Indian Journal of Gastroenterology (September–October 2022) 41(5):475–482
https://doi.org/10.1007/s12664-022-01263-2

ORIGINAL ARTICLE

A machine-learning approach for nonalcoholic steatohepatitis susceptibility estimation

Fatemeh Ghadiri[1,2] · Abbas Ali Husseini[3] · Oğuzhan Öztaş[1]

Received: 11 August 2021 / Accepted: 2 May 2022 / Published online: 11 November 2022

Abstract

Background Nonalcoholic steatohepatitis (NASH), a severe form of nonalcoholic fatty liver disease, can lead to advanced liver damage and has become an increasingly prominent health problem worldwide. Predictive models for early identification of high-risk individuals could help identify preventive and interventional measures. Traditional epidemiological models with limited predictive power are based on statistical analysis. In the current study, a novel machine-learning approach was developed for individual NASH susceptibility prediction using candidate single nucleotide polymorphisms (SNPs).

Methods A total of 245 NASH patients and 120 healthy individuals were included in the study. Single nucleotide polymorphism genotypes of candidate genes including two SNPs in the cytochrome P450 family 2 subfamily E member 1 (CYP2E1) gene (rs6413432, rs3813867), two SNPs in the glucokinase regulator (GCKR) gene (rs780094, rs1260326), rs738409 SNP in patatin-like phospholipase domain-containing 3 (PNPLA3), and gender parameters were used to develop models for identifying at-risk individuals. To predict the individual's susceptibility to NASH, nine different machine-learning models were constructed. These models involved two different feature selections including Chi-square, and support vector machine recursive feature elimination (SVM-RFE) and three classification algorithms including k-nearest neighbor (KNN), multi-layer perceptron (MLP), and random forest (RF). All nine machine-learning models were trained using 80% of both the NASH patients and the healthy controls data. The nine machine-learning models were then tested on 20% of both groups. The model's performance was compared for model accuracy, precision, sensitivity, and *F* measure.

Results Among all nine machine-learning models, the KNN classifier with all features as input showed the highest performance with 86% *F* measure and 79% accuracy.

Conclusions Machine learning based on genomic variety may be applicable for estimating an individual's susceptibility for developing NASH among high-risk groups with a high degree of accuracy, precision, and sensitivity.

Keywords Algorithm · Artificial intelligence · Disease susceptibility · Fatty liver · Gene · Machine learning · Neural network model · Nonalcoholic fatty liver disease · Nonalcoholic steatohepatitis · Single nucleotide polymorphism · Support vector machine

✉ Fatemeh Ghadiri
fateme.ghadiry@gmail.com

[1] Department of Computer Engineering, Istanbul University Cerrahpaşa, 34320 Istanbul, Turkey

[2] Computer Programming, Vocational School, Nişantaşı University, 1453 Istanbul, Turkey

[3] Life Science, and Biomedical Engineering Application and Research Center, Istanbul Gelisim University, 34310 Istanbul, Turkey

Bullet points of the study highlights

What is already known?

- Currently, no reliable predictive model is available to predict nonalcoholic steatohepatitis (NASH) risk.
- The genetic factors affecting the development of NASH and NASH-derived hepatocellular carcinoma, are not well-delineated.

What is new in this study?

- In the current study, we developed a novel machine learning-based method to predict the individual susceptibility to develop NASH based on the candidate gene/single nucleotide polymorphisms (SNPs).

What are the future clinical and research implications of the study findings?

- Machine learning based on genomic variety may be applicable for estimating susceptibility for developing NASH among high-risk groups with a high degree of accuracy, precision, and sensitivity.

Introduction

Nonalcoholic fatty liver disease (NAFLD) is a health problem that is rising globally with a prevalence of 25% to 30% [1, 2]. The prevalence of NAFLD among high-risk populations may exceed 70% to 90% [1]. NAFLD has a wide spectrum of abnormalities. Although simple steatosis, which is considered a less severe form of NAFLD, covers the main share of the spectrum, approximately 7% to 30% of patients develop nonalcoholic steatohepatitis (NASH), which is the severe form of NAFLD [1]. This can lead to advanced hepatocellular damage, inflammation, liver fibrosis, or cirrhosis or even hepatocellular carcinoma [1]. Numerous conditions including genetic aptitude and metabolic syndromes such as obesity, insulin resistance, type 2 diabetes, dyslipidemia, and hypertension may act in parallel to impact the development of NASH [2]. Despite the fact that NASH is increasing globally, the disease is underdiagnosed due to the absence of clear symptoms and the lack of reliable markers. There is a need to intervene early and identify NASH patients before advanced fibrosis and irreversible liver damage occur.

