used in practice to flag those patients with early NAFLD at high risk of CVD. As such, our model delineates those patients with early NAFLD along with age >59 years, hypertension, and high waist circumference with a sedentary lifestyle and specific SNPs as being high risk for CVD. Therefore, our model goes beyond the current literature, by identifying patients with early NAFLD at risk for CVD.

Metabolic-associated fatty liver disease was recently suggested to better define fatty liver disease and metabolic dysfunction, rather than using the term “nonalcoholic.” Metabolic-associated fatty liver disease reflects a heterogeneous phenotype that is influenced by multiple factors, including age, sex, hormonal status, ethnicity, diet, alcohol intake, smoking, genetic predisposition, the microbiota, and metabolic status, which all interact with each other in a reciprocal manner and reflect the fact that modifying these factors may ultimate influence the disease course and future complications.[54] Similarly, in our study, we showed that when integrating clinical and genetic data, outcome prediction (in this case, CVD) is better than analyzing each risk factor separately.

A median age >59 years (95% CI, 54.00–63.00 years) was the strongest predictor for CVD among the clinical data parameters in our model, indicating its importance as a strong risk factor for CVD in the population with NAFLD. Age is a major and well-established risk factor for CVD,[3,55,56] exposing an individual to

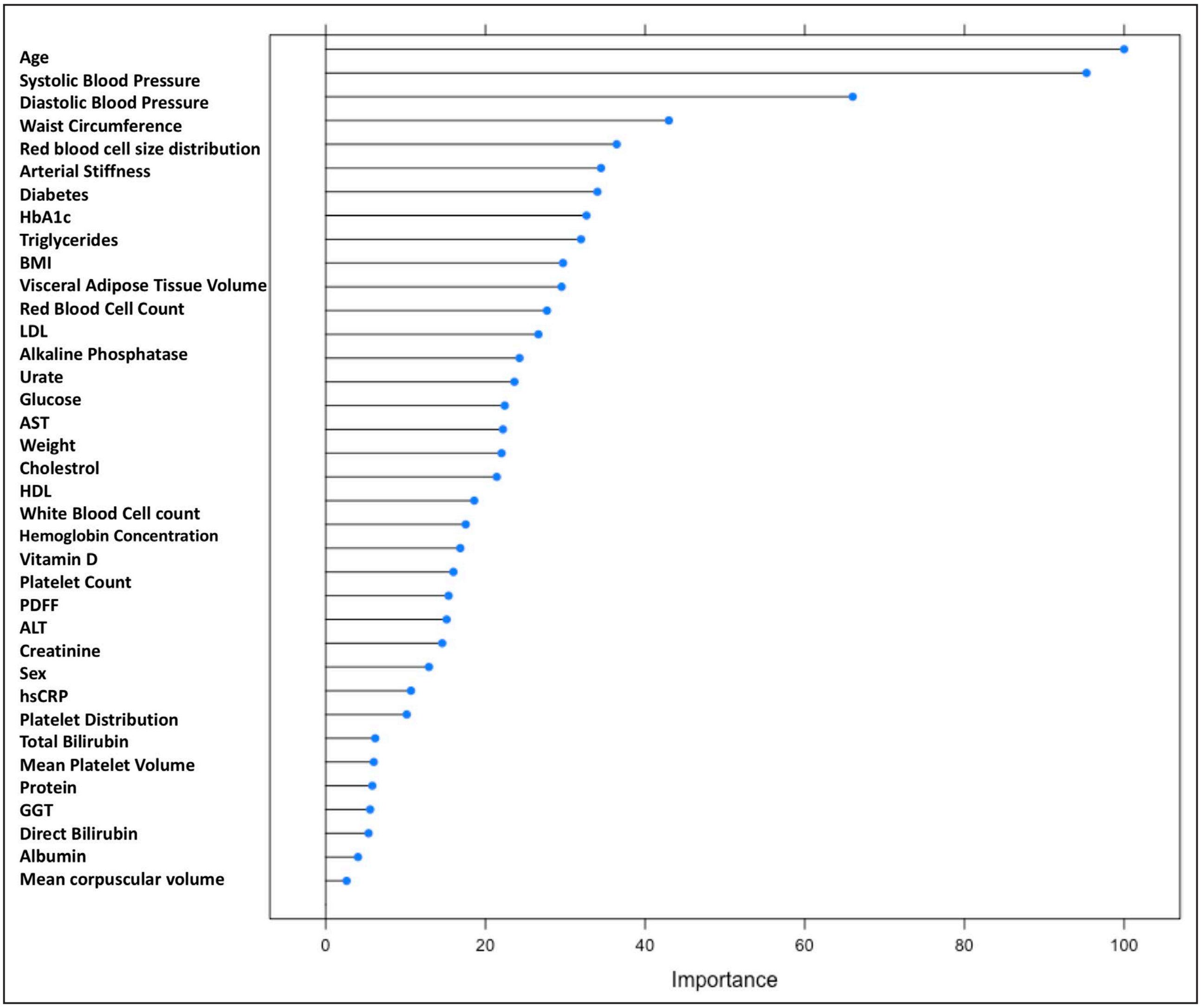


Figure 6. Variable importance plot, demonstrating the importance of clinical variables obtained through the machine learning modeling on the clinical data.

ALT indicates alanine aminotransferase; AST, aspartate aminotransferase; BMI, body mass index; GGT, γ-glutamyl transferase; HbA1c, hemoglobin A1c; HDL, high-density lipoprotein; hs-CRP, high-sensitivity C-reactive protein; LDL, low-density lipoprotein; and PDFF, proton density fat fraction.

metabolic and environmental risk factors for a longer duration. Hypertension, with a systolic blood pressure ≥145 mm Hg (95% CI, 134.0–156.0 mm Hg) and a diastolic blood pressure ≥85 mm Hg (95% CI, 79.00–93.00 mm Hg), and waist circumference >98 cm (95% CI, 91.00–105.00 cm) were the next strongest variables in our model. Hypertension has been established as the strongest risk factor for CVD. Furthermore, there is a gradual increase in coronary artery calcium, and the risk for CVD progression increases alongside increases in systolic blood pressure values.[57,58] Waist circumference, which represents abdominal adiposity, is strongly associated with cardiovascular mortality to a much larger extent than body mass index alone. Diabetes and triglycerides were significant contributors to the model, however, to a lesser extent. In our total cohort of individuals diagnosed with NAFLD, 24% were diagnosed with type 2 diabetes and 30% of those with CVD had diabetes. Markers of inflammation, such as hs-CRP, high-density lipoprotein, albumin, arterial stiffness, and visceral adipose tissue volume, had only a modest contribution to the model. Most of the patients in our cohort had normal levels of both alanine aminotransferase and aspartate aminotransferase, where alanine aminotransferase had a median of 26.22 and an interquartile range of 19.70 to 35.41 U/L and aspartate aminotransferase had a median of 26.00 and an interquartile range of 22.32 to 31.20 U/L, suggesting,

