Progress in Artificial Intelligence
https://doi.org/10.1007/s13748-025-00387-8

REGULAR PAPER

On the role of explanations in machine learning prediction of nonalcoholic steatohepatitis

Nikolay Babakov[1] · Elena Rezgova[2] · Ehud Reiter[3] · Alberto Bugarín-Diz[1]

Received: 8 August 2024 / Accepted: 13 July 2025
© The Author(s) 2025

Abstract

Nonalcoholic steatohepatitis (NASH) is a common disease that ultimately can lead to the development of end-stage liver disease, cirrhosis, or hepatocellular carcinoma. An early prediction of NASH provides an opportunity to make an appropriate strategy for prevention, early diagnosis, and treatment. The most accurate approach for NASH diagnostics is a liver biopsy, which can lead to various complications for the patient. Many papers have studied non-invasive machine learning (ML)-driven approaches to early non-invasive NASH prediction; however, to the best of our knowledge, none of the works considered the problem of explainability of the trained ML models to the medical experts. In this work, we address this issue. We use the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) nonalcoholic fatty liver disease adult database to train different ML models and propose the technique to explain their predictions. We compare the explanations obtained from a transparent model (Decision Tree) and a non-transparent model (Random Forest). Furthermore, we analyze the quality of explanation prediction by objective means and with a user study involving 11 medical practitioners. Our findings show that there is no significant difference in the perception of explanation obtained from transparent and non-transparent models, and that the explanation of the models' predictions slightly increases their usability and trustworthiness for real practitioners, enhancing their practical adoption in clinical settings.

Keywords Explainable Artificial Intelligence · Natural Language Generation · Textual explanations · Explainable Machine Learning · Non-Alcoholic Fatty Liver Disease (NAFLD) · NonAlcoholic SteatoHepatitis (NASH)

1 Introduction

Non-Alcoholic Fatty Liver Disease (NAFLD) is a leading cause of chronic liver diseases around the world [1]. NAFLD is strongly associated with metabolic syndrome and is considered the hepatic manifestation of the metabolic syndrome [2]. It can manifest as pure fatty liver disease (hepato-steatosis) or as non-alcoholic steatohepatitis (NASH), an evolution of the former in which steatosis is associated with inflammation and hepatocellular damage and with fibrogenic activation that can lead to cirrhosis and the onset of hepatocarcinoma [3]. In its ultimate phase, NASH can lead to the development of end-stage liver disease, cirrhosis, or hepatocellular carcinoma (HCC) [4]. However, NASH remains significantly underdiagnosed because of the lack of specific clinical symptoms, low awareness among patients, and the lack of treatments explicitly approved for NASH. Identifying patients with a high probability of NASH is the first step toward risk stratification and diagnosis for disease management.

The early diagnosis of NASH and elimination of its cause can stop further liver damage, increase the chances of transplant success, and also reduce mortality rates [5]. The most accurate way to detect NASH is a liver biopsy. However, it is an invasive method that can be risky and can lead to various complications [6]. This naturally creates the need for non-invasive testing [7].

There are already many works aiming to apply Artificial Intelligence (AI) techniques to diagnose NASH, relying on image [8–10] or numerical data [11–14]. In our work, we consider numerical data obtained from the results of patient

These authors contributed equally to this work.

✉ Nikolay Babakov
nikolay.babakov@usc.es

[1] Centro Singular de Investigacion en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain

[2] Health Technologies LLC, Saint Petersburg, Russian Federation

[3] University of Aberdeen, Aberdeen, United Kingdom

Published online: 25 July 2025

Table 1 Comparison of our work with the existing works about NASH diagnosis. "Global explanation" indicates whether the trained model predictions were explained and analyzed on a global level. The data or model is assumed to be available only if it is open-sourced or can be requested clearly. If it is available only by requesting the authors it is labeled as "AR". "User study" indicates whether the predictions of the trained model were assessed by medical practitioners.

Paper	Cohort size	Data available	Global explanation	Model available	User study
[14]	3,000	-	+	-	-
[13]	10,000	AR	+	-	-
[15]	500	-	-	-	-
[16]	26,000	+	+	-	-
[17]	100	+	+	-	-
[11]	700	+	+	-	-
[12]	100	-	-	-	-
[18]	200	-	+	+	-
[19]	600	-	-	-	-
[20]	5,156	+	+	AR	-
[21]	141,293	-	+	-	-
Our work	600	+	+	+	+

examinations. Refer to Table 1, where we compare our work with the existing ones.

Even though AI-driven non-invasive NASH diagnosis has been widely studied, most of the existing papers and the AI models presented in them have certain shortcomings. First, it is pretty common that NASH-related studies are performed on closed-source datasets that have to be either requested from the authors of the study or impossible to request at all, which makes the results of such studies unreproducible. Second, the models trained in such studies are rarely made open-sourced. This problem is related not only to the AI models applied to NASH but also to many medical AI papers [22], which significantly hinders their adaptability in real practice [23]. The most important drawback is that the explainability of the models trained in such works is not verified with real medical practitioners (see Table 1). In some works, global explainability analysis (which studies the significance of different features for the predictions) of the trained models is considered; however, to the best of our knowledge, none of the existing works dedicated to NASH diagnosis engaged medical experts in the examination of the explainability of such models in their real practice. The lack of verification of explainable AI (XAI) techniques in medicine by practitioners is also applicable not only to NASH but to all medical XAI [24, 25].

It is generally known that the final decision about a patient is up to a human expert, who cannot blindly rely on AI-model prediction [26], which raises the need for XAI methods [27]. An AI model may only be treated as a clinical decision support system, which must be capable of providing clinicians with the knowledge to enhance medical decision-making [28]. Still, the experts may resist using a system if it does not provide a relevant explanation of the reasons behind its decision or capture the nuances of human thinking [29–31]. Thus, in our work, we address the aforementioned drawbacks and set the following research questions:

- RQ1. Does the transparency of a model increase the acceptability of the model to medical professionals?
- RQ2. Does the explanation of the AI model's prediction increase the acceptability of the model to medical professionals?

To answer these questions, we use the NIDDK dataset[1], a high-quality clinical resource collected by the US National Institute of Diabetes and Digestive and Kidney Diseases. While this dataset is not open-access, it has a clearly defined process for academic use, making it a reliable benchmark for medical AI research.

Unlike most previous studies on AI-based NASH prediction-which often overlook the explainability of model decisions or fail to involve real clinicians in the evaluation process-our work explicitly focuses on how explanations are perceived by medical professionals and whether they contribute to the practical usability of AI predictions. We propose a systematic methodology to assess this through both quantitative analysis and a dedicated user study.

Our main contributions are as follows:

- We design and implement a method to explain NASH predictions from both a transparent model (Decision Tree) and a non-transparent model (Random Forest), using SHAP-based visual and verbal explanations tailored for clinical relevance.
- We conduct a detailed analysis of explanation quality through both user-agnostic metrics and a user study with 11 medical practitioners experienced in NASH diagnosis.

[1] https://repository.niddk.nih.gov/home/

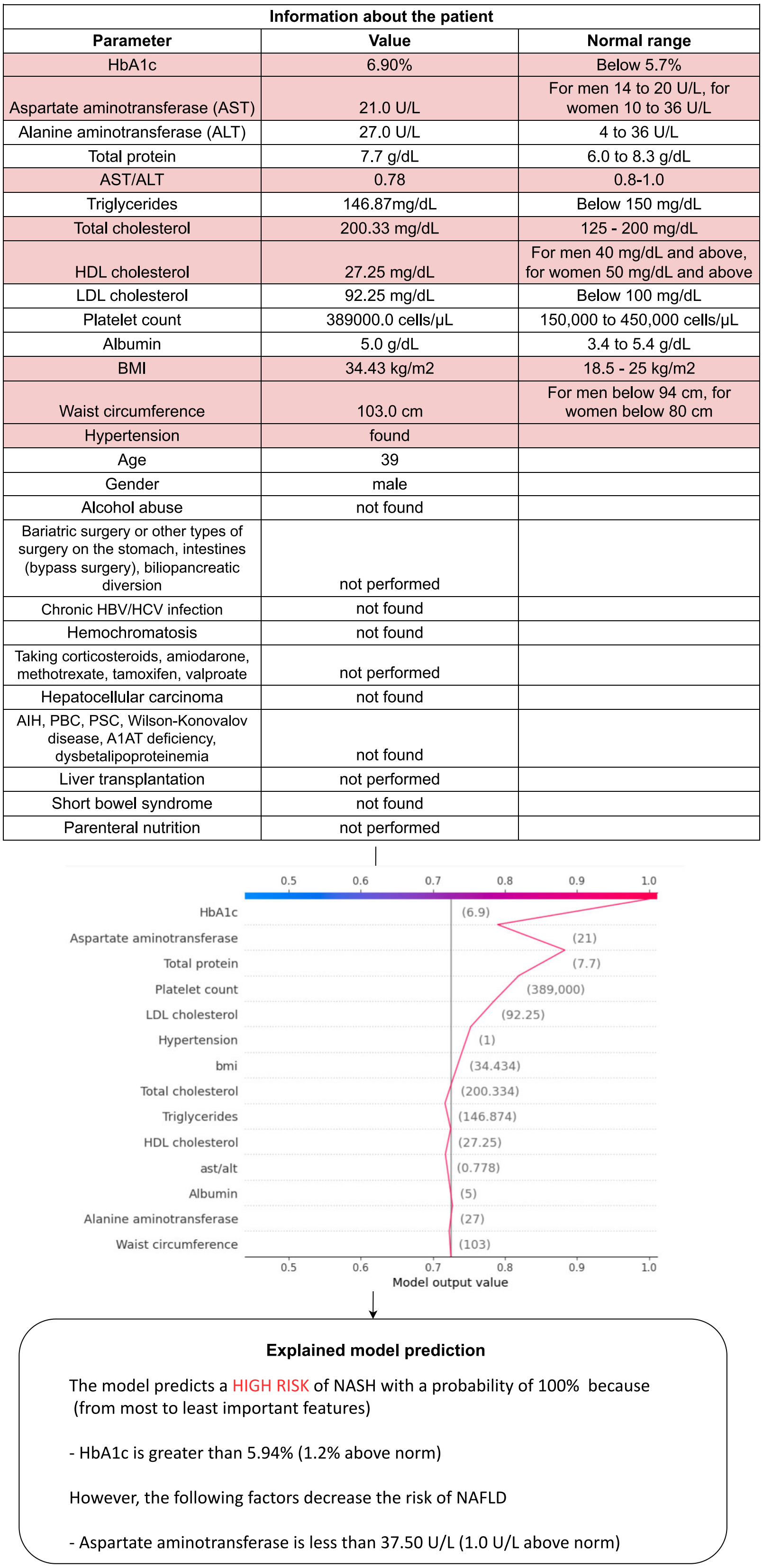


Fig. 1 Decision Tree explanation example. The explanation interface consists of three parts. **Information about the patient:** table containing two groups of features. The first group (from HbA1c to Hypertension) shows the features used by the model for predictions. The second group (from Age to Parenteral nutrition) indicates that the patient is related to a certain cohort which implies that the patient does not have certain diseases, does not take certain medicines, etc (that is why most of the features in this group have "not found" value). **Shapley values-based decision plot:** the plot showing how the model has arrived at the conclusion, where the features in the top part play a more significant role and the direction of the line indicates whether the feature increased or decreased the risk of NASH. **Verbal explanation of the model's decision:** Verbal explanation of the model's decision containing the features increased and decreased the risk of NASH. The exact form of the verbal explanation varies w.r.t. the transparency of the model.