Risk stratification has emerged as a fundamental issue in preventive strategies and disease management [3]. Although there are many clinical, biochemical, metabolic, and lipid biomarkers used to predict NASH, currently, no reliable predictive model is available [4].

Genome-wide association studies and candidate gene studies have provided insight into a part of the genetic variants associated with NASH development among different populations. Disease prediction based on a combination of single nucleotide polymorphism (SNP) parameters and clinical factors has been modeled previously [5–7]. Although several studies show a strong association of patatin-like phospholipase domain-containing 3 (PNPLA3) and transmembrane 6 superfamily member 2 (TM6SF2) with development of NASH, the genetic factors affecting the development of NASH and NASH–derived hepatocellular carcinoma are not well-delineated [5, 6].

Currently, all prediction modeling techniques, which use SNP associated with NASH have adopted a traditional statistical test. Machine learning emerges as a unique technique for uncovering potential biological interactions for better prediction and diagnosis of complex diseases like NASH [8].

In the current study, a novel machine-learning-based method is developed to predict the individual susceptibility to development of NASH based on the candidate gene/SNPs including two SNPs in the cytochrome P450 family 2 subfamily E member 1 (CYP2E1) gene (rs6413432, rs3813867), two SNPs in the glucokinase regulator (GCKR) gene (rs780094, rs1260326), rs738409 SNP in PNPLA3, and gender. The method used a two-step approach that involves feature selection and classification. The method was subsequently trained and tested to compare the performance metrics (Fig. 1).

Methods

Study population and SNP genotyping

The medical records of NASH patients who had histologic and biochemical evidences (e.g. abnormal aspartate aminotransferase [AST], alanine aminotransferase [ALT], gamma-glutamyl transferase [GGT], and hyperbilirubinemia) were studied. Individuals who were biopsy-proven NASH patients using the steatosis, activity and fibrosis (SAF) scoring system were included in the patient group. The age range of