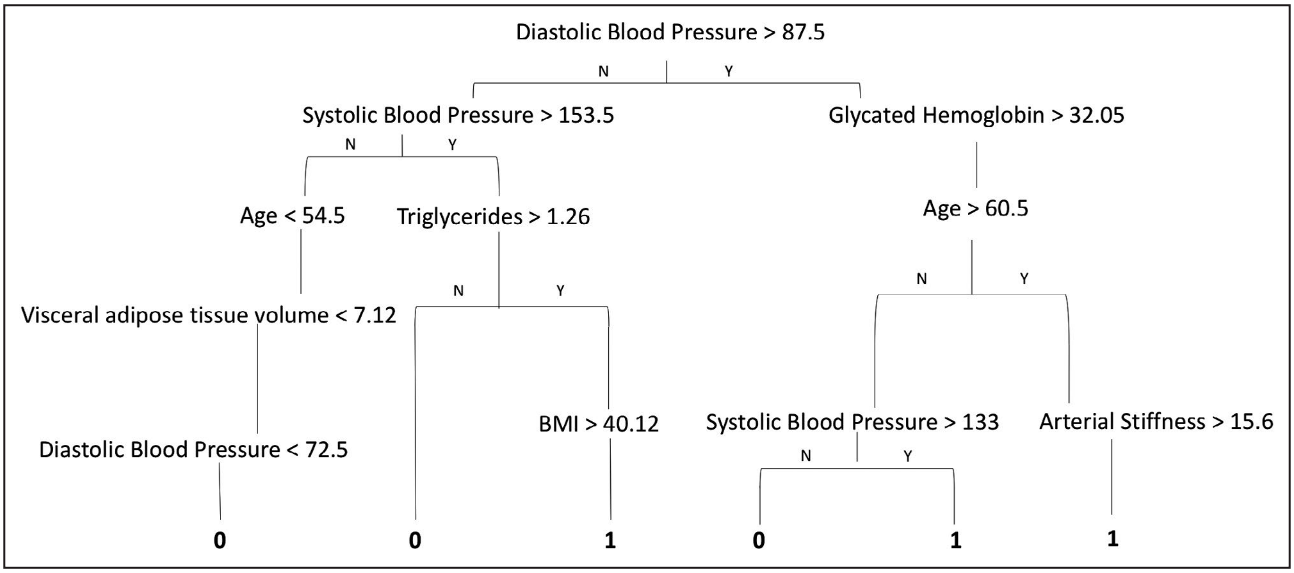

Diastolic Blood Pressure > 87.5
N
Y
Systolic Blood Pressure > 153.5
N
Y
Glycated Hemoglobin > 32.05
Age < 54.5
Triglycerides > 1.26
Age > 60.5
N
Y
N
Y
Visceral adipose tissue volume < 7.12
BMI > 40.12
Systolic Blood Pressure > 133
N
Y
Arterial Stiffness > 15.6
Diastolic Blood Pressure < 72.5
0
0
1
0
1
1

Figure 7. Example tree illustrating some set of rules and thresholds from the dense random forest tree used for classification in the analysis.

The "0" and "1" on the leaf node represent no risk of cardiovascular disease (CVD) and risk of CVD in the subject, respectively. BMI indicates body mass index; N, no; and Y, yes.

however not proving, that our cohort mostly experienced simple steatosis rather than nonalcoholic steatohepatitis.

From the lifestyle parameters, time spent in front of the TV was associated with the highest risk for developing CVD among patients with NAFLD, with an AUC of 0.65 in the ML model, followed by salt intake, smoking, and physical activity. We considered time spent in front of the TV as a marker of sedentary lifestyle, recently recognized as an independent cardiovascular risk factor,[59,60] consistent with previous studies showing a clear association between sedentary lifestyle and NAFLD mainly secondary to obesity.[61–63]

Several studies have aimed to link certain genes to CVD in the population with NAFLD. For instance, *PNPLA3* and *TM6SF2* might decrease the risk and possibly protect from CVD,[64] whereas variants in *GCKR* may be associated with increased CVD risk.[65] However, despite robust investigations, studies have failed to allocate specific genetic components that link NAFLD to CVD in terms of causality.[66,67] We observed 2 SNPs in IL16 and ANKLE1 genes as highly associated with CVD. Studies have shown a significant correlation between increased levels of IL16, body mass index, and waist circumference. Furthermore, IL16 mRNA is reflective of the inflammatory process in individuals with overweight/obesity, which is strongly associated with NAFLD and CVD.[68–70] ANKLE1 SNPs are expressed in hematopoietic tissues in human and have previously been associated with genomic instability in colorectal and breast cancer, but not CVD to date.[71]

Some of the limitations of our study include the inability to bucket our risk prediction in a 5- or 10-year risk timeline because of variability in the duration from baseline to the time of clinical or subclinical CVD in the cohort. However, with ML modeling, we could still allocate those patients with NAFLD at risk of CVD, with high AUC of 0.849. In the future, we will explore studies with data focused on longer follow-up, to provide a better insight into predicting CVD in specific time frames.

Also, the data in our study were taken at a single point of time, and longitudinal assessment for patterns could not be performed; however, the data were representative enough to offer sound observations. In the future, we would explore cohorts with follow-up data to establish a longitudinal prediction model for CVD. Another limitation of our data was the small number of cases diagnosed with NAFLD compared with the large number of participants in the UK Biobank as the diagnosis was based on MRI-PDFF rather than on abnormal biochemistry or ultrasound findings. MRI-PDFF is the gold standard for quantifying hepatic fat; however, the number of patients whose MRI-PDFF data were available for analysis was rather small compared with the total number of participants in the UK Biobank. However, we observed that the number of subjects in the study was sufficient for our ML analysis to yield good performance (AUC = 0.849) while classifying subjects with risk of CVD and justify the purpose of integrative ML modeling for CVD prediction. Also, because of the fact that most of the patients had normal enzymes, Fibrosis-4 (FIB-4) and NAFLD fibrosis scores, which were validated to assess the degree of fibrosis, were not calculated; and as a result, we assume that the degree of fibrosis was low in most of the cohort.

NAFLD-Related CVD Risk Stratification

There are currently no guidelines for CVD screening in patients with NAFLD. Global cardiovascular risk assessment scores available for the general population, including the Framingham risk score, atherosclerotic CVD, and others, use multiple traditional cardiovascular risk factors for risk assessment in all asymptomatic adults without a clinical history of CVD. NAFLD, however, is not included in these scores.[72,73] Other cardiovascular risk scores have been suggested in NAFLD,[74] such as the coronary artery calcium scores, Leaman scores,[75] and one based on age, mean platelet volume, and diabetes. None of these scores is yet validated, and it is uncertain to which patients with NAFLD they should be applied. As CVD risk increases in concordance with NAFLD severity, it is reasonable to screen high-risk groups with obesity and diabetes.[76–79] As shown in our study, it may be worth screening those individuals with early NAFLD who are present with certain clinical and genetic risk factors and could potentially benefit from early interventions that would prevent cardiovascular complications.