Information about the patient

Parameter	Value	Normal range
HbA1c	6.90%	Below 5.7%
Aspartate aminotransferase (AST)	21.0 U/L	For men 14 to 20 U/L, for women 10 to 36 U/L
Alanine aminotransferase (ALT)	27.0 U/L	4 to 36 U/L
Total protein	7.7 g/dL	6.0 to 8.3 g/dL
AST/ALT	0.78	0.8-1.0
Triglycerides	146.87mg/dL	Below 150 mg/dL
Total cholesterol	200.33 mg/dL	125 - 200 mg/dL
HDL cholesterol	27.25 mg/dL	For men 40 mg/dL and above, for women 50 mg/dL and above
LDL cholesterol	92.25 mg/dL	Below 100 mg/dL
Platelet count	389000.0 cells/μL	150,000 to 450,000 cells/μL
Albumin	5.0 g/dL	3.4 to 5.4 g/dL
BMI	34.43 kg/m2	18.5 - 25 kg/m2
Waist circumference	103.0 cm	For men below 94 cm, for women below 80 cm
Hypertension	found	
Age	39	
Gender	male	
Alcohol abuse	not found	
Bariatric surgery or other types of surgery on the stomach, intestines (bypass surgery), biliopancreatic diversion	not performed	
Chronic HBV/HCV infection	not found	
Hemochromatosis	not found	
Taking corticosteroids, amiodarone, methotrexate, tamoxifen, valproate	not performed	
Hepatocellular carcinoma	not found	
AIH, PBC, PSC, Wilson-Konovalov disease, A1AT deficiency, dysbetalipoproteinemia	not found	
Liver transplantation	not performed	
Short bowel syndrome	not found	
Parenteral nutrition	not performed	

Explained model prediction

The model predicts a HIGH RISK of NASH with a probability of 100% because (from most to least important features)

- HbA1c is greater than 5.94% (1.2% above norm)

However, the following factors decrease the risk of NAFLD

- Aspartate aminotransferase is less than 37.50 U/L (1.0 U/L above norm)

- We empirically investigate whether model transparency or explanation alone increases clinicians' trust and willingness to rely on AI predictions in a realistic diagnostic context.
- We release the trained models and explanation code to the research community to encourage reproducibility and reuse in future clinical AI studies.

2 Related literature

AI is widely applied in various fields of healthcare, such as health services management, predictive medicine, clinical decision-making, and patient data and diagnostics (a lot of application examples are listed in [32]). In the field of medicine, where risks are high and responsibilities are substantial, AI models are primarily employed as decision support systems. Ultimately, the final decision rests with a human medical expert [33]. This naturally raises the interest in XAI methods [27].

Several XAI techniques have been developed to interpret model predictions in different domains. SHAP [34] assigns each feature an importance score based on its marginal contribution to the prediction, using principles from cooperative game theory; it provides consistent, additive explanations that are both local and global. LIME [35] approximates the model's behavior locally around a specific prediction by fitting a simple interpretable model (usually linear) to perturbed samples near the input point. Grad-CAM [36] is tailored for convolutional neural networks and highlights the most influential regions of an input image by computing the gradients of the prediction with respect to spatial feature maps.

XAI methods may be roughly divided into global and local [37]. Global explanation allows us to determine the extent to which each feature contributes to the model's decisions across all predictions, whereas local explanations focus on a single prediction and are useful for understanding specific model outputs. In the majority of medical papers, Shapley-values-based (SHAP) [34] explanation techniques are used [27]. This is natural because SHAP has a well-maintained open-sourced implementation[2], and understanding the generated explanation requires only a basic knowledge of the Shapley values concept. However, most medical papers apply only global explainability analysis to the trained models [38–42] and do not analyze the predictions on the local level. The same is applicable to the papers applying ML to NASH prediction: XAI techniques are used either only for global explanation of the models [11, 13, 14, 16–18] or not used at all [12, 15, 19].

This seems pretty critical because, as we mentioned above, the most natural role of AI models in medicine is decision support, which is applied to specific cases by the human practitioner. That is why verification of the usability of the local explanation methods seems crucial to the integration of XAI into real practice. However, the process of XAI method evaluation is also a complicated area that deserves discussion.

In one of the first papers on XAI evaluation [43] four key perspectives of evaluation were proposed: goodness, user satisfaction, comprehension, and the impact on expert performance. Still, the evaluation is based not only on the right questions for participants, but also on the task demonstrated before the questions are asked. [44] showed that the XAI evaluation setup must be based on the real decision task rather than on artificial proxy tasks. Moreover, apart from collecting feedback from experts, there are still several ways to evaluate the XAI techniques objectively in a user-agnostic manner [45].

XAI evaluation in medicine has also been studied from different perspectives [46]. [47] delves into the critical consideration of various XAI failures and their potential implications on decision-making for individual patients. The most relevant medical domain paper that served as a motivation for the user study setup is [31], which is specifically dedicated to collecting feedback from medical experts regarding their expectations from XAI. In this work, it was shown that the following factors play a crucial importance in the usability of XAI techniques in the decision-making process: feature importance, level of a model's uncertainty, and transparent design of the model.

3 Materials and Methods

In this section, we describe the data used for our experiments, the process of selecting the patient cohort relevant to our study, data pre-processing techniques, and feature selection strategies. We also present the machine learning models trained on the processed data and explain the methodology used to generate interpretable predictions for the user study. An overview of the entire experimental pipeline-from cohort selection to explanation assessment by clinicians-is illustrated in Figure 2, and an example of the explanation interface is shown in Figure 1.

3.1 Dataset

We use the dataset obtained from The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), which is based on the 5-year study of 1441 patients, 1066 of whom have from one to three NASH or non-NASH diagnoses confirmed through liver biopsy and histological assessment. The dataset has a complex structure; it consists of 18 files,

[2] https://shap.readthedocs.io/en/latest/, retrieved on February 13th, 2024

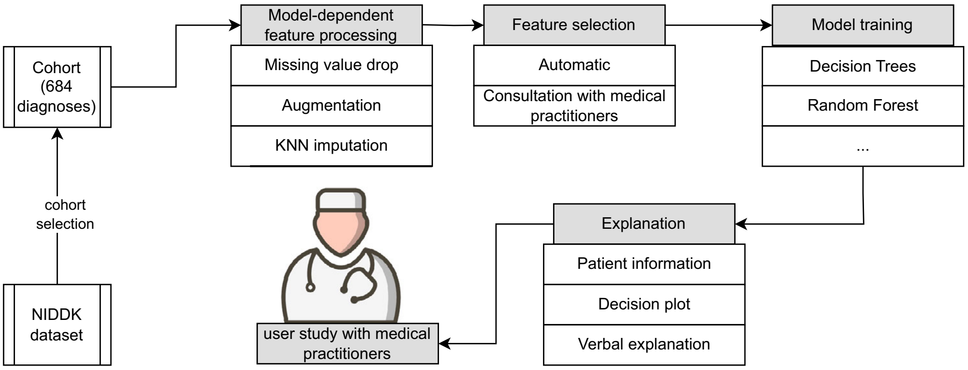


Fig. 2 Overview of the experimental pipeline used in this study. The process begins with cohort selection from the NIDDK dataset, resulting in 684 biopsy-confirmed diagnoses. Pre-processed data is generated using model-dependent pipelines, which include options such as missing value dropping, KNN-based imputation, and data augmentation. Feature selection is carried out using a combination of automatic techniques and input from medical practitioners. Trained models (including a transparent Decision Tree and a non-transparent Random Forest) are then explained using a unified interface that includes patient information, SHAP-based decision plots, and verbal explanations. These explained predictions are presented to clinicians in a user study to evaluate their perceived trustworthiness and usefulness.

each corresponding to the form filled out either by a patient or practitioner during initial screening or follow-up visits that take place five times every 24 weeks. Below is a list of the forms that play the most significant role in our study:

- *Central Histology Review.* Record results of the NASH Pathology Committee review of liver biopsy slides. This form contains the decision of the Committee answering the question "Is this steatohepatitis?" with three possible answers: "No", "Suspicious", and "Yes, definite".
- *Laboratory results.* The form contains various laboratory results taken during the study; it includes such important tests as HbA1c, AST, ALT, and other factors having a direct relation to NASH.
- *Physical examination.* Record-focused physical exam findings, which may be used as a source for such important patient properties as BMI and waist circumference,.
- *Alcohol Use Disorders Identification Test.* A self-administered form containing general information about a patient's alcohol-consuming habits is necessary for further considering the patient's suitability for the NASH study.

Other forms are related to either general information (initial registration, medical background, physical activity, etc.) or other observations of the patient (liver imaging studies, symptoms of liver disease).