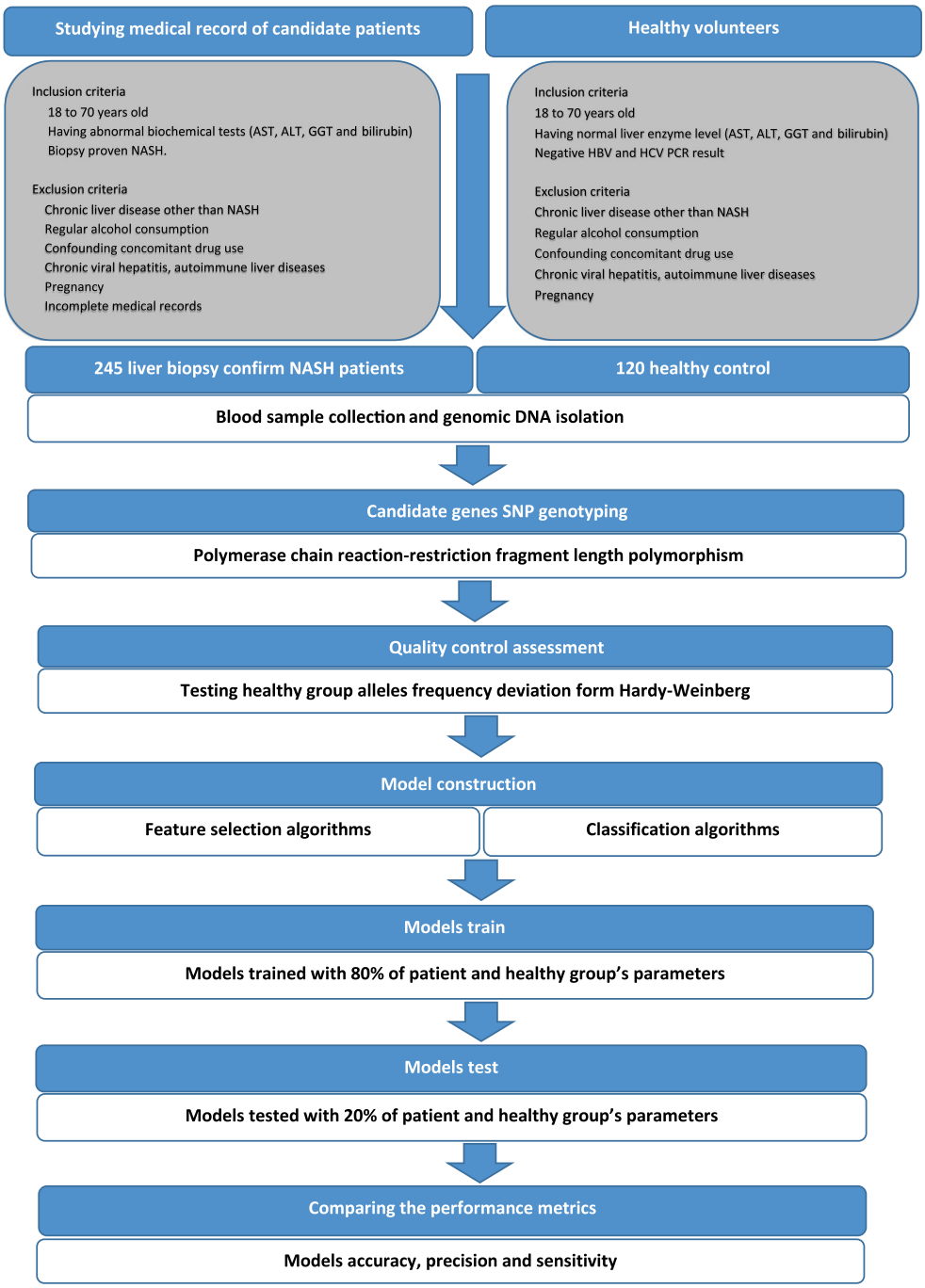

Fig. 1 Graphical abstract of novel machine-learning model to predict an individual's susceptibility to nonalcoholic steatohepatitis. *AST* aspartate aminotransferase, *ALT* alanine aminotransferase, *GGT* gamma-glutamyl transferase, *NASH* nonalcoholic steatohepatitis, *HBV* hepatitis B virus, *HCV* hepatitis C virus, *PCR* polymerase chain reaction, *SNP* single nucleotide polymorphism, *DNA* deoxyribonucleic acid

participants was 18–70 years. Individuals who had chronic liver disease other than NASH such as chronic viral hepatitis, autoimmune liver diseases, and metabolic liver diseases (e.g. Wilson's disease, Crigler-Najjar/Rotor syndrome, and hemochromatosis); confounding concomitant drug users (e.g. glucocorticoids, aspirin, tamoxifen, synthetic estrogens, methotrexate, and calcium-channel blockers); regular alcohol consumers; pregnancy; or incomplete medical records were all excluded. Finally, 245 patients were enrolled in the study and were invited to participate in the study and provide blood samples.

One hundred and twenty healthy volunteers with the same age range who showed normal liver enzyme levels, normal bilirubin, negative results for hepatitis B virus and hepatitis C virus polymerase chain reaction (PCR) tests, and no historical evidence of metabolic diseases that meet the exclusion criteria were included in the study as a healthy control group.

An in-house-modified salt-out method for deoxyribonucleic acid (DNA) extraction from whole blood was used to extract

genomic DNA from 20-mL peripheral blood samples taken from participants. Subsequently, SNP genotyping from the candidate genes involving two SNPs in the CYP2E1 gene (rs6413432, rs3813867), two SNPs in the GCKR gene (rs780094, rs1260326), and one known rs738409 SNP in the PNPLA3 gene were determined by using a polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP) approach.