The clinical and lifestyle variables that were included in the model can be easily and routinely collected during clinic visits. Moreover, the most significant variables identified in the model are those that can be retrieved by the general practitioner in the community setting and hence flag patients at risk as early as possible. On the other hand, genetic testing is not performed in patients with NAFLD as part of routine clinical care. It is still relatively new, and its utility would need to be clearly demonstrated for implementation, particularly in a public health care framework. Our study demonstrates that genetic data are additive to clinical and lifestyle data in predicting CVD among individuals with NAFLD.

In conclusion, our ML model integrates important clinical, lifestyle, and genetic risk factors to efficiently identify CVD in the population with early NAFLD, thereby flagging those patients who will derive the greatest benefit from CVD screening and treatment of metabolic risk factors. This has the potential to help reduce the morbidity and mortality associated with CVD as the most common complications of NAFLD.

ARTICLE INFORMATION

Received May 22, 2021; accepted November 12, 2021.

Affiliations

Department of Biostatistics, Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada (D.S., W.X.); Division of Adult Gastroenterology, University Health Network, Toronto General Hospital, Toronto, Ontario, Canada (N.G.); Peter Munk Cardiac Centre, Heart and Stroke Richard Lewar Centre, University of Toronto, Ontario, Canada (M.E.F.); Division of Gastroenterology, University Health Network, Toronto General Hospital, Toronto, Ontario, Canada (K.P.); Biostatistics Division, Dalla Lana School of Public Health, University of Toronto, Ontario, Canada (W.X.); and Department of Medicine, Multi-Organ Transplant Program, Toronto General Hospital, Toronto, Ontario, Canada (M.B.).

Acknowledgments

We acknowledge the assistance of Mei Dong for the genome-wide association study analysis. Guarantor of the article: Dr Bhat. All declared authors read and approved the final manuscript submitted for publication.

Sources of Funding

This work was funded by Toronto General and Western Hospital Foundation. None of the authors of this article has a financial or personal relationship with other people or organizations that could inappropriately influence the content of the article.

Disclosures

Dr Farkouh received research grant support from Amgen, Novartis, and Novo Nordisk. The remaining authors have no disclosures to report.

Supplemental Material

Tables S1–S7
Figure S1–S2

REFERENCES

1. Vernon G, Baranova A, Younossi ZM. Systematic review: the epidemiology and natural history of non-alcoholic fatty liver disease and non-alcoholic steatohepatitis in adults. *Aliment Pharmacol Ther*. 2011;34:274–285. doi: 10.1111/j.1365-2036.2011.04724.x
2. Younossi ZM, Marchesini G, Pinto-Cortez H, Petta S. Epidemiology of nonalcoholic fatty liver disease and nonalcoholic steatohepatitis: implications for liver transplantation. *Transplantation*. 2019;103:22–27.
3. Younossi ZM, Koenig AB, Abdelatif D, Fazel Y, Henry L, Wymer M. Global epidemiology of nonalcoholic fatty liver disease—meta-analytic assessment of prevalence, incidence, and outcomes. *Hepatology*. 2016;64:73–84.
4. Estes C, Razavi H, Loomba R, Younossi Z, Sanyal AJ. Modeling the epidemic of nonalcoholic fatty liver disease demonstrates an exponential increase in burden of disease. *Hepatology*. 2018;67:123–133. doi: 10.1002/hep.29466
5. Mikolasevic I, Filipec-Kanizaj T, Mijic M, Jakopcic I, Milic S, Hrstic I, Sobocan N, Stimac D, Burra P. Nonalcoholic fatty liver disease and liver transplantation-where do we stand? *World J Gastroenterol*. 2018;24:1491. doi: 10.3748/wjg.v24.i14.1491
6. Noureddin M, Vipani A, Bresee C, Todo T, Kim IK, Alkhouri N, Setiawan VW, Tran T, Ayoub WS, Lu SC, et al. NASH leading cause of liver transplant in women: updated analysis of indications for liver transplant and ethnic and gender variances. *Am J Gastroenterol*. 2018;113:1649–1659.
7. Byrne CD, Targher G. NAFLD: a multisystem disease. *J Hepatol*. 2015;62:S47–S64. doi: 10.1016/j.jhep.2014.12.012
8. Anstee QM, Mantovani A, Tilg H, Targher G. Risk of cardiomyopathy and cardiac arrhythmias in patients with nonalcoholic fatty liver disease. *Nat Rev Gastroenterol Hepatol*. 2018;15:425–439.
9. Targher G, Byrne CD, Lonardo A, Zoppini G, Barbui C. Non-alcoholic fatty liver disease and risk of incident cardiovascular disease: a meta-analysis. *J Hepatol*. 2016;65:589–600.
10. Assy N, Djibre A, Farah R, Grosovski M, Marmor A. Presence of coronary plaques in patients with nonalcoholic fatty liver disease. *Radiology*. 2010;254:393–400.
11. Mirbagheri SA, Rashidi A, Abdi S, Saedi D, Abouzari M. Liver: an alarm for the heart? *Liver Int*. 2007;27:891–894.
12. Katsiki N, Perez-Martinez P, Anagnostis P, Mikhailidis DP, Karagiannis A. Is nonalcoholic fatty liver disease indeed the hepatic manifestation of metabolic syndrome? *Curr Vasc Pharmacol*. 2018;16:219–227.
13. Juneja A. Non-alcoholic fatty liver disease (NAFLD)—the hepatic component of metabolic syndrome. *JAPI*. 2009;57:201.
14. Lonardo A, Nascimbeni F, Mantovani A, Targher G. Hypertension, diabetes, atherosclerosis and NASH: cause or consequence? *J Hepatol*. 2018;68:335–352.