3.2 Cohort selection

We consider only adult patients with at least one biopsy-confirmed NASH or non-NASH diagnosis and also with the results of alcohol use disorder tests applicable to non-alcoholic steatohepatitis. We also exclude patients who have other forms of chronic liver disease (e.g., hepatocellular carcinoma), have undergone bariatric surgery (e.g., stapling or banding of the stomach), or take certain medicines likely to interfere with our study (e.g., corticosteroids). After applying all inclusion and exclusion criteria, we get 582 patients with corresponding 684 biopsy-confirmed NASH diagnoses (some patients have several diagnoses). 495 of the diagnoses are either certain or borderline NASH. The whole list of inclusion and exclusion criteria with the sequence of applications and the number of dropped patients corresponding to each step is available in Appendix A Table 5.

3.3 Data pre-processing

As mentioned in Section 3.1, the NIDDK dataset consists of multiple files, so preparing the training-ready dataset requires certain pre-processing. To create fixed-size samples, we use NASH diagnoses as a target variable, and according to the corresponding histology review, we collect the latest data available to the NASH Pathology Committee by the date of the review. After the fixed-size samples are prepared, we apply the following data pre-processing techniques:

- *Missing values processing.* 70% of the values from all the data files are missing. This requires certain steps to handle the missing values. We consider two approaches. First, we simply drop the features with a percentage of non-available values above a certain threshold. Alternatively, we use K-nearest neighbor imputation based on 10 neighbors.
- *Feature processing.* We normalize continuous values and do not apply pre-processing to integer values.
- *Target variable.* Originally, the target variable consisted of four states. Two of them mean either a "certain yes" or a "certain no" NASH diagnosis. Two other states are used for certain degrees of borderline decisions. Such

a granular decision could be made if the results of the biopsy are available. However, in this study, we develop a model that relies on the data obtained by non-invasive methods. Thus, to make the training task feasible, we interpret all borderline diagnoses as positive NASH, so the target variable takes binary form.

- *Augmentation*. After cohort selection, the obtained dataset turns out to be imbalanced, and since the overall number of samples capable of training the model is pretty low, we apply the well-known augmentation technique. It is done by nullifying existing values and further imputing them with a KNN algorithm, which results in "new" samples in a dataset. To prevent nullification of any data from samples with too many missing values, we consider only such samples that have more than half of the non-missing values. We add 5 "new" samples to the data if the element has positive NASH and 10 "new" ones otherwise.
- *Train-test splitting*. We apply stratified 5-fold cross-validation to ensure that the proportion of positive and negative target values is preserved across each fold. All augmented samples are excluded from test splits to avoid data leakage.

We always use feature processing, train-test splitting, and target variable pre-processing steps. Other approaches (missing value processing and augmentation) serve as hyperparameters in our pre-processing pipeline. We experiment with various combinations of these approaches and ultimately report the most effective one for each specific model.

We also acknowledge that while KNN imputation and data augmentation allow us to make better use of limited clinical data, they also introduce potential limitations. Imputed values may not fully capture the variability of true patient data, and synthetic samples created through augmentation could introduce biases if overused. However, given the small sample size after cohort filtering, these methods were necessary to enhance model robustness. To mitigate their impact, augmented samples were excluded from evaluation splits during cross-validation.

3.4 Feature selection

Our study focuses on enhancing model explainability, and within data pre-processing, feature selection plays a crucial role. This is because, even if a feature set yields high metrics for the trained model, it may still prove impractical for real-world practitioners. So the features should be suitable for a demonstration to the practitioners and, at the same time, should result in proper metrics for the model.

As an initial reference, we considered the feature set proposed in [11], as it achieved strong performance and was aligned with medical reasoning. However, we did not rely on it exclusively. In our study, we applied several standard feature selection techniques-including recursive feature elimination, removal of low-variance features, and univariate feature selection-both using the referenced set as a starting point and independently, in order to assess the robustness and relevance of features from multiple perspectives. The experiments with such techniques gave us an initial feature set. However, after preliminary consultations with a group of medical experts, we slightly adjusted these features, taking into account their feedback. First, we remove white blood cell count and hematocrit because they do not look useful for real diagnostics, and height because it is encoded in BMI and does not yield significant information alone. Second, we do not consider the gender of the patient as a standalone feature for model training; however, we keep it for demonstration during the explanation because the normal ranges of some analyses vary depending on the gender (e.g., AST). Finally, we add cholesterol (total, HDL, and LDL) and waist circumference. The final feature set and the corresponding data statistics are shown in Table 2.

3.5 Machine-learning models

We consider various ML models. First, we use such baseline methods as Logistic Regression (LR), support vector machines (SVM), naïve Bayes (NB), k-nearest neighbor (KNN), and Decision Trees (DT). The more sophisticated models included in our experiments are Random Forest (RF) [48], Bayesian Networks (BN) [49], XGBoost (XGB) [50], CatBoost [51], and MLP [52].

3.6 Explanation technique

In our study, we experimented with explaining transparent and non-transparent models. The main motivation for this model selection is that transparent models are naturally explainable, but as a rule have less complex structure than non-transparent models, which may result in lower performance compared to more complex non-transparent models. Thus, we study the usability of more transparent but less accurate models applied to the specific task of NASH prediction.

For transparent models, we select the Decision Trees, because they can be naturally explained [53, 54] which have been widely used in medical practice [55–57]. We also use a non-transparent model, because normally they have a more complex structure that yields better performance on unseen data. For both types of models, the general explanation interface consists of three main parts: information about the patient, a Shapley values-based decision plot, and a verbal explanation of the model's decision. Refer to Figure 1 for the DT explanation interface and C Figure 8 for non-transparent explanation interface examples.

Table 2 Statistics of the data selected for ML model training: mean and standard deviation of NASH and non-NASH samples, and null values' fraction. For Hypertension, only a percentage of patients with the found hypertension is reported. p-value is calculated with an independent t-test between the NASH and non-NASH groups of samples.

	NASH (N=495)	Non-NASH (N=189)	Empty, %	p-value
HbA1c (%)	6.41 ± 1.43	5.84 ± 1.13	4	<0.001
Aspartate aminotransferase, AST (U/L)	50.66 ± 32.23	39.49 ± 25.94	1	<0.001
Alanine aminotransferase, ALT (U/L)	66.83 ± 51.95	58.04 ± 51.76	1	0.049
Total protein (g/dL)	7.32 ± 0.6	7.14 ± 0.55	1	0
AST/ALT	0.87 ± 0.35	0.86 ± 0.49	1	0.796
Triglycerides (mg/dL)	179.98 ± 137.89	163.96 ± 128.8	1	0.17
Total cholesterol (mg/dL)	189.49 ± 43.72	198.15 ± 39.27	1	0.018
HDL cholesterol (mg/dL)	44.22 ± 12.11	45.81 ± 11.78	1	0.128
LDL cholesterol (mg/dL)	113.08 ± 36.14	123.92 ± 33.56	4	0.001
Platelet count (cells/mcL)	234183.64 ± 77844.49	245244.68 ± 77734.5	1	0.099
Albumin (g/dL)	4.25 ± 0.47	4.17 ± 0.42	1	0.054
BMI (kg/m2)	33.35 ± 6.22	33.41 ± 6.14	56	0.939
Waist circumference (cm)	106.87 ± 13.77	106.69 ± 14.28	57	0.922
Hypertension	1.0 ± 0.0	1.0 ± 0.0	48	

3.6.1 Information about the patient

In the upper part of the explanatory interface, we show the information about the patient corresponding to the selected features. Each parameter has a corresponding value and a normal range. In cases where the value of the parameter is out of the normal range, the corresponding line in a table is highlighted in red to make studying patient data more convenient for medical experts. Moreover, we show the information that clarifies that the patient has passed the exclusion criteria used in our study (age, alcohol consumption, other forms of chronic liver disease, etc.).

3.6.2 Shapley values-based decision plot

A wide range of techniques have been developed to explain the predictions of complex machine learning models, especially in high-stakes domains such as healthcare. Among the most widely used are LIME (Local Interpretable Model-agnostic Explanations) [35] and SHAP (SHapley Additive exPlanations) [34]. Both aim to make individual model decisions interpretable by attributing importance scores to input features.

LIME explains a model's prediction by approximating it locally with a simpler interpretable model-typically a sparse linear model-trained on randomly perturbed samples around the instance of interest. The idea is that small changes in the input should reveal how each feature influences the model's output. However, as shown in [35], LIME explanations can exhibit high variance across runs due to their reliance on random sampling. Moreover, since LIME is inherently local and fitted around a single point, it is not suitable for producing global explanations-i.e., summarizing feature importance across the entire dataset.

In contrast, SHAP is grounded in cooperative game theory and assigns each feature an importance value based on its contribution to the model's output, considering all possible combinations of features. The general intuition is that each feature value is treated as a "player" in a game where the model prediction is the "payout", and the Shapley value determines how fairly to attribute that payout among the features. This approach satisfies several desirable properties such as consistency and local accuracy.

Given the need for consistency in explanations, especially in a clinical setting where reproducibility is essential, and our intent to perform global analyses of model behavior beyond individual cases, we adopt SHAP for both individual and aggregate explanation of predictions. SHAP's theoretical grounding and stable outputs make it well-suited for this purpose.

To visualize the Shapley values-based feature importance, we use a decision plot that shows how complex models arrive at their prediction. It displays the average of the model's base values and shifts the SHAP values accordingly to accurately reproduce the model's predictions.

3.6.3 Verbal explanation of the model's decision

Even though the SHAP values decision plot may give initial intuition to a medical expert about how the model made the particular prediction, we cannot be absolutely sure that it will be correctly interpreted by the expert, especially if they are not familiar with the concept of SHAP and with the basic principles of AI models. Thus, we assume that the SHAP values visualization may serve only as a supplementary explanation tool, whereas the core explanation should be delivered in textual form.

The textual explanation consists of a clear identification of what exactly a model predicts and what its confidence is in the prediction. The decision of the model (i.e., "High risk" or "Low risk") is highlighted with red or green correspondingly to make the prediction text more readable. Then the factors increasing and decreasing the risk of NASH are shown to an expert (we use SHAP values to calculate that). The order of factors in the demonstration depends on the final prediction. So, if a model predicts high risk, we show the factors increasing the risk first and then the decreasing ones. In the case of low-risk prediction, we demonstrate them in the opposite order. Moreover, we include only the factors that are out of the normal range to the ones increasing the risk.

The exact way of verbalizing the factors varies depending on the type of model we explain. In the case of any non-transparent model explanation, we can only list the names of the corresponding factors and their values.