A fragment of each region was amplified using specific primers in the optimized conditions via in-house-developed protocols. The rs6413432 and rs3813867 SNPs inside the CYP2E1 gene were amplified in 236-base-pair (bp) and 244-bp fragments that respectively were restricted from the TTT/AAA and AG/CT cut sites using DraI and AluI enzymes. In the same way, rs780094 and rs1260326 SNPs, which place at the GCKR gene, were amplified in 196 bp and 199 bp length via conventional PCR and subsequently were restricted from the allele-specific site (A/CATGT, C/CGG), using PCiI and MspI enzymes, respectively. In addition, a 168-bp fragment, involving rs738409 SNP in PNPLA3, was amplified and afterward was restricted via the NlaIII enzyme from the CATG/ cut site (Table 1). Later, restricted fragments were run over the 2% agarose gel electrophoresis to determine the genotypes of each allele and their frequency among NASH patients and those in the control group. The data set used in this study contained phenotype information, allele distribution of each SNP, and participant gender (all enzymes purchased from MilliporeSigma, Massachusetts, USA).

Data quality control

Allele frequency distributions were checked for Hardy-Weinberg proportions (HWP) using the following formula:

$$P = \frac{2 \times obs(AA) + obs(Aa)}{2 \times (obs[AA] + obs[Aa] + obs[aa])}$$
$$q = 1 - p$$

$$\text{Hardy–Weinberg equilibrium} = p2 + 2pq + q2 = 1$$

where A and a are the symbols for dominant and recessive relevant allele nucleotide variants, p^2 is the relative frequency of homozygotes for allele A, $2pq$ is the relative frequency of heterozygotes for alleles A and a, and q^2 is the relative frequency of homozygotes for allele a [9].

To test deviation from HWP and the comparability of observed genotype frequencies between NASH patients and the healthy control group, the differences between the observed and expected allele frequency in the healthy group were assessed via Pearson's Chi-square test with 95% confidence level. The significance level was defined as p-value ≤ 0.05.

Therefore, the Hardy-Weinberg allele frequency expectation was calculated via the following formula:

$$EXP\ (AA) = p^2 n$$
$$EXP\ (Aa) = 2pqn$$
$$EXP(aa) = q^2 n$$

where n is the total number of healthy participants.

Feature selection and classification algorithms

For the development of a high-metrics prediction model, feature selection algorithms including filter-based method (e.g. Chi-square) and wrapper method-like support vector machines-recursive feature elimination (SVM-RFE) were implemented to identify the features, which are highly associated with the phenotype. Additionally, three classification algorithms including k-nearest neighbor (kNN), multi-layer perceptron (MLP), and random forest (RF) were used for classification of patients and heathy individuals. All the algorithms were developed using Python version 3.8.0 (Python Software Foundation, Delaware, USA).

Model construction

To predict the individuals' susceptibility to NASH, nine different machine-learning models were constructed. These

Table 1 Candidate genes/single nucleotide polymorphisms and characteristics of the amplified fragment

Gene	SNP	Primers: 5′……..›3′	Fragment size	Cut sit	Restriction enzyme
CYP2E1	rs6413432	Forward: AGGCTCGTCAGTTCCTGAAA Reverse: ACCACACCCGGCTACTTTTT	236 bp	5′.....TTT↓AAA …3′ 3′... AAA↑TTT….5′	Dra I
	rs3813867	Forward: CCAGTCGAGTCTACATTGTCA Reverse: TTCATTCTGTCTTCTAACTGG	244 bp	5′...AG↓CT…3′ 3′....TC↑GA…5′	Alu I
GCKR	rs780094	Forward: GATTGTCTCAGGCAAACCTGGTAG Reverse: CATGTTGGCTAGGCTTGTTG	196 bp	5′….A↓CATGT …3′ 3′...TGTAC↑A...5′	PCi I
	rs1260326	Forward: CTGGATGGTGAGAGGGAAGAT Reverse: CCCTACAGCCTTGGGTTTTT	199 bp	5′….C↓CGG …3′ 3′...GGC↑C...5′	Msp I
PNPLA3	rs738409	Forward: GCCCTGCTCACTTGGAGAAA Reverse: TGAAAGGCAGTGAGGCATGG	168 bp	5′…..CATG↓…3′ 3′....↑GTAC…..5′	Nla III

SNP single nucleotide polymorphism, *CYP2E1* cytochrome P450 family 2 subfamily E member 1, *GCKR* glucokinase regulator, *PNPLA3* patatin-like phospholipase domain-containing 3, *bp* base pair

models included six integrated machine learning and three additional classification algorithms.