15. Oni ET, Agatston AS, Blaha MJ, Fialkow J, Cury R, Sposito A, Erbel R, Blankstein R, Feldman T, Al-Mallah MH, et al. A systematic review: burden and severity of subclinical cardiovascular disease among those with nonalcoholic fatty liver; should we care? *Atherosclerosis*. 2013;230:258–267.
16. Stols-Gonçalves D, Hovingh GK, Nieuwdorp M, Holleboom AG. NAFLD and atherosclerosis: two sides of the same dysmetabolic coin? *Trends Endocrinol Metab*. 2019;30:891–902.
17. Ichikawa K, Miyoshi T, Osawa K, Miki T, Toda H, Ejiri K, Yoshida M, Nanba Y, Yoshida M, Nakamura K, et al. Prognostic value of non-alcoholic fatty liver disease for predicting cardiovascular events in patients with diabetes mellitus with suspected coronary artery disease: a prospective cohort study. *Cardiovasc Diabetol*. 2021;20:1–10.
18. Targher G, Corey KE, Byrne CD. NAFLD, and cardiovascular and cardiac diseases: factors influencing risk, prediction and treatment. *Diabetes Metab*. 2021;47:101215. doi: 10.1016/j.diabet.2020.101215
19. Lee SB, Park G-M, Lee J-Y, Lee BU, Park JH, Kim BG, Jung SW, Du Jeong I, Bang S-J, Shin JW, et al. Association between non-alcoholic fatty liver disease and subclinical coronary atherosclerosis: an observational cohort study. *J Hepatol*. 2018;68:1018–1024.
20. Wójcik-Cichy K, Koślińska-Berkan E, Piekarska A. The influence of NAFLD on the risk of atherosclerosis and cardiovascular diseases. *Clin Exp Hepatol*. 2018;4:1.
21. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One*. 2017;12:e0174944. doi: 10.1371/journal.pone.0174944
22. Zhao J, Feng Q, Wu P, Lupu RA, Wilke RA, Wells QS, Denny JC, Wei W-Q. Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction. *Sci Rep*. 2019;9:1–10.
23. Yang L, Wu H, Jin X, Zheng P, Hu S, Xu X, Yu W, Yan J. Study of cardiovascular disease prediction model based on random forest in eastern China. *Sci Rep*. 2020;10:1–8. doi: 10.1038/s41598-020-62133-5
24. Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK Biobank participants. *PLoS One*. 2019;14:e0213653. doi: 10.1371/journal.pone.0213653
25. Chiuve SE, Cook NR, Shay CM, Rexrode KM, Albert CM, Manson JE, Willett WC, Rimm EB. Lifestyle-based prediction model for the prevention of CVD: the Healthy Heart Score. *J Am Heart Assoc*. 2014;3:e000954. doi: 10.1161/JAHA.114.000954
26. Collins R. What makes UK Biobank special? *Lancet*. 2012;9822:1173–1174. doi: 10.1016/S0140-6736(12)60404-8
27. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12:e1001779. doi: 10.1371/journal.pmed.1001779
28. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*. 2015;12:e1001779.
29. Wilman HR, Kelly M, Garratt S, Matthews PM, Milanesi M, Herlihy A, Gyngell M, Neubauer S, Bell JD, Banerjee R, et al. Characterisation of liver fat in the UK Biobank cohort. *PLoS One*. 2017;12:e0172921. doi: 10.1371/journal.pone.0172921
30. Dulai PS, Sirlin CB, Loomba R. MRI and MRE for non-invasive quantitative assessment of hepatic steatosis and fibrosis in NAFLD and NASH: clinical trials to clinical practice. *J Hepatol*. 2016;65:1006–1016.
31. Gu J, Liu S, Du S, Zhang Q, Xiao J, Dong Q, Xin Y. Diagnostic value of MRI-PDFF for hepatic steatosis in patients with non-alcoholic fatty liver disease: a meta-analysis. *Eur Radiol*. 2019;29:3564–3573.
32. Caussy C, Alquiraish MH, Nguyen P, Hernandez C, Cepin S, Fortney LE, Ajmera V, Bettencourt R, Collier S, Hooker J, et al. Optimal threshold of controlled attenuation parameter with MRI-PDFF as the gold standard for the detection of hepatic steatosis. *Hepatology*. 2018;67:1348–1359.
33. Castera L, Friedrich-Rust M, Loomba R. Noninvasive assessment of liver disease in patients with nonalcoholic fatty liver disease. *Gastroenterology*. 2019;156:1264–1281.
34. Naqvi TZ, Mendoza F, Rafii F, Gransar H, Guerra M, Lepor N, Berman DS, Shah PK. High prevalence of ultrasound detected carotid atherosclerosis in subjects with low Framingham risk score: potential implications for screening for subclinical atherosclerosis. *J Am Soc Echocardiogr*. 2010;23:809–815.
35. Caughey MC, Qiao Y, Windham BG, Gottesman RF, Mosley TH, Wasserman BA. Carotid intima-media thickness and silent brain infarctions in a biracial cohort: the Atherosclerosis Risk in Communities (ARIC) study. *Am J Hypertens*. 2018;31:869–875.
36. Nambi V, Chambless L, Folsom AR, He M, Hu Y, Mosley T, Volcik K, Boerwinkle E, Ballantyne CM. Carotid intima-media thickness and presence or absence of plaque improves prediction of coronary heart disease risk: the ARIC (Atherosclerosis Risk In Communities) study. *J Am Coll Cardiol*. 2010;55:1600–1607.
37. Nezu T, Hosomi N, Aoki S, Matsumoto M. Carotid intima-media thickness for atherosclerosis. *J Atheroscler Thromb*. 2016;23:18–31.
38. Mancia G, Fagard R, Narkiewicz K, Redon J, Zanchetti A, Böhm M, Christiaens T, Cifkova R, De Backer G, Dominiczak A, et al. 2013 ESH/ESC practice guidelines for the management of arterial hypertension: ESH-ESC the task force for the management of arterial hypertension of the European Society of Hypertension (ESH) and of the European Society of Cardiology (ESC). *Blood Press*. 2014;23:3–16.
39. Jarvis H, Craig D, Barker R, Spiers G, Stow D, Anstee QM, Hanratty B. Metabolic risk factors and incident advanced liver disease in non-alcoholic fatty liver disease (NAFLD): a systematic review and meta-analysis of population-based observational studies. *PLoS Medicine*. 2020;17:e1003100. doi: 10.1371/journal.pmed.1003100
40. Younossi Z, Anstee QM, Marietti M, Hardy T, Henry L, Eslam M, George J, Bugianesi E. Global burden of NAFLD and NASH: trends, predictions, risk factors and prevention. *Nat Rev Gastroenterol Hepatol*. 2018;15:11.
41. Suykens JAK, Vandewalle J. Least squares support vector machine classifiers. *Neural Process Lett*. 1999;9:293–300.
42. Liaw A, Wiener M. Classification and regression by randomForest. *R News*. 2002;2:18–22.
43. Priddy KL, Keller PE. *Artificial neural networks: an introduction*. Washington, USA: SPIE Press; 2005;68:1–163.
44. Wright RE. *Logistic Regression*. American Psychological Association; 1995.
45. Tibshirani R. Regression shrinkage and selection via the lasso. *J Roy Stat Soc: Ser B (Methodol)*. 1996;58:267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x
46. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*. 1970;12:55–67.
47. Rish I. An empirical study of the naive bayes classifier. In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*. New York: IBM. 2001;3:41–46.
48. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44:837–845. doi: 10.2307/2531595
49. Dietterich TG. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*. Springer; 2000:1–15.
50. El-Sappagh S, Alonso JM, Islam SR, Sultan AM, Kwak KS. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. *Sci Rep*. 2021;11:1–26.
51. Badgeley MA, Zech JR, Oakden-Rayner L, Glicksberg BS, Liu M, Gale W, McConnell MV, Percha B, Snyder TM, Dudley JT. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ Digital Medicine*. 2019;2:1–10.
52. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3:32–35.
53. Sedgwick P. Multiple significance tests: the Bonferroni correction. *BMJ*. 2012;344:e509. doi: 10.1136/bmj.e509
54. Eslam M, Sanyal AJ, George J, Sanyal A, Neuschwander-Tetri B, Tiribelli C, Kleiner DE, Brunt E, Bugianesi E, Yki-Järvinen H, et al. MAFLD: a consensus-driven proposed nomenclature for metabolic associated fatty liver disease. *Gastroenterology*. 2020;158:1999–2014.
55. Costantino S, Paneni F, Cosentino F. Ageing, metabolism and cardiovascular disease. *J Physiol*. 2016;594:2061–2073.
56. Marchisello S, Di Pino A, Scicali R, Urbano F, Piro S, Purrello F, Rabuazzo AM. Pathophysiological, molecular and therapeutic issues of nonalcoholic fatty liver disease: an overview. *Int J Mol Sci*. 2019;20:1948.
57. Whelton SP, McEvoy JW, Shaw L, Psaty BM, Lima JAC, Budoff M, Nasir K, Szklo M, Blumenthal RS, Blaha MJ. Association of normal systolic blood pressure level with cardiovascular disease in the absence of risk factors. *JAMA Cardiol*. 2020;5:1011–1018.
58. Fuchs FD, Whelton PK. High blood pressure and cardiovascular disease. *Hypertension*. 2020;75:285–292.
59. Vasankari V, Husu P, Vähä-Ypyä H, Suni J, Tokola K, Halonen J, Hartikainen J, Sievänen H, Vasankari T. Association of objectively