However, in the case of DT explanation, we may provide the expert with a deeper look at the decision-making process, as shown in Figure 3. As far as the nature of DTs provides the exact decision path of the model's prediction, we may verbalize it [58]. The only problem is that straightforward verbalization may not be readable because the decision path may have repetitive features. We take the idea similar to the one described in [59]. In this work, the authors verbalize the DT decision path by showing either one of the biggest or the smallest boundaries of the feature value, or the interval of the values. They rely on the dataset used for training a particular DT to calculate which features favor the final prediction and which ones are against it. As far as we demonstrate the SHAP-values-based decision plot, we rely on SHAP values to calculate which group the feature should be shown in (in favor or against the final prediction). Apart from dropping the features that have their value within the normal range from the group, increasing NASH risk, we also do not show the features that are not present in the DT decision path.

Finally, in both the DT and non-transparent models' explanations, we show the reminder of whether the value of this factor is outside the normal range. To prevent showing insignificant and unnecessary information for each group of factors, we calculate the sum of modules of SHAP values and verbalize only the ones whose SHAP value is above the manually selected 5% threshold of this sum.

3.7 User study design

To the best of our knowledge, there is no unified way to evaluate XAI techniques, either in medical or any other applications. Thus, we design our user study by joining the best existing XAI evaluation practices (discussed in Section 2) and our study's peculiarities.

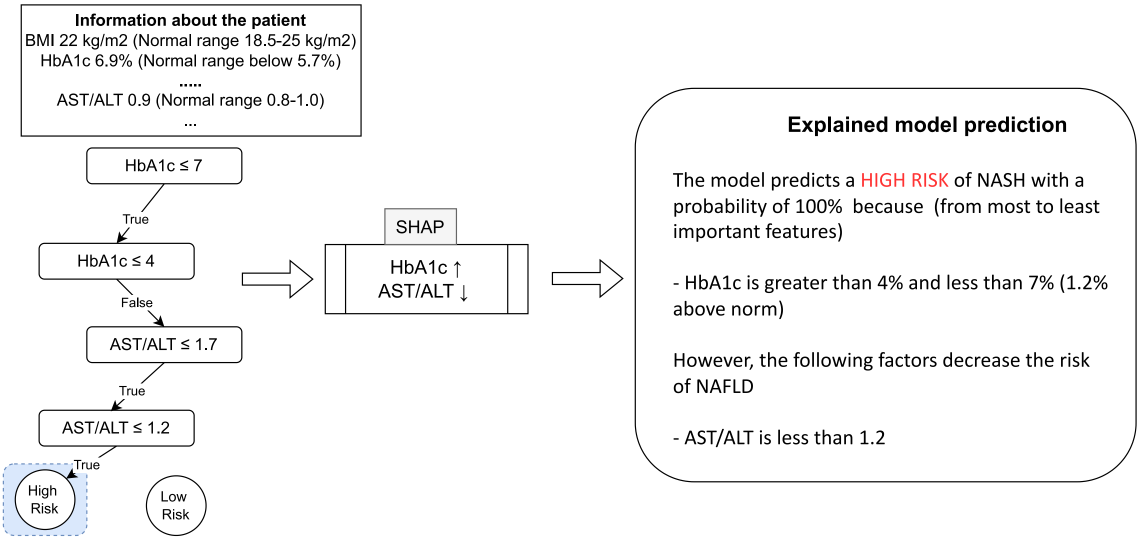
The case of each patient is shown on two pages. On the first page, the patient information is shown with the verbalization of the model's prediction without any explanation (see C Figure 9). On this page, we only ask the expert one question: "What is their diagnosis or solution for this patient (NASH, non-NASH, impossible to determine, or other)?" This approach ensures that the doctor engages with the model's predictions in a manner consistent with real-world practice.

On the second page, we show the patient information accompanied by an explained prediction similar to the one shown in Figure 1 or C Figure 8. After that, we ask about the diagnosis again. We also show the following statements related to personal perception of the explanation and ask the participants to rate them using a 5-point Likert scale with answers that vary from "Strongly agree" to "Strongly disagree" (see D Figure 11):

- Explanation makes sense from the medical knowledge point of view
- Explanation increases the trustworthiness of the model's prediction
- Explanation is easy to understand
- Explanation helps me assess whether the prediction is correct
- I would consider this explanation when making decisions about real patients

As mentioned in Section 3.6, we use DT and a best-performing non-transparent model (the specific model will be indicated in Section 4.1) for predictions. We take three samples from the NIDDK dataset for this user study. These samples are extracted from train and test splits. Moreover, we ensure that the patient corresponding to this sample has only one histology review result associated with it to prevent data contamination. Additionally, we use two patients from the real practice of one of the authors. Overall, two selected cases have negative NASH, and three of them have positive. We ensure that all selected models make correct predictions for all the selected cases because this study aims to evaluate the explanations and, therefore, it only makes sense to explain correct predictions. An expert can most probably recognize an incorrect prediction, which could yield a negative bias in scoring the quality of explanations.

The experts are engaged from the professional network of one of the authors. Before the expert starts participating in the user study, we ask for their informed consent to participation (see B Figure 7). Then we ask them about their medical specialty, years of active experience, frequency of



Information about the patient
BMI 22 kg/m2 (Normal range 18.5-25 kg/m2)
HbA1c 6.9% (Normal range below 5.7%)
.....
AST/ALT 0.9 (Normal range 0.8-1.0)
...

Explained model prediction

The model predicts a HIGH RISK of NASH with a probability of 100% because (from most to least important features)

- HbA1c is greater than 4% and less than 7% (1.2% above norm)

However, the following factors decrease the risk of NAFLD

- AST/ALT is less than 1.2

Fig. 3 The illustration of the approach for Decision Tree path verbalization. In this example, the HbA1c feature passed two decision steps with borders 4 and 7, and as far as its value is between these borders it is verbalized as "greater than 4% and less than 7%". AST/ALT passed two decision steps, and it turned out to be less than both engaged border values, so it is demonstrated with the least of these values - "less than 1.2". The features not engaged in the DT decision path are not verbalized. SHAP-values are used to decide whether to show the verbalized feature as increasing or decreasing the NASH risk.

work with patients with NASH, and their common practices for making a NASH diagnosis (see D Figure 10). Potential participants who indicate that they have never worked with NASH patients are automatically excluded from the study. We also use the manual input about their practices in NASH diagnostics to make sure that the participant possesses the minimal necessary knowledge about NASH. Finally, we track the time of page completion of each page presented to a participant to further control the quality of the replies.

To make sure that the participants provide relevant answers that do not have any sequence-related bias, we shuffle the cases and present them to the participant in a random order. Moreover, we ensure that each patient can be shown only once to one participant (with either DT or non-transparent explained prediction).

The target audience for the user study is qualified medical professionals with little time available. To avoid such problems as tiredness or lack of motivation for long questionnaires, we aim to show only four cases in total to each participant. However, as it is difficult to predict how much cognitive load will be required for each case, after two cases are shown, we ask the participant whether they want to go on the study or they would like to quit it. This is supposed to prevent collecting irrelevant answers from experts who lose the motivation to participate due to excessive study time.

4 Results

4.1 Trained models performance

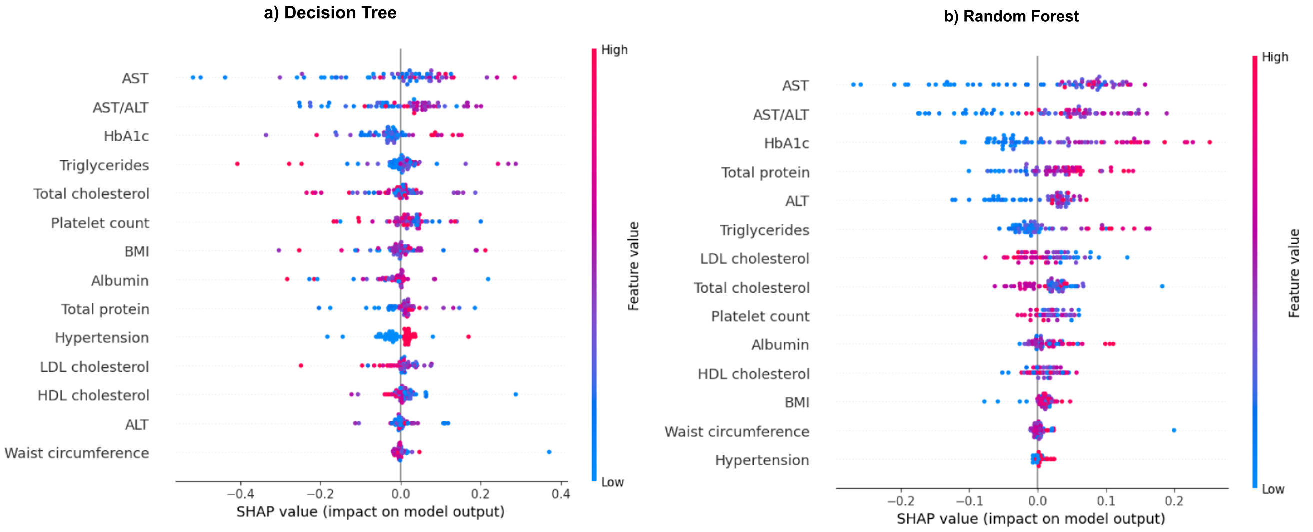
The selected ML models are trained and tested on a 5-fold train-test split. We apply grid search for both model and data pre-processing hyperparameters. In Table 3 we report the type of model and its corresponding best metrics. The precise hyperparameters resulting in the corresponding metrics are reported in E Table 6. RF performed best of all according to most of the metrics. So we use it for further explainability experiments.

We could also consider BNs as a possible candidate for explanation. BNs are sometimes referred to as gray-box models [60] because their internal structure can be inspected, but generating clear explanations is not very straightforward. There are several BN explanation methods that may be interesting to be tested in a user study similar to ours [61, 62]. However, the BNs trained with the selected dataset do not show proper metrics, and what is more important, the structure learned from the cohort did not make much medical sense, so we finally decided not to consider them in our user study. The most probable reason for the poor quality of structure learned from the data is its limited amount, which is known to be a significant obstacle for many structure learning algorithms [63].