Each integrated machine-learning model was made up of one feature selection and one classification algorithm. Selected parameters in the feature selection algorithms subsequently were used as inputs to classifiers. Moreover, all parameters used as inputs were the only classifier algorithms applied.

All nine machine-learning models were trained with randomly selected input parameters of 80% of patients and 80% of the healthy control group. Subsequently, the models were tested with 20% of NASH patients and 20% of healthy samples' parameters. The performance of the models was compared for model accuracy, precision, sensitivity, and *F* measure metrics.

Results

The patients' and healthy individuals' ages were in the range of 18–70, and the mean age was 44.4 ± 15.1 and 42.7 ± 14.8 years, respectively. Overall, with 245 NASH patients and 120 healthy controls, the female/male ratio was 129/116 in the NASH group and 79/41 in the control group.

The gene/SNP alleles and their frequencies are presented in Table 2. Allele frequencies in all SNPs were within HWP. No deviation from the HWP was detected in the genotype distribution of healthy groups (p-value > 0.05). Among the six parameters used as input, rs738409, rs3813867, rs1260326, and gender showed a high association with NASH (p-value ≤ 0.05) when a Chi-square filter was applied. The SVM-RFE algorithm determined that rs738409, rs3813867, rs780094, rs1260326 SNPs, and gender variables were the best parameters for a high-performance predicting model (Table 3).

Among all nine machine-learning models, the KNN classifier algorithm with no feature selection demonstrated the highest accuracy (79%) followed by SVM-RFE-MLP, SVM-RFE-RF, Chi-square-RF, and RF (78%). In contrast, Chi-square MLP (85%) and RF (85%) models showed the highest precision. Moreover, the SVM-RFE-MLP (94%), KNN (92%), and Chi-square-RF (90%) models demonstrated the highest sensitivity. However, the *F* measures were over 80% and were approximately close. Howbeit, the KNN classifier with all features as input showed the highest performance among all models with an 86% *F* measure and 79% accuracy (Fig. 2).

Table 2 Candidate gene/single nucleotide polymorphism allele's frequency distribution

Gene/SNP	Alleles	Allele frequency	
		NASH patients (*n*)	Healthy (*n*)
CYP2E1; rs6413432	AA	3	2
	TA	38	19
	TT	204	99
CYP2E1; rs3813867	GC	6	9
	GG	239	111
	CC	0	0
GCKR; rs780094	AA	6	4
	GA	3	0
	GG	236	116
GCKR; rs1260326	CC	33	28
	CT	116	60
	TT	96	32
PNPLA3; rs738409	CC	84	67
	GC	72	43
	GG	89	10

SNP single nucleotide polymorphism, *CYP2E1* cytochrome P450 family 2 subfamily E member 1, *GCKR* glucokinase regulator, *PNPLA3* patatin-like phospholipase domain-containing 3, *NASH* nonalcoholic steatohepatitis

Discussion

Traditional epidemiological analysis that relies heavily on logistic regression to anticipate the potential association of clinical variables or genetic factors to a specific disease has limited predictive power [10, 11]. In the past decade, genome association studies have been used to identify genetic variants related to diseases. Machine-learning algorithms have emerged as an effective method for risk prediction of complex diseases due to their ability to handle multi-dimensional data [12].

The existing machine-learning models developed regarding NASH are mainly developed based on socio-demographic, laboratory, and clinical parameters that can be applied for diagnosis, not for early prediction. Machine-learning models with laboratory parameter input for the diagnosis of NAFLD in the general population have been developed with 87% overall accuracy, 92% (86% to 96%) sensitivity, and 90% (86% to 93%) precision [13]. In another model, developed by Canbay et al. (2019), the training dataset included the serum parameters AST, ALT, AST/ALT ratio, GGT, albumin, total cholesterol, triacylglycerols, fasting blood sugar, hemoglobin A1c, thrombocyte count, caspase-cleaved serum CK-18, and adiponectin, as well as socio-demographic parameters such as gender, age, height, weight, and body mass index [14]. Fialoke et al. have developed a machine-learning model based on demographic properties (age, gender, and race), type 2 diabetes status, and longitudinal lab parameters of ALT, AST, and platelets to distinguish NASH patients from healthy controls [15]. Perakakis et al. presented a machine-learning-based predictive algorithm using omics data (lipidomics,