measured sedentary behaviour and physical activity with cardiovascular disease risk. *Eur J Prev Cardiol*. 2017;24:1311–1318.

60. Patterson R, McNamara E, Tainio M, Hérick de Sá T, Smith AD, Sharp SJ, Edwards P, Woodcock J, Brage S, Wijndaele K. Sedentary behaviour and risk of all-cause, cardiovascular and cancer mortality, and incident type 2 diabetes: a systematic review and dose response meta-analysis. *Springer*. 2018;9:811–829. doi: 10.1007/s10654-018-0380-1
61. Helajärvi H, Pahkala K, Heinonen OJ, Juonala M, Oikonen M, Tammelin T, Hutri-Kähönen N, Kähönen M, Lehtimäki T, Mikkilä V, et al. Television viewing and fatty liver in early midlife: the Cardiovascular Risk in Young Finns Study. *Ann Med*. 2015;47:519–526.
62. Satapathy SK, Sanyal AJ. Epidemiology and natural history of nonalcoholic fatty liver disease. *Semin Liver Dis*. 2015;35:221–235.
63. Anstee QM, Reeves HL, Kotsiliti E, Govaere O, Heikenwalder M. From NASH to HCC: current concepts and future challenges. *Nat Rev Gastroenterol Hepatol*. 2019;16:411–428.
64. Wu J-T, Liu S-S, Xie X-J, Liu Q, Xin Y-N, Xuan S-Y. Independent and joint correlation of PNPLA3 I148M and TM6SF2 E167K variants with the risk of coronary heart disease in patients with non-alcoholic fatty liver disease. *Lipids Health Dis*. 2020;19:1–7.
65. Simons PIHG, Simons N, Stehouwer CDA, Schalkwijk CG, Schaper NC, Brouwers MCGJ. Association of common gene variants in glucokinase regulatory protein with cardiorenal disease: a systematic review and meta-analysis. *PLoS One*. 2018;13:e0206174. doi: 10.1371/journal.pone.0206174
66. Pirola CJ, Sookoian S. The dual and opposite role of the TM6SF2-rs58542926 variant in protecting against cardiovascular disease and conferring risk for nonalcoholic fatty liver: a meta-analysis. *Hepatology*. 2015;62:1742–1756.
67. Brouwers MCGJ, Simons N, Stehouwer CDA, Koek GH, Schaper NC, Isaacs A. Relationship between nonalcoholic fatty liver disease susceptibility genes and coronary artery disease. *Hepatol Commun*. 2019;3:587–596.
68. Tarantino G, Citro V, Conforti P, Balsano C, Capone D. Is there a link between basal metabolic rate, spleen volume and hepatic growth factor levels in patients with obesity-related NAFLD? *J Clin Med*. 2019;8:1510.
69. Tong Z, Li Q, Zhang J, Wei Y, Miao G, Yang X. Association between interleukin 6 and interleukin 16 gene polymorphisms and coronary heart disease risk in a Chinese population. *J Int Med Res*. 2013;41:1049–1056.
70. Chen Y, Huang H, Liu S, Pan L-A, Zhou B, Zhang L, Zeng Z. IL-16 rs11556218 gene polymorphism is associated with coronary artery disease in the Chinese Han population. *Clin Biochem*. 2011;44:1041–1044.
71. Braun J, Meixner A, Brachner A, Foisner R. The GIY-YIG type endonuclease ankyrin repeat and LEM domain-containing protein 1 (ANKLE1) is dispensable for mouse hematopoiesis. *PLoS One*. 2016;11:e0152278. doi: 10.1371/journal.pone.0152278
72. Wallace ML, Ricco JA, Barrett B. Screening strategies for cardiovascular disease in asymptomatic adults. *Prim Care*. 2014;41:371–397.
73. Lindbohm JV, Sipilä PN, Mars NJ, Pentti J, Ahmadi-Abhari S, Brunner EJ, Shipley MJ, Singh-Manoux A, Tabak AG, Kivimäki M. 5-Year versus risk-category-specific screening intervals for cardiovascular disease prevention: a cohort study. *Lancet Public Health*. 2019;4:e189–e199. doi: 10.1016/S2468-2667(19)30023-4
74. Abeles RD, Mullish BH, Forlano R, Kimhofer T, Adler M, Tzallas A, Giannakeas N, Yee M, Mayet J, Goldin RD, et al. Derivation and validation of a cardiovascular risk score for prediction of major acute cardiovascular events in non-alcoholic fatty liver disease; the importance of an elevated mean platelet volume. *Aliment Pharmacol Ther*. 2019;49:1077–1085.
75. Meyersohn NM, Mayrhofer T, Corey KE, Bittner DO, Staziaki PV, Szilveszter B, Hallett T, Lu MT, Puchner SB, Simon TG, et al. Association of hepatic steatosis with major adverse cardiovascular events, independent of coronary artery disease. *Clin Gastroenterol Hepatol*. 2020;19:1480–1488.
76. Kim D, Kim WR, Kim HJ, Therneau TM. Association between noninvasive fibrosis markers and mortality among adults with nonalcoholic fatty liver disease in the United States. *Hepatology*. 2013;57:1357–1365.
77. Choudhary NS, Duseja A. Screening of cardiovascular disease in nonalcoholic fatty liver disease: whom and how? *J Clin Exp Hepatol*. 2019;9:506–514.
78. Siddiqui MS, Fuchs M, Idowu MO, Luketic VA, Boyett S, Sargeant C, Stravitz RT, Puri P, Matherly S, Sterling RK, et al. Severity of nonalcoholic fatty liver disease and progression to cirrhosis are associated with atherogenic lipoprotein profile. *Clin Gastroenterol Hepatol*. 2015;13:1000–1008.
79. Kumar R, Priyadarshi RN, Anand U. Non-alcoholic fatty liver disease: growing burden, adverse outcomes and associations. *J Clin Transl Hepatol*. 2020;8:76.