We analyze the feature importance of DT and RF models using SHAP values in Figure 4. On these plots, the higher the feature is placed on the Y-axis, the more importance it has. Moreover, the colors of the features encode their value

Table 3 Model performance on 5-fold train-test splits of NIDDK dataset.

Model	F-score	Accuracy	Precision	Recall	Specificity	AUC
RF	**0.69 ± 0.05**	**0.75 ± 0.03**	**0.69 ± 0.04**	**0.69 ± 0.06**	0.55 ± 0.13	0.72 ± 0.04
XGB	0.67 ± 0.05	0.74 ± 0.04	0.68 ± 0.05	0.66 ± 0.05	0.48 ± 0.08	**0.73 ± 0.04**
CatBoost	0.66 ± 0.02	0.72 ± 0.02	0.67 ± 0.01	0.65 ± 0.03	0.48 ± 0.02	0.70 ± 0.07
MLP	0.66 ± 0.03	0.74 ± 0.03	0.68 ± 0.04	0.66 ± 0.01	0.49 ± 0.05	0.71 ± 0.03
NB	0.65 ± 0.05	0.71 ± 0.04	0.65 ± 0.05	0.65 ± 0.05	0.53 ± 0.06	0.70 ± 0.04
SVM	0.65 ± 0.03	0.71 ± 0.03	0.65 ± 0.03	0.65 ± 0.03	0.52 ± 0.07	0.65 ± 0.03
KNN	0.64 ± 0.02	0.70 ± 0.02	0.64 ± 0.02	0.66 ± 0.02	**0.57 ± 0.07**	0.68 ± 0.03
BN	0.64 ± 0.06	0.69 ± 0.05	0.64 ± 0.06	0.65 ± 0.06	0.55 ± 0.12	0.68 ± 0.06
DT	0.64 ± 0.01	0.68 ± 0.02	0.63 ± 0.01	0.65 ± 0.01	0.57 ± 0.04	0.64 ± 0.03
LR	0.60 ± 0.03	0.70 ± 0.01	0.61 ± 0.02	0.60 ± 0.03	0.36 ± 0.07	0.69 ± 0.05



a) Decision Tree

b) Random Forest

Fig. 4 Shapley plots for Decision Tree and Random Forest models. The higher the feature is placed on the Y-axis, the more importance it has. The colors of the features encode their value (light corresponds to low, and dark to high), and the position relative to the X-axis encodes whether the corresponding feature increases (when shifted to the right) or decreases (when shifted to the left) the model output.

(light corresponds to low, and dark to high), and the position relative to the X-axis encodes whether the corresponding feature increases (when shifted to the right) or decreases (when shifted to the left) the model output.

In the case of DT, we can see that hypertension has a clear impact on the increase in NASH risk; however, this impact is not among the most significant ones. Other factors do not have clear dependence, so higher and lower values of the corresponding feature may lead to both high-risk and low-risk predictions. In the case of RF, we can see more clusters that seem explicit: increased values of AST, AST/ALT, total protein, ALT, BMI, and hypertension yield “High risk” predictions of RF.

4.2 User-agnostic analysis of explanations

When we select a certain XAI approach, we may expect some tradeoff between the transparency of the model’s design and its accuracy, which however not always take place [64]. In Table 3 we can see that DT, which is transparent by its nature, performs worse than non-transparent RF and XGB, which can be explained only on the feature-importance level, like all other non-transparent models. Thus, it seems natural to analyze whether the transparency of DT is valuable enough to sacrifice the performance compared to the more sophisticated non-transparent models. One of the possible approaches to this analysis could be a human-agnostic analysis [65] of the sensibility of the explanation from the medical knowledge point of view.

Refer to Figure 5, which visualizes how well the borders verbalized during the explanation correlate with the experts’

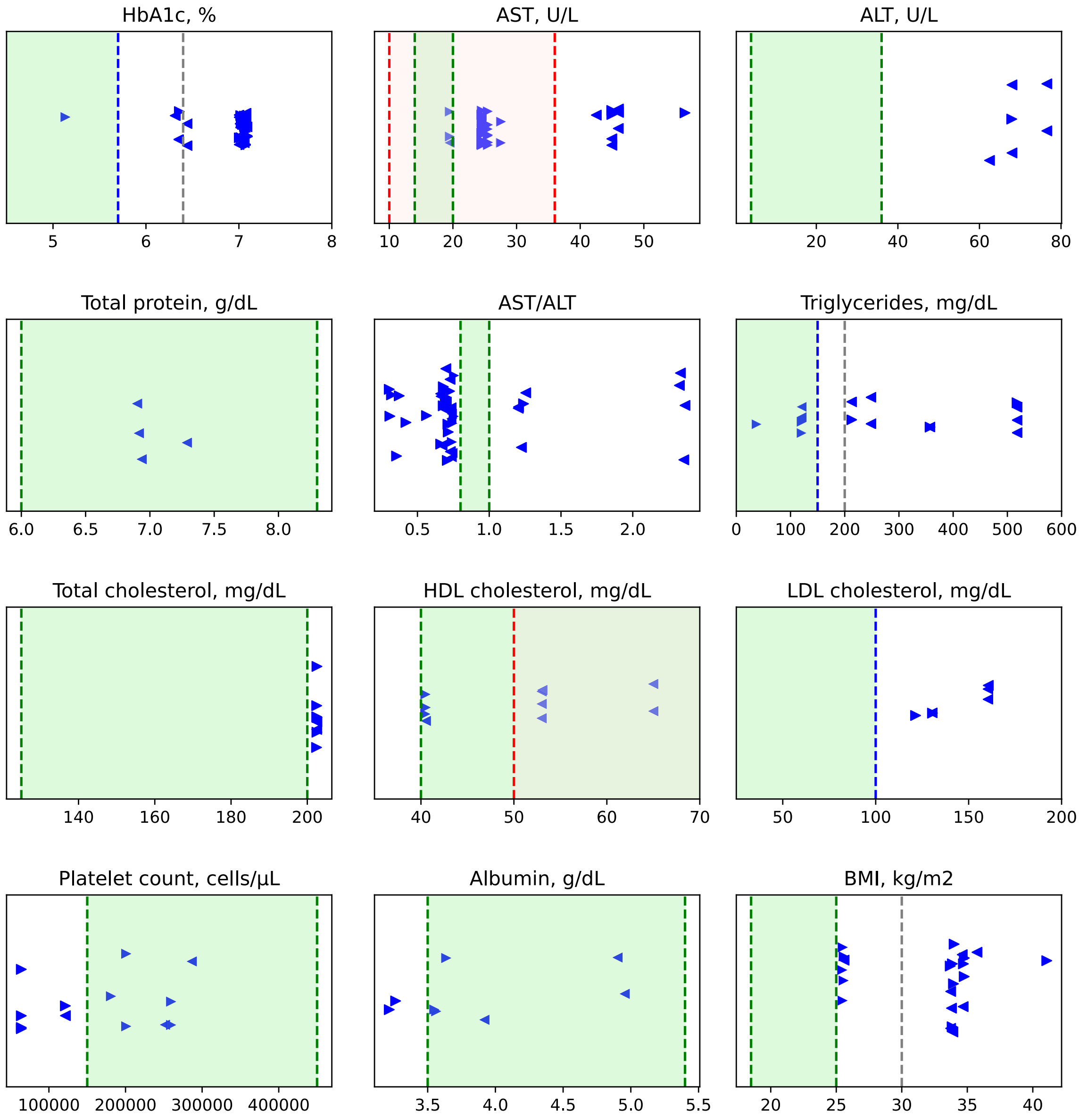


Fig. 5 Analysis of the borders used in the explanations of Decision Tree predictions. ▶ and ◀ signs mean that the textual explanation referred to “greater than” or “less than” a particular value of the feature, respectively. The vertical lines and the green zone between them correspond to the normal range of the parameter. In cases where the normal range depends on gender, the women’s range is highlighted with a light red color. For some analysis (e.g., HbA1c, Triglycerides, and BMI), we draw additional gray lines corresponding to the “increased risk” zone.

knowledge about the normal range of each particular property of the patient. To create this plot, we take the best-performing DT, run it on the corresponding test set, and collect the statistics of what exactly is verbalized to explain the prediction. Each subplot of Figure 5 corresponds to the feature engaged in the prediction. We use the ▶ and ◀ signs to refer to the fact that during the explanation it is mentioned that the feature is “greater than” or “less than” some value. For example, if the explanation of the prediction of NASH contains the phrase “... because (from most to least important features) HbA1c is greater than 5.2%” we place the ▶ sign in the HbA1C subplot on the X-coordinate corresponding to the 5.2% (note that we slightly vary the X and Y coordinates to make the signs readable on the plots).

When the values referred to during the DT prediction explanations are visualized, we may compare them to the common knowledge of the normal ranges of these parameters. The exact borders were collected from the guidelines frequently used in one of the author’s medical practice experience, who is a medical practitioner. We indicate the borders of normal ranges with vertical dot lines, and draw the area corresponding to the normal range with a green color. When the normal ranges for men and women vary, we use a light red color to highlight the normal range for women. Additionally, in some cases, we draw a “High risk” border of the feature if it is frequently used in real practice.

From this plot, we can see that the sensibility of the borders shown during the explanation varies. In some cases, like “Total cholesterol”, “HDL cholesterol” and “AST/ALT” the values show a significant correlation with the known normal range borders. Some features like “HbA1C”, “BMI”, or “Trigliceridies” do engage meaningful values for DT explanation, but still, a significant amount of such values seem to be placed pretty far from the commonly known normal ranges. Finally, there are a significant number of features whose values, while being used for DT explanation, do not seem to make much sense from the medical point of view, e.g., “ALT”, “Total protein”, “Platelet count”.

Note that we do not show the “Waist circumference” feature on Figure 5, because it turned out to not be engaged in any explanation of the selected test set. This corresponds to its low importance, as shown in the previous section on Figure 4a.

4.3 User study of explanations perception

Whereas the user-agnostic analysis performed in Section 4.2 is important, it is clear that the addressee of an explanation is a medical expert who will finally decide the patient’s diagnosis (probably) relying on the explained model’s prediction. That is why it is crucial to verify the usability of the generated explanations with the experts.