Table 3 Feature selection methods and selected parameters

Feature selection methods	Parameters					
	Gender	rs6413432	rs3813867	rs780094	rs1260326	rs738409
Chi-square	√		√		√	√
SVM-RFE algorithm	√		√	√	√	√

SVM-RFE support vector machine recursive feature elimination

glycomics), and hormone values that can differentiate NASH patients from healthy controls with high accuracy (up to 90%) [16]. Chiappini et al. developed a random forest–based machine-learning method that allowed discriminating NASH with 100% sensitivity and specificity based on characterizing a signature of 32 lipids [17].

In the current study, nine machine-learning models were developed to evaluate the SNP genotype–based predictive model for early detection of high-risk individuals among at-risk groups. The results from the screening model revealed the gender, rs738409, rs3813867, rs780094, and rs1260326 as the most discriminative features in the data set.

PNPLA3 rs738409 polymorphism is a well-known predisposing factor for NAFLD, fibrosis, and alcoholic cirrhosis. The PNPLA3 gene variants show significant association with high serum ALT [18]. Several studies demonstrate that G-allele carriers of rs738409, particularly of the GG genotype, are highly associated with a greater risk of progressive NASH, fibrosis, and hepatocellular carcinoma. G-allele was found to be significantly associated with high AST, ALT, ferritin levels, and the fibrosis stage in patients with NAFLD [18–21]. It is presumed that there is an association between the rs780094 and rs1260326 variations in GCKR and predisposition to NASH. Such variations along with PNPLA3 rs738409 polymorphism interact to increase an individual's susceptibility [22]. The CYP2E1 gene involves genetic polymorphisms with a high variety in frequency among different ethnic groups [23]. However, the CYP2E1 rs3813867 polymorphism is not considered as a genomic factor associated with developing NASH. The current study suggests that rs3813867 SNP inside the CYP2E1 gene is a highly associated biomarker for developing NASH. Studies show age and gender differences in prevalence and severity of NAFLD and NASH. Gender differences in NASH susceptibility have been demonstrated in an animal study [24].

Currently, NASH is one of the leading causes for liver transplantation, particularly in females [25]. Although the prevalence of NASH in younger ages is more common among men, the disease is becoming more common among women in older ages, particularly in those over 60 years. In individuals who develop NASH, 37.6% of them develop progressive fibrosis [26].

Model accuracy is the rate of true predictions made from all predictions and is widely used because it is one single measure for summarizing model performance. However, to decide whether model robustness is enough, accuracy alone is not adequate to predict NASH. The true-positive rate is a more accurate predictor. Low precision and sensitivity, respectively, may result in excessive false positives and false negatives. Precision and sensitivity as a measure of a model's exactness and completeness play a critical role in decision-making about the predisposition and diagnosis of NASH. Models that have