Supplemental Material

Table S1. Subjects categorized using their medication intake. The second column represents the number of subjects in each category and third column represents their respective percentage in the total population.

Medications category	N	% (N/846)
Cholesterol lowering medication	204	24.11
Blood pressure medication	135	15.95
Insulin	2	0.2
Hormone replacement therapy	30	3.5
Oral contraceptive pill or minipill	5	0.59

Table S2. Distribution of subclinical and clinical cardiovascular events among patients with NAFLD

Cardiovascular disease/Events	No. of Cases (%)
Heart Attack	86 (21.5%)
Angina	71 (17.75%)
Heart Failure	1 (0.25%)
Heart/cardiac problem	114 (28.5%)
Carotid intima-medial thickness (CIMT) > 0.9 mm	194 (48.5%)
Transmural myocardial infarction	1 (0.25%)
Acute myocardial infarction	5 (1.25%)
Myocardial infarction inferior wall	1 (0.25%)
Old Myocardial Infarction	2 (0.5%)
Myocardial infarction anterior wall	1 (0.25%)
Subendocardial myocardial infarction	1 (0.25%)
Ischemic Heart Disease	2 (0.5%)

Table S3. Baseline characteristics of variables in the study. The distribution is summarized based on cardiovascular disease outcome categories of cases and controls. The summary of the continuous variables is represented using median and Inter-Quartile Range (IQR) of their distribution. The summary of categorical variables if represented using frequency and percentage.

Variables	Controls (N=446) (Median [IQR])	Cases (N=400) (Median [IQR])	p-value
Age (years)	Median :55.00 [1st Qu.:48.00 3rd Qu.:60.00]	Median :59.00 [1st Qu.:54.00 3rd Qu.:63.00]	1.74E-15
Weight (kgs)	Median : 83.20 [1st Qu.: 73.30 3rd Qu.: 94.00]	Median : 86.55 [1st Qu.: 75.90 3rd Qu.: 95.78]	7.91E-03
Body Mass Index	Median :28.46 [1st Qu.:25.99 3rd Qu.:31.08]	Median :29.23 [1st Qu.:26.80 3rd Qu.:31.94]	1.55E-03
Proton Density Fat Fraction (%)	Median : 8.694 [1st Qu.: 6.305 3rd Qu.:13.099]	Median : 9.291 [1st Qu.: 6.858 3rd Qu.:13.895]	4.92E-02
Alanine aminotransferase (U/L)	Median : 25.60 [1st Qu.: 19.09 3rd Qu.: 34.06]	Median : 27.24 [1st Qu.: 20.18 3rd Qu.: 36.80]	3.96E-02

Aspartate aminotransferase (U/L)	Median : 25.60 [1st Qu.: 22.12 3rd Qu.: 30.70]	Median : 26.60 [1st Qu.: 22.60 3rd Qu.: 31.73]	4.52E-02
Arterial stiffness	Median : 9.429 [1st Qu.: 7.695 3rd Qu.:11.089]	Median : 10.056 [1st Qu.: 8.237 3rd Qu.: 11.467]	5.85E-03
Mean Maximum CIMT (micrometers)	Median :721.2 [1st Qu.:615.8 3rd Qu.:789.4]	Median : 869.8 [1st Qu.: 721.5 3rd Qu.: 972.0]	1.68E-39
high-sensitivity C-reactive protein (mg/L)	Median : 1.685 [1st Qu.: 0.980 3rd Qu.: 3.083]	Median : 1.830 [1st Qu.: 1.030 3rd Qu.: 3.510]	1.76E-01
LDL (mmol/L)	Median :3.739 [1st Qu.:3.187 3rd Qu.:4.340]	Median :3.529 [1st Qu.:2.956 3rd Qu.:4.133]	1.63E-04
Triglycerides (mmol/L)	Median :1.941 [1st Qu.:1.403 3rd Qu.:2.717]	Median :1.978 [1st Qu.:1.393 3rd Qu.:2.805]	1.00E+00
Age high blood pressure diagnosed (years)	Median :46.00 [1st Qu.:34.50 3rd Qu.:51.50]	Median :50.00 [1st Qu.:40.00 3rd Qu.:55.00]	7.52E-02
Glucose (mmol/L)	Median : 4.969 [1st Qu.: 4.626 3rd Qu.: 5.363]	Median : 5.041 [1st Qu.: 4.632 3rd Qu.: 5.617]	4.73E-02
Amount of alcohol drunk (1 unit= 8mL)	Median :2.00 [1st Qu.:1.00 3rd Qu.:3.00]	Median : 2.000 [1st Qu.: 1.000 3rd Qu.: 3.000]	4.54E-02
LV end diastolic volume (mL)	Median : 136.0 [1st Qu.: 114.0 3rd Qu.: 160.0]	Median :134.0 [1st Qu.:112.0 3rd Qu.:158.0]	8.24E-01
LV end systolic volume (mL)	Median : 60.00 [1st Qu.: 47.00 3rd Qu.: 72.25]	Median : 58.00 [1st Qu.: 47.00 3rd Qu.: 72.00]	5.51E-01
Gamma glutamyltransferase (U/L)	Median : 33.20 [1st Qu.: 22.60 3rd Qu.: 51.95]	Median : 34.10 [1st Qu.: 23.82 3rd Qu.: 50.08]	2.46E-01
Glycated haemoglobin (HbA1c) (mmol/mol)	Median :35.20 [1st Qu.:32.65 3rd Qu.:37.70]	Median :35.90 [1st Qu.:33.40 3rd Qu.:39.20]	9.51E-04
HDL cholesterol (mmol/L)	Median :1.240 [1st Qu.:1.064 3rd Qu.:1.474]	Median :1.254 [1st Qu.:1.081 3rd Qu.:1.417]	8.25E-01