We engaged 15 medical experts in the user study using the professional network of one of the authors. Participation in the survey was voluntary and not paid. That is why we manually analyze the collected responses to make sure that they are sensible. In particular, we checked the time spent answering each task page, the declared experience with NASH, and the free-form answers to the question about the practices they use for the treatment and diagnosis of NASH.

We dropped the results of 4 participants as far as all the answers to Likert-scale-based questions from them were similar within each page, and the submission time of such pages was pretty small (normally below 20 seconds). This was the only exclusion criteria applied to filter the experts, i.e., all other participants indicated that they at least sometimes work with NASH patients, provided sensible descriptions of the practices used for NASH diagnosis, and their answers to Likert-scale-based questions varied and were submitted within adequate time (normally 60 seconds or more).

After filtering, the number of participants decreased to 11 medical experts with different experiences, which varied from 3 to 32 years (median 8 years). 7 of them were gastroenterologists, 3 were cardiologists, and 1 was a therapist. They gave 10 and 14 answers about DT and RF-based predictions, respectively.

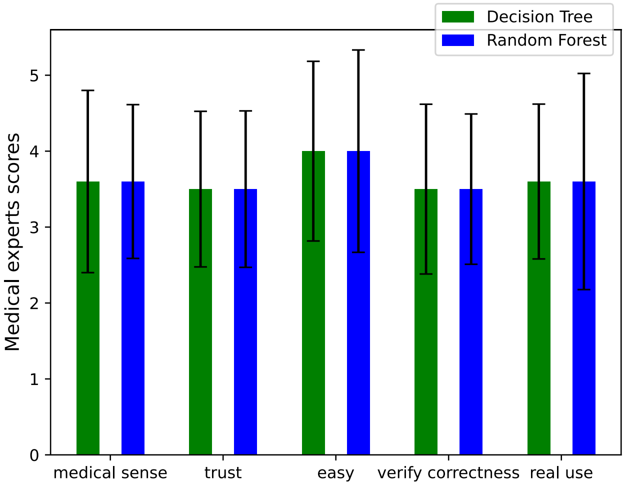


Fig. 6 Aggregated statistics of a user study on the quality of the DT and RF models explanations. The labels correspond to the questions about the personal perception of the explanation shown in Section 3.7.

We map the verbal Likert-scale answers to the integers from 1 to 5, where 1 means “Strongly disagree” and 5 means “Strongly agree”. We show the aggregated scores in Figure 6. On the plot, it can be seen that the scores of DT and RF corresponding to all 5 questions do not differ significantly from each other. To analytically study the significance of the difference between the scores of the two approaches, we use a dependent t-test for paired samples because the evaluated and explained predictions are obtained from the data of similar patients. This test confirms the visual intuition - in none of the 5 questions, the difference between DT and RF scores turns out to be significant (the p-value is greater than 0.05).

We also analyze the comments provided by the participants during the survey. Typical comments are related to the unclarity of SHAP-value-based grouping of the features’ importance. Here are some examples of such comments:

- *It is not very clear why deviations that are not specific are noted in the explanation of low risk, and not, on the contrary, normal indicators that reduce this risk are highlighted*
- *It’s not entirely clear why albumin is placed in the first place and LDL in the third*

Indeed, SHAP values are particularly useful for taking a look into the logic of certain decisions in the model. However, it is not guaranteed that the model itself learned all the peculiarities of the diagnosis correctly, which sometimes results in not very sensible priorities for the highlighted features.

In addition to SHAP-related feedback, several practitioners also commented on the lack of certain clinical features they considered important for diagnosis. However, these suggestions varied considerably between individuals, with no consistent pattern in the requested features. This observation supports the idea that the feature set we constructed-with input from a dedicated team of medical practitioners and supported by automatic selection techniques-likely covers the general diagnostic needs, though specific clinical scenarios may require more tailored or extensive information.

By this moment, we have enough information to answer the *RQ1 (Does the transparency of a model increase the acceptability of the model to medical professionals?)*. First, in Section 4.2 we show that even though some borders used to explain DT predictions make sense from a medical knowledge point of view, in many cases such borders are not very clear. Second, as mentioned above, there is no statistically significant difference in any studied property of medical experts' perception of the explanation obtained from DT and RF. Finally, given the aforementioned lack of sensibility of some feature borders used for DT prediction explanation, we may expect some relevant comments about this from the experts who participated in the study. However, this is not the case; no one asks any questions about why the explanation includes the reference to "greater than" or "less than" this particular value. The possible explanation for this lack of interest in these sometimes arguable borders yielded by the DT decision path is that the experts do not pay specific attention to what borders are highlighted in the explanation. Instead, they seem to analyze just the names of the features referred to in the explanation text and what group (increasing or decreasing NASH risk) they belong to. Thus *the answer to RQ 1 is No*.

We may also answer *RQ2 (Does the explanation of the AI model's prediction increase the acceptability of the model to medical professionals?)* using the results of the user study. Recall that we do not explicitly ask participants about their perception of the model prediction without explanation, as far as we assume that the blind belief in the model in the medical domain tasks is unacceptable [45]. Furthermore, posing questions that inherently require a demonstration of the explanation behind a prediction might compromise the quality of responses to inquiries about the explained predictions, which are essential for our study.

To study the general increase in usability of the models' predictions with an explanation, we analyze the answers to the questions about the explained predictions in Table 4.

Table 4 Analysis of the different properties of explained predictions perception obtained from DT and RF separately and jointly. p_i refers to the p-value calculated with a one-sample t-test with the population mean i. "*" sign indicates that the null hypothesis of equality between the population mean and the mean of the analyzed distribution for the population mean i is accepted.

Expl. perception	$p - val_2$	$p - val_3$	$p - val_4$
$\mathrm{DT}_{medicalsense}$	0.003	0.168*	0.343*
$\mathrm{RF}_{medicalsense}$	0.0008	0.459*	0.015
$\mathrm{DT{+}RF}_{medicalsense}$	4.73e-06	0.119*	0.013
DT_{trust}	0.002	0.177*	0.177*
RF_{trust}	0.0005	0.336*	0.027
$\mathrm{DT{+}RF}_{trust}$	1.63e-06	0.095*	0.008
DT_{easy}	0.0006	0.032	1.0*
RF_{easy}	0.0004	0.075*	0.453*
$\mathrm{DT{+}RF}_{easy}$	5.29e-07	0.005	0.539*
$\mathrm{DT}_{verifycorrectness}$	0.003	0.213*	0.213*
$\mathrm{RF}_{verifycorrectness}$	0.001	0.612*	0.008
$\mathrm{DT{+}RF}_{verifycorrectness}$	5.89e-06	0.2*	0.004
$\mathrm{DT}_{realuse}$	0.001	0.111*	0.269*
$\mathrm{RF}_{realuse}$	0.009	0.596*	0.068*
$\mathrm{DT{+}RF}_{realuse}$	3.36e-05	0.175*	0.029

We apply a one-sample t-test to all scores collected about DT and RF-explained predictions in isolation and jointly. The idea of this test is to compare the distribution of the scores collected in terms of a user study with the mean of the unknown population. The null hypothesis is that the mean of the collected scores and observations is equal to the given population mean. We perform a comparison to the following means: 2 ("Somewhat disagree"), 3 ("Neither agree nor disagree"), and 4 ("Somewhat agree"). We mark cases where the null hypothesis is accepted with "*" which means that the mean of the distribution is equal to the particular value.

In all the cases of the first column ($p - val_2$) the null hypothesis that the mean of the distribution equals 2 is rejected. For the second and third columns, the null hypothesis is accepted either for mean value 3 or 4 (or both), which has the general interpretation that none of the models received bad marks (in the rank below 3) about any explanation perception property. Looking at the data in more detail, we can see that in most cases, the mean of the experts' scores is 3 ("Neither agree nor disagree"). However, for the readiness to use the explained predictions in real practice, the null hypothesis was accepted for both population means 3 and 4. The only property for which the null hypothesis was accepted for the mean value 4 is the easiness of perception.

In general, these significant results indicate that the expert markings are, in most cases, "Neither agree nor disagree." Therefore, we cannot conclude that the medical practitioners in this case express a significant, general, and consistent preference for the explanations of the prediction models. Still, we may see that none of the explanations obtained "Strongly disagree" or "Somewhat disagree" scores. So *the obtained results are promising, but they do not allow us to conclusively answer RQ2 positively, and further evidence needs to be gathered in this regard.*

One possible approach to getting a more precise answer about the real usability of the explained prediction could be running a study where one group of medical professionals scores unexplained predictions and answers some questions specific to this type of prediction, and another group answers the questions only about explained predictions. However, it is clear that for such a study, a larger number of participants will be needed to get significant results.

5 Discussion

In this section, we discuss several insights collected from our study.

5.1 Experts had their own views of important features, different from models.

In the era of rapid development of AI, even the conservative medical field is turning towards modern technologies; more and more medical data is collected specifically for ML model training. However, we may hardly assume that every possible peculiarity of the phenomena to which a particular dataset is dedicated may be visited with the cases included in such datasets. This makes cross-disciplinary collaboration between AI and medical specialists crucial to developing a model that could at least theoretically be useful in real practice. This is especially important in terms of selecting the features for training the model.

Our variable selection process included several iterations. We started with the feature set proposed in [11]. In this work, the authors pre-selected the initial set of variables manually. Then they relied on a recursive feature elimination algorithm to select variables that maximized the metrics of the trained models. However, after several attempts to show the explained predictions based on these features to an initially small group of experts, we had to significantly modify the list of features: almost all preliminary interviewed experts highlighted the lack of Cholesterol and waist circumference, whereas the importance of those features could hardly be learned from pure data. Refer to Table 2 where it is shown that the distributions of HDL cholesterol and waist circumference of NASH and non-NASH patients do not have a significant difference, and moreover, the waist circumference data is not available in more than half of the cases from the selected cohort.

5.2 Explanations of DT are rated no better than explanations of the better-performing non-transparent model, so use the latter.

Our study directly explores the commonly cited trade-off between model transparency and predictive performance [64, 66]. Transparent models like Decision Trees are inherently easier to interpret and offer detailed, rule-based like explanations. However, this interpretability often comes at the cost of reduced accuracy, as confirmed in our experiments, where Decision Trees consistently underperformed compared to more complex, non-transparent models like Random Forests. In theory, this loss in performance might be justified if transparency meaningfully improves practitioner trust or understanding. However, our user study revealed that clinicians did not express a clear preference for explanations from the transparent model, nor did they appear to engage with its specific decision thresholds. This suggests that, in practice, the gain in interpretability did not translate into added value for end users. While methods like constraint-regularized DTs [67] could enhance the clarity of explanations, they still assume that users actively engage with such structure-which was not observed in our study. As a result, in this setting, prioritizing performance through more accurate models like Random Forests appears to be the more effective choice.