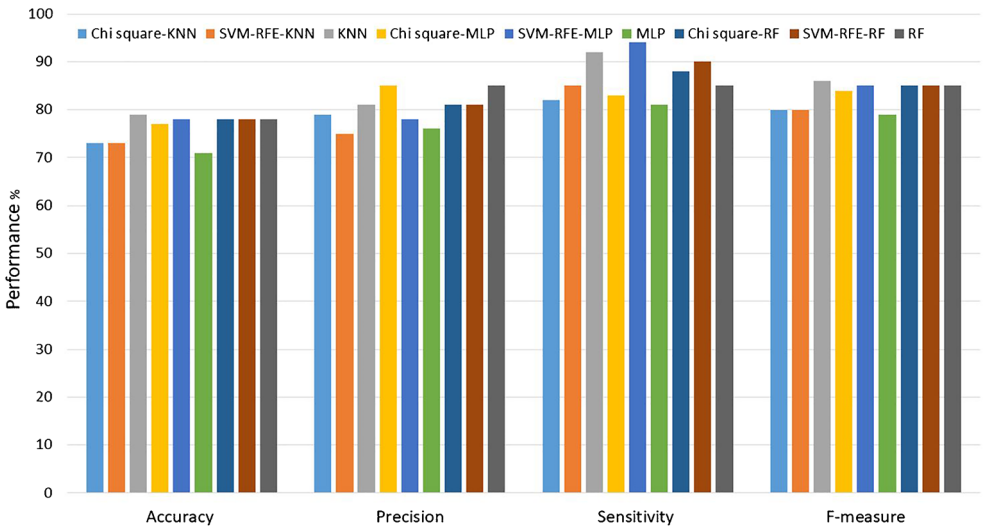

Fig. 2 Comparison of machine-learning models' performance developed in this study to predict individuals' susceptibility to non-alcoholic steatohepatitis. Nine different models were compared in terms of accuracy, precision, sensitivity, and *F* measure. KNN classifier showed the highest performance with an *F* measure of 86%, which is the harmonic mean of sensitivity and precision, and an accuracy of 79%. *KNN* k-nearest neighbor, *SVM-RFE* support vector machine recursive feature elimination, *MLP* multilayer perceptron, *RF* random forest

high precision with low sensitivity or models with both low precision and low sensitivity do not provide enough parameters for making a true decision. Therefore, *F* measure, the harmonic mean of precision and sensitivity, provides a way to express both concerns with a single score of sensitivity and precision using a factor that controls their relative importance [27].

We developed superlative algorithms for the prediction of individual NASH development susceptibility. Moreover, the current study demonstrated that machine learning based on genomic variety may be applicable for estimating susceptibility for developing NASH among high-risk groups with a high degree of accuracy, precision, and sensitivity.

Author contribution Concept: FG, AAH; design: FG; supervision: AAH, OÖ; materials: AAH; data collection and/or analysis: AAH, FG; literature search: FG; writing: FG, AAH; critical reviews: AAH

Data availability The datasets analyzed during the current study are available in the ZENODO repository and can be accessed from https://doi.org/10.5281/zenodo.4686908.

Compliance with ethical standards

Competing interests FG, AAH and OÖ declare no competing interests.

Ethics statement The study was performed conforming to the Helsinki declaration of 1975, as revised in 2000 and 2008 concerning human and animal rights, and the authors followed the policy concerning informed consent as shown on Springer.com.

Ethics approval The ethics committee of Istanbul Gelişim University approved this study (ethical code: 77366270-302.08.01-E.12978, date: 16.11.2020).

Consent to participate Consent forms were signed by all the participants before being included in the study.

Consent for publication Not applicable.

Disclaimer The authors are solely responsible for the data and the contents of the paper. In no way is the honorary editor in chief, editorial board members, the Indian Society of Gastroenterology or the printer/publishers responsible for the results/findings and content of this article.