Waist circumference (cm)	Median : 95.00 [1st Qu.: 88.00 3rd Qu.:101.00]	Median : 98.00 [1st Qu.: 91.00 3rd Qu.:105.00]	1.37E-05
Diastolic Blood pressure (mmHg)	Median : 82.00 [1st Qu.: 76.00 3rd Qu.: 89.00]	Median : 85.00 [1st Qu.: 79.00 3rd Qu.: 93.00]	6.92E-07
Systolic Blood Pressure (mmHg)	Median :135.0 [1st Qu.:125.0 3rd Qu.:146.5]	Median :145.0 [1st Qu.:134.0 3rd Qu.:156.0]	4.23E-16
Liver iron (mg/g)	Median :1.304 [1st Qu.:1.204 3rd Qu.:1.427]	Median :1.3183 [1st Qu.:1.1797 3rd Qu.:1.4448]	8.88E-01
Liver inflammation factor (units)	Median :0.8413 [1st Qu.:0.7474 3rd Qu.:1.1046]	Median :0.8107 [1st Qu.:0.7246 3rd Qu.:0.9977]	2.66E-01
Visceral adipose tissue volume (cm3)	Median : 5.366 [1st Qu.: 4.090 3rd Qu.: 6.788]	Median : 5.806 [1st Qu.: 4.253 3rd Qu.: 7.331]	4.34E-03
White blood cell leukocyte count (10^9 cells/Litre)	Median : 6.700 [1st Qu.: 5.700 3rd Qu.: 7.777]	Median : 6.890 [1st Qu.: 6.000 3rd Qu.: 7.980]	2.81E-02
Red blood cell erythrocyte count (10^12 cells/Litre)	Median :4.624 [1st Qu.:4.345 3rd Qu.:4.865]	Median :4.670 [1st Qu.:4.390 3rd Qu.:4.905]	1.18E-01
Haemoglobin concentration (grams/decilitre)	Median :14.50 [1st Qu.:13.80 3rd Qu.:15.38]	Median :14.59 [1st Qu.:13.80 3rd Qu.:15.50]	3.63E-01
Mean corpuscular volume (femtolitres)	Median : 91.16 [1st Qu.: 88.40 3rd Qu.: 94.02]	Median : 90.86 [1st Qu.: 88.55 3rd Qu.: 93.60]	4.97E-01
Red blood cell erythrocyte distribution width (%)	Median :13.24 [1st Qu.:12.88 3rd Qu.:13.70]	Median :13.35 [1st Qu.:12.90 3rd Qu.:13.83]	1.39E-02
Platelet count (10^9 cells/Litre)	Median :250.0 [1st Qu.:213.0 3rd Qu.:288.7]	Median :244.1 [1st Qu.:209.8 3rd Qu.:289.1]	3.03E-01
Mean platelet thrombocyte volume (femtolitres)	Median : 9.100 [1st Qu.: 8.450 3rd Qu.: 9.800]	Median : 9.100 [1st Qu.: 8.500 3rd Qu.: 9.900]	4.97E-01
Platelet distribution width (%)	Median :16.49 [1st Qu.:16.16 3rd Qu.:16.80]	Median :16.48 [1st Qu.:16.21 3rd Qu.:16.84]	3.37E-01

Albumin (g/L)	Median :45.41 [1st Qu.:43.70 3rd Qu.:47.15]	Median :45.52 [1st Qu.:43.84 3rd Qu.:47.24]	4.63E-01
Alkaline phosphatase (U/L)	Median : 78.90 [1st Qu.: 67.70 3rd Qu.: 94.67]	Median : 82.10 [1st Qu.: 68.42 3rd Qu.: 97.47]	4.27E-02
Direct bilirubin (umol/L)	Median :1.620 [1st Qu.:1.310 3rd Qu.:2.038]	Median :1.640 [1st Qu.:1.310 3rd Qu.:2.190]	1.87E-01
Cholesterol (mmol/L)	Median :5.824 [1st Qu.:5.103 3rd Qu.:6.599]	Median : 5.589 [1st Qu.: 4.840 3rd Qu.: 6.332]	6.45E-04
Creatinine (umol/L)	Median : 72.70 [1st Qu.: 62.98 3rd Qu.: 82.30]	Median : 73.95 [1st Qu.: 64.20 3rd Qu.: 83.55]	1.33E-01
C.reactive protein (mg/L)	Median : 1.685 [1st Qu.: 0.980 3rd Qu.: 3.083]	Median : 1.830 [1st Qu.: 1.030 3rd Qu.: 3.510]	1.76E-01
Total bilirubin (umol/L)	Median : 8.310 [1st Qu.: 6.675 3rd Qu.:10.523]	Median : 8.290 [1st Qu.: 6.655 3rd Qu.:10.373]	9.96E-01
Total protein (g/L)	Median :72.37 [1st Qu.:70.08 3rd Qu.:74.52]	Median :72.56 [1st Qu.:70.14 3rd Qu.:75.08]	3.95E-01
Urate (umol/L)	Median :334.9 [1st Qu.:281.8 3rd Qu.:386.8]	Median :348.2 [1st Qu.:294.1 3rd Qu.:404.9]	6.53E-03
Vitamin D (nmol/L)	Median : 46.30 [1st Qu.: 32.55 3rd Qu.: 59.10]	Median : 47.30 [1st Qu.: 32.60 3rd Qu.: 62.40]	3.79E-01
Salt Intake Never Sometimes Usually Always	256 (57.4%) 131 (29.3%) 17 (3.8%) 14 (3.1%)	232 (58%) 133 (33.25%) 11 (2.7%) 7 (1.75%)	0.001
Sex Female (N(%)) Male (N(%))	213 (47.7%) 223 (52.3%)	174 (43.5%) 226 (56.5%)	0.2
Diabetes Yes (N(%)) No (N(%))	80 (17.9%) 366 (82.1%)	132(33%) 268 (67%)	1.83E-06
Ethnicity Caucasian East Asian South east Asian Black Other	433 (97%) 4 (0.9%) 6 (1.3%) 0 (0%) 3 (0.7%)	392 (98%) 2 (0.5%) 3 (0.75%) 1 (0.25%) 3 (0.75%)	0.978

Moderate Physical Activity Yes (N(%)) No (N(%))	 147 (32.95%) 284 (65.91%)	 137 (34.2%) 261 (65.2%)	 0.81
Alcohol Consumption Status Yes (N(%)) No (N(%))	 405 (90.6%) 40 (9.1%)	 373 (93.2%) 26 (6.5%)	 0.48
Smoking Status Yes (N(%)) No (N(%))	174 (39%) 265 (59.41%)	190 (47.5%) 207 (51.75%)	0.02
Watching TV <1h (N(%)) 1h - 4h (N(%)) >4h (N(%))	253 (56.7%) 155 (34.8%) 25 (5.6%)	143 (35.7%) 110 (27.5%) 16 (4%)	1.98E-04
Using Computer <1h (N(%)) 1h - 4h (N(%)) >4h (N(%))	252 (56.5%) 121 (27.1%) 51 (11.43%)	226 (56.5%) 110 (27.5%) 34 (8.5%)	0.61

Table S4. Performance of 4 integrative ensemble modelling techniques on integrating the clinical and genetic domain (second column) and clinical, genetic and lifestyle domains (third column). The top row in bold shows that Naïve Bayes methodology performed the best in terms of mean AUC and 95% confidence intervals for predicting cardiovascular outcome in NAFLD subjects using the integrative methodology.

Methods	Integration (Clinical + Genetic) AUC [95% CI]	Integration (Clinical+ Genetic + Lifestyle) AUC [95% CI]
Naïve Bayes	**0.820 [0.811, 0.828]**	**0.849 [0.840, 0.855]**
Random Forest	0.811 [0.791, 0.829]	0.841 [0.828, 0.853]
Ada Boost	0.817 [0.800, 0.831]	0.835 [0.821, 0.845]
Bagging	0.800 [0.783, 0.818]	0.838 [0.823, 0.852]

Table S5. AUC values comparing performance of ML approaches in the clinical CVD subgroup to the sub-clinical CVD subgroup.

Methods	**Clinical CVD subgroup (Cases=285)**	**Sub-clinical CVD (CIMT) subgroup (Cases=194)**
Random Forest	0.687	0.617
Lasso Regression	0.687	0.621
Ridge Regression	0.698	0.613
Naïve Bayes	0.688	0.617
Support Vector Machines	0.683	0.611
Logistic Regression	0.698	0.629
Neural Network	0.635	0.589

Table S6. Mean AUC values comparing ML approaches after adding medications to the input data while prediction as opposed to considering AUC without medications.

Methods	AUC without Medications	AUC with Medications
Random Forest	0.799	0.808
Lasso Regression	0.753	0.754
Ridge Regression	0.747	0.751
Naïve Bayes	0.744	0.752
Support Vector Machines	0.743	0.754
Logistic Regression	0.742	0.750
Neural Network	0.728	0.731

Table S7. p-values obtained through Univariate analysis of the variables with Cardiovascular diseases as the outcome.

Variables	p-value
Age	7.12E-15
Ethnicity	0.978753
Weight	0.010588
BMI	0.009055
BMI_categorical	0.016817
Alcohol Drinker Status	0.484727
Proton Density Fat Fraction	0.214678
Alanine aminotransferase	0.291229
Aspartate aminotransferase	0.354439
Liver Inflammation factor	0.414832
Arterial stiffness	0.021088
Mean Maximum CIMT	4.30E-21

high-sensitivity C-reactive protein	0.531247
LDL	0.000247
Triglycerides	0.565215
Age high blood pressure diagnosed	0.035106
Glucose	0.024351
Amount of alcohol drunk	0.108914
LV end diastolic volume	0.95738
LV end systolic volume	0.98677
Gamma glutamyltransferase	0.631237
Glycated haemoglobin (HbA1c)	0.009174
HDL cholesterol	0.535533
Waist circumference	3.83E-05
Diastolic Blood pressure	4.05E-07
Systolic Blood Pressure	4.69E-14
Smoking Status	0.006624
Liver iron	0.845729
Visceral adipose tissue volume	0.005375
White blood cell leukocyte count	0.079771
Red blood cell erythrocyte count	0.085268
Haemoglobin concentration	0.351264
Mean corpuscular volume	0.31351
Red blood cell erythrocyte distribution width	0.017247
Platelet count	0.367095
Mean platelet thrombocyte volume	0.433889
Platelet distribution width	0.326936
Albumin	0.408899
Alkaline phosphatase	0.049853
Direct bilirubin	0.070975
Cholesterol	0.00047
Creatinine	0.03331
C reactive protein	0.531247
Total bilirubin	0.603702
Total protein	0.276138
Urate	0.005244
Vitamin D	0.479191
Diabetes	1.83E-06

Figure S1. 95% confidence intervals obtained for the mean AUC values for 10 times 10-fold cross validation on the training set comprising of the clinical variables for the study cohort.

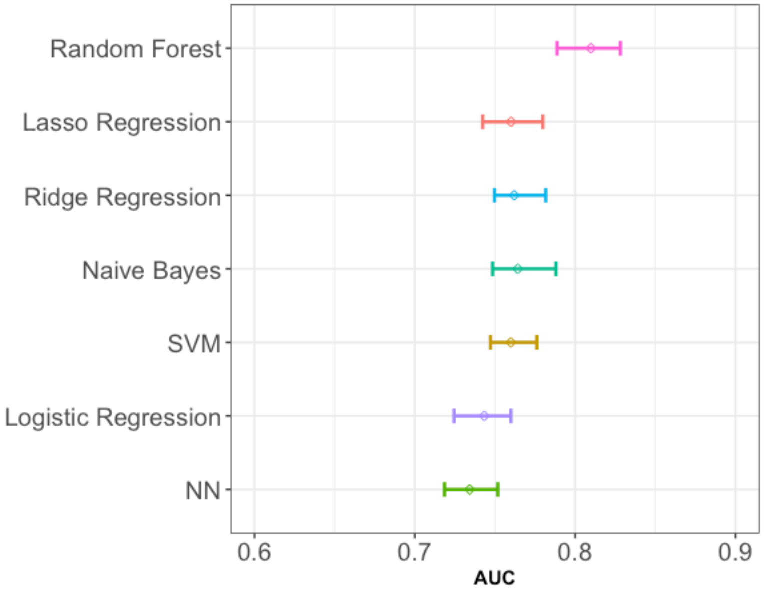
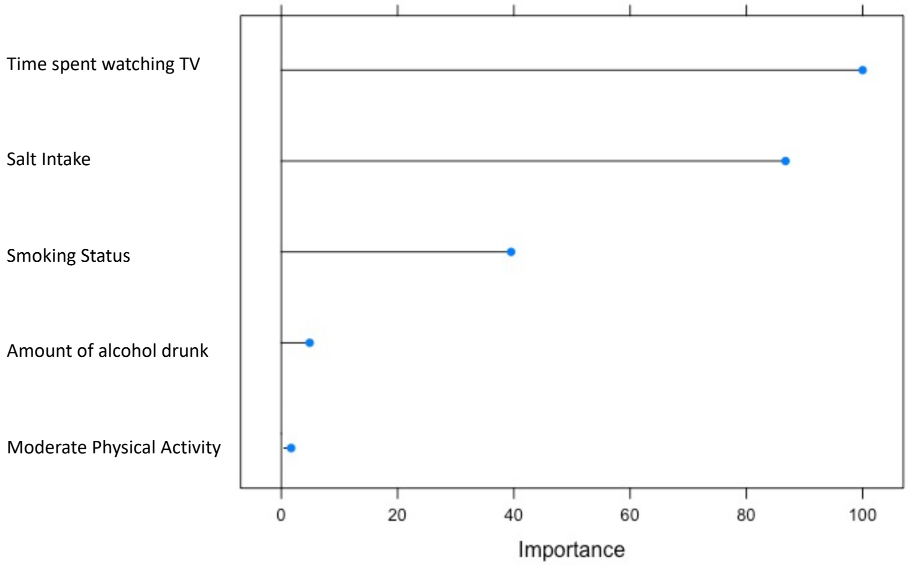


Figure S2. Variable importance plot demonstrating the importance of Lifestyle data variables obtained through the Machine learning modelling on the lifestyle data.


Time spent watching TV
Salt Intake
Smoking Status
Amount of alcohol drunk
Moderate Physical Activity
Importance