5.3 Explainability may not be enough if the model's performance is far from perfect

Overall, even though the idea of the usability of the explanation seems intuitive, it is necessary to remember that the explanation depends on the model. In the case of our work, it seems that the models trained on the data were far from ideal. This can be seen from the metrics in Table 3. Even the best-performing non-transparent models failed to overcome the 0.7 thresholds in most of the metrics, which certainly leaves much room for improvement. This problem is most probably caused by the small amount of cases available to us. Thus, the explanation techniques applied to both DT and RF may allow us to understand the model's logic, but this logic may not always be right.

However, even with the aforementioned limitations, we can see that the participants in the user study gave mostly positive feedback. Still, we believe that if more data of similar quality is available, the results of the experts' feedback can become significantly better.

6 Limitations and Future Work

Our study has several limitations that open up directions for future research. First, the dataset used for model training was derived from the NIDDK repository and filtered using strict inclusion and exclusion criteria. While this ensures internal consistency and clinical focus, it limits the generalizability of the model to broader patient populations, particularly those with comorbidities or atypical profiles. Future studies could address this by evaluating the proposed approach on more heterogeneous cohorts or across multiple datasets, including multi-center or commercial data such as OPTUM.

Second, the overall size of the dataset is relatively small, which may constrain the model's ability to capture complex patterns and increases the risk of overfitting. Although this was partially mitigated through data augmentation and stratified cross-validation, access to larger and more diverse datasets would enable more robust training and validation.

Third, our user study involved 11 medical practitioners, which is in line with similar studies but still limited in scale. Participation was voluntary and unpaid, and while we applied quality control measures, we cannot fully rule out variability in attention or engagement. Expanding the user study to include more participants and a broader range of clinical backgrounds-potentially through multi-institutional collaborations-would strengthen the findings and allow for deeper subgroup analyses.

Fourth, while we incorporated SHAP for generating explanations due to its stability and compatibility with structured data, other XAI techniques-such as counterfactuals, contrastive explanations, or domain-specific visualizations-may offer complementary insights. Future work should investigate how these methods compare in terms of clinician trust and usability.

Finally, future studies could explore the use of more complex models, including deep neural networks, TabNet, or TabTransformer, to examine whether performance gains justify the reduced transparency. However, incorporating such models into clinician-centered evaluation would require re-running human studies with new explanation formats, which presents significant logistical challenges.

7 Conclusion

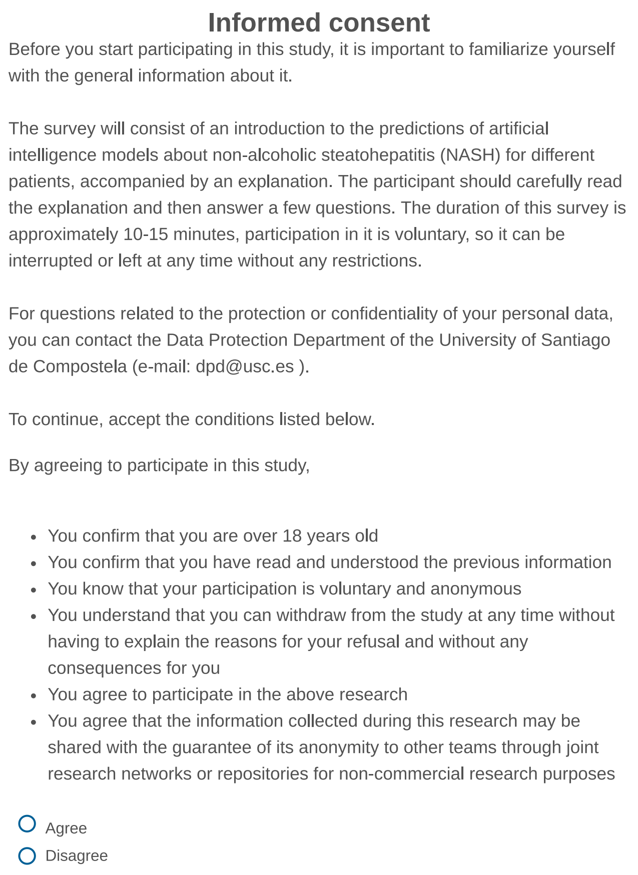
In this work, we train several machine learning models predicting non-alcoholic steatohepatitis (NASH) on the features manually selected to be maximally relevant to the ones used in the real practice of NASH diagnostics. We propose several explanation techniques for transparent and non-transparent models and show that for the task of non-invasive NASH diagnosis, transparent models are less useful because they have worse metrics than non-transparent models, and their transparency does not significantly improve the perceived quality and trustworthiness of the obtained explanations. We also show that overall, the explanation of the models' predictions slightly increases their usability for real practitioners. Finally, we open-source the trained models and code for explanation generation to make it easily re-usable for the research community. Given the general approach to structured clinical data, our methodology could be adapted to other diagnostic tasks where model explainability is essential for building clinician trust and supporting decision-making.

Appendix A Inclusion and exclusion criteria

Table 5 Inclusion and exclusion criteria were applied in our study. Seq. Indicates the serial number corresponding to the sequence of applying the particular criterion.

Seq.	Skipped patients	Criterion
1	385	Patient is over 18 y.o.
2	142	At least one histology review is available
3	54	Any alcohol use disorder
4	3	Hepatocellular carcinoma
5	3	Hepatitis B
6	0	Hepatitis C
7	5	Autoimmune hepatitis
8	0	Autoimmune cholestatic liver disorder
9	0	Wilson's disease
10	0	Alpha-1 antitrypsin deficiency
11	82	Iron overload
12	1	Dysbetalipoproteinemia
13	9	Stapling or banding of the stomach
14	6	Jejunoileal bypass
15	0	Biliopancreatic diversion
16	1	Total parenteral nutrition
17	0	Short bowel syndrome
18	0	Has the patient ever received a liver transplant
19	0	Hemophilia
20	159	Any corticosteroids taken (6 month before study or during the study)
21	5	Amiodarone
22	1	Methotrexate
23	3	Tamoxifen
24	0	Valproate sodium

Appendix B Informed consent



Informed consent

Before you start participating in this study, it is important to familiarize yourself with the general information about it.

The survey will consist of an introduction to the predictions of artificial intelligence models about non-alcoholic steatohepatitis (NASH) for different patients, accompanied by an explanation. The participant should carefully read the explanation and then answer a few questions. The duration of this survey is approximately 10-15 minutes, participation in it is voluntary, so it can be interrupted or left at any time without any restrictions.

For questions related to the protection or confidentiality of your personal data, you can contact the Data Protection Department of the University of Santiago de Compostela (e-mail: dpd@usc.es).

To continue, accept the conditions listed below.

By agreeing to participate in this study,

- You confirm that you are over 18 years old
- You confirm that you have read and understood the previous information
- You know that your participation is voluntary and anonymous
- You understand that you can withdraw from the study at any time without having to explain the reasons for your refusal and without any consequences for you
- You agree to participate in the above research
- You agree that the information collected during this research may be shared with the guarantee of its anonymity to other teams through joint research networks or repositories for non-commercial research purposes

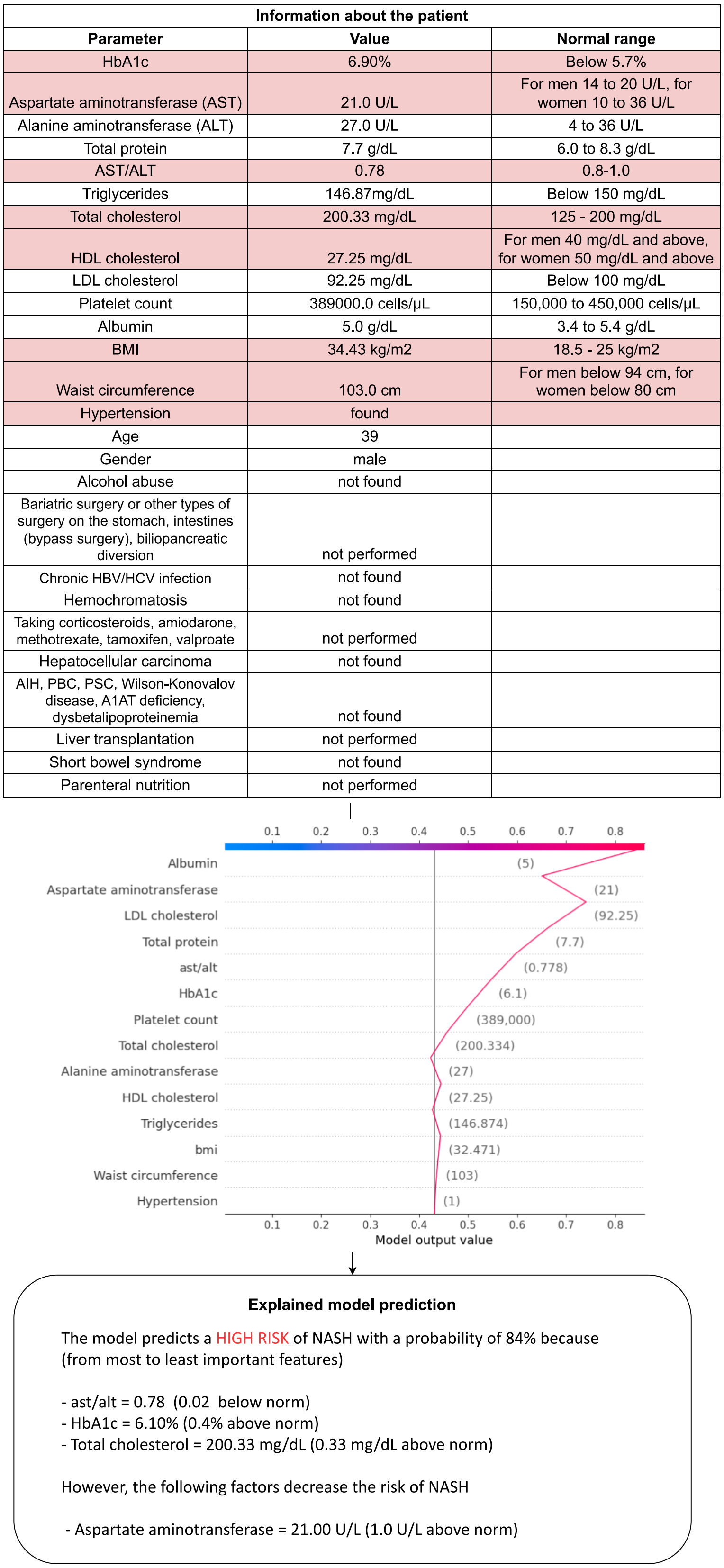
○ Agree

○ Disagree

Fig. 7 Informed consent demonstrated to the participants before the start of the survey.

Appendix C Explanation interfaces

Fig. 8 Non-transparent model explanation example.



Information about the patient		
Parameter	**Value**	**Normal range**
HbA1c	6.90%	Below 5.7%
Aspartate aminotransferase (AST)	21.0 U/L	For men 14 to 20 U/L, for women 10 to 36 U/L
Alanine aminotransferase (ALT)	27.0 U/L	4 to 36 U/L
Total protein	7.7 g/dL	6.0 to 8.3 g/dL
AST/ALT	0.78	0.8-1.0
Triglycerides	146.87mg/dL	Below 150 mg/dL
Total cholesterol	200.33 mg/dL	125 - 200 mg/dL
HDL cholesterol	27.25 mg/dL	For men 40 mg/dL and above, for women 50 mg/dL and above
LDL cholesterol	92.25 mg/dL	Below 100 mg/dL
Platelet count	389000.0 cells/μL	150,000 to 450,000 cells/μL
Albumin	5.0 g/dL	3.4 to 5.4 g/dL
BMI	34.43 kg/m2	18.5 - 25 kg/m2
Waist circumference	103.0 cm	For men below 94 cm, for women below 80 cm
Hypertension	found	
Age	39	
Gender	male	
Alcohol abuse	not found	
Bariatric surgery or other types of surgery on the stomach, intestines (bypass surgery), biliopancreatic diversion	not performed	
Chronic HBV/HCV infection	not found	
Hemochromatosis	not found	
Taking corticosteroids, amiodarone, methotrexate, tamoxifen, valproate	not performed	
Hepatocellular carcinoma	not found	
AIH, PBC, PSC, Wilson-Konovalov disease, A1AT deficiency, dysbetalipoproteinemia	not found	
Liver transplantation	not performed	
Short bowel syndrome	not found	
Parenteral nutrition	not performed	

Explained model prediction

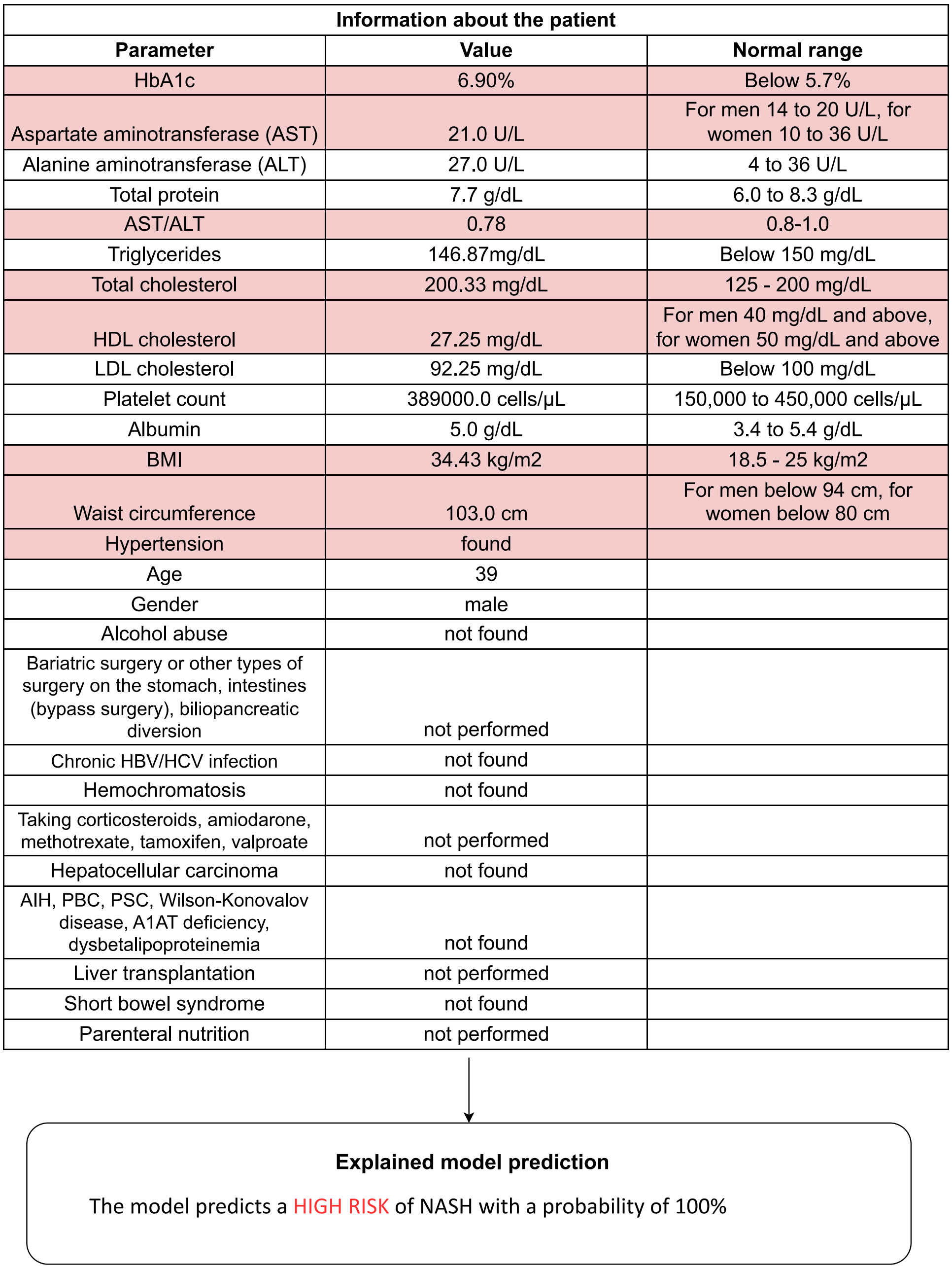
The model predicts a HIGH RISK of NASH with a probability of 84% because (from most to least important features)

- ast/alt = 0.78 (0.02 below norm)
- HbA1c = 6.10% (0.4% above norm)
- Total cholesterol = 200.33 mg/dL (0.33 mg/dL above norm)

However, the following factors decrease the risk of NASH

- Aspartate aminotransferase = 21.00 U/L (1.0 U/L above norm)

Fig. 9 The example of a Decision Tree prediction demonstration interface without explanation (shown to the user, study participants before the explained prediction is shown).

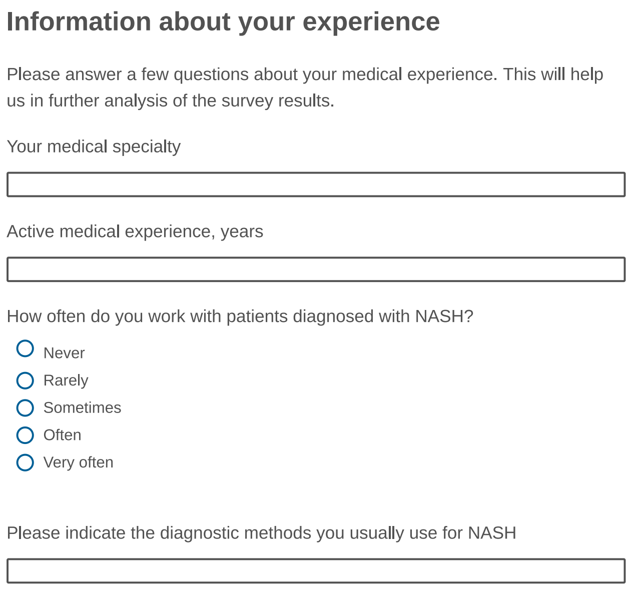


Information about the patient		
Parameter	**Value**	**Normal range**
HbA1c	6.90%	Below 5.7%
Aspartate aminotransferase (AST)	21.0 U/L	For men 14 to 20 U/L, for women 10 to 36 U/L
Alanine aminotransferase (ALT)	27.0 U/L	4 to 36 U/L
Total protein	7.7 g/dL	6.0 to 8.3 g/dL
AST/ALT	0.78	0.8-1.0
Triglycerides	146.87mg/dL	Below 150 mg/dL
Total cholesterol	200.33 mg/dL	125 - 200 mg/dL
HDL cholesterol	27.25 mg/dL	For men 40 mg/dL and above, for women 50 mg/dL and above
LDL cholesterol	92.25 mg/dL	Below 100 mg/dL
Platelet count	389000.0 cells/μL	150,000 to 450,000 cells/μL
Albumin	5.0 g/dL	3.4 to 5.4 g/dL
BMI	34.43 kg/m2	18.5 - 25 kg/m2
Waist circumference	103.0 cm	For men below 94 cm, for women below 80 cm
Hypertension	found	
Age	39	
Gender	male	
Alcohol abuse	not found	
Bariatric surgery or other types of surgery on the stomach, intestines (bypass surgery), biliopancreatic diversion	not performed	
Chronic HBV/HCV infection	not found	
Hemochromatosis	not found	
Taking corticosteroids, amiodarone, methotrexate, tamoxifen, valproate	not performed	
Hepatocellular carcinoma	not found	
AIH, PBC, PSC, Wilson-Konovalov disease, A1AT deficiency, dysbetalipoproteinemia	not found	
Liver transplantation	not performed	
Short bowel syndrome	not found	
Parenteral nutrition	not performed	

Explained model prediction

The model predicts a HIGH RISK of NASH with a probability of 100%

Appendix D Questions interface examples



Information about your experience

Please answer a few questions about your medical experience. This will help us in further analysis of the survey results.

Your medical specialty