References

1. Caligiuri A, Gentilini A, Marra F. Molecular pathogenesis of NASH. Int J Mol Sci. 2016;17:1575.
2. Adams LA, Feldstein AE. Nonalcoholic steatohepatitis: risk factors and diagnosis. Expert Rev Gastroenterol Hepatol. 2010;4:623–35.
3. Vespasiani-Gentilucci U, Gallo P, Dell'Unto C, Volpentesta M, Antonelli-Incalzi R, Picardi A. Promoting genetics in non-alcoholic fatty liver disease: combined risk score through polymorphisms and clinical variables. World J Gastroenterol. 2018;24:4835–45.
4. Vilar-Gomez E, Chalasani N. Non-invasive assessment of non-alcoholic fatty liver disease: clinical prediction rules and blood-based biomarkers. J Hepatol. 2018;68:305–15.
5. Anstee QM, Seth D, Day CP. Genetic factors that affect risk of alcoholic and nonalcoholic fatty liver disease. Gastroenterology. 2016;150:1728–44.e7.
6. Kawaguchi T, Shima T, Mizuno M, et al. Risk estimation model for nonalcoholic fatty liver disease in the Japanese using multiple genetic markers. PLoS One. 2018;13:e0185490.
7. Koo BK, Joo SK, Kim D, et al. Development and validation of a scoring system, based on genetic and clinical factors, to determine risk of steatohepatitis in Asian patients with nonalcoholic fatty liver disease. Clin Gastroenterol Hepatol. 2020;18:2592-9.e10.
8. Gaudillo J, Rodriguez JJR, Nazareno A, et al. Machine learning approach to single nucleotide polymorphism-based asthma prediction. PLoS One. 2019;14:e0225574.
9. Ostrovski V. New equivalence tests for Hardy–Weinberg equilibrium and multiple alleles. Stats. 2020;3:34–9.
10. Wang X, Strizich G, Hu Y, Wang T, Kaplan RC, Qi Q. Genetic markers of type 2 diabetes: progress in genome-wide association studies and clinical application for risk prediction. J Diabetes. 2016;8:24–35.
11. Ma H, Xu CF, Shen Z, Yu CH, Li YM. Application of machine learning techniques for clinical predictive modeling: A cross-sectional study on nonalcoholic fatty liver disease in China. Biomed Res Int. 2018;2018:4304376.
12. Ho DSW, Schierding W, Wake M, Saffery R, O'Sullivan J. Machine learning SNP based prediction for precision medicine. Front Genet. 2019;10:267.
13. Yip TC, Ma AJ, Wong VW, et al. Laboratory parameter-based machine learning model for excluding non-alcoholic fatty liver disease (NAFLD) in the general population. Aliment Pharmacol Ther. 2017;46:447–56.
14. Canbay A, Kälsch J, Neumann U, et al. Non-invasive assessment of NAFLD as systemic disease-a machine learning perspective. PLoS One. 2019;14:e0214436.
15. Fialoke S, Malarstig A, Miller MR, Dumitriu A. Application of machine learning methods to predict non-alcoholic steatohepatitis (NASH) in non-alcoholic fatty liver (NAFL) patients. AMIA Annu Symp Proc. 2018;2018:430–9.
16. Perakakis N, Polyzos SA, Yazdani A, et al. Non-invasive diagnosis of non-alcoholic steatohepatitis and fibrosis with the use of omics and supervised learning: a proof of concept study. Metabolism. 2019;101:154005.
17. Chiappini F, Coilly A, Kadar H, et al. Metabolism dysregulation induces a specific lipid signature of nonalcoholic steatohepatitis in patients. Sci Rep. 2017;7:46658.
18. Dai G, Liu P, Li X, Zhou X, He S. Association between PNPLA3 rs738409 polymorphism and nonalcoholic fatty liver disease (NAFLD) susceptibility and severity: A meta-analysis. Medicine (Baltimore). 2019;98:e14324.
19. Vespasiani-Gentilucci U, Gallo P, Porcari A, et al. The PNPLA3 rs738409 C>G polymorphism is associated with the risk of progression to cirrhosis in NAFLD patients. Scand J Gastroenterol. 2016;51:967–73.
20. Hotta K, Yoneda M, Hyogo H, et al. Association of the rs738409 polymorphism in PNPLA3 with liver damage and the development of nonalcoholic fatty liver disease. BMC Med Genet. 2010;11:172.
21. Liu YL, Patman GL, Leathart JB, et al. Carriage of the PNPLA3 rs738409 C>G polymorphism confers an increased risk of non-alcoholic fatty liver disease associated hepatocellular carcinoma. J Hepatol. 2014;61:75–81.
22. Tan HL, Zain SM, Mohamed R, et al. Association of glucokinase regulatory gene polymorphisms with risk and severity of non-alcoholic fatty liver disease: an interaction study with adiponutrin gene. J Gastroenterol. 2013;49:1056–64.

23. Ulusoy G, Arinç E, Adali O. Genotype and allele frequencies of polymorphic CYP2E1 in the Turkish population. Arch Toxicol. 2007;81:711–8.
24. Matsushita N, Hassanein MT, Martinez-Clemente M, et al. Gender difference in NASH susceptibility: roles of hepatocyte Ikkβ and Sult1e1. PLoS One. 2017;12:e0181052.
25. Noureddin M, Vipani A, Bresee C, 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;113:1649–59.
26. Hashimoto E, Tokushige K. Prevalence, gender, ethnic variations, and prognosis of NASH. J Gastroenterol. 2011;46 Suppl 1:63–9.
27. Soleymani R, Granger E, Fumera G. F-measure curves: a tool to visualize classifier performance under imbalance. Pattern Recognition. 2020;107146:107146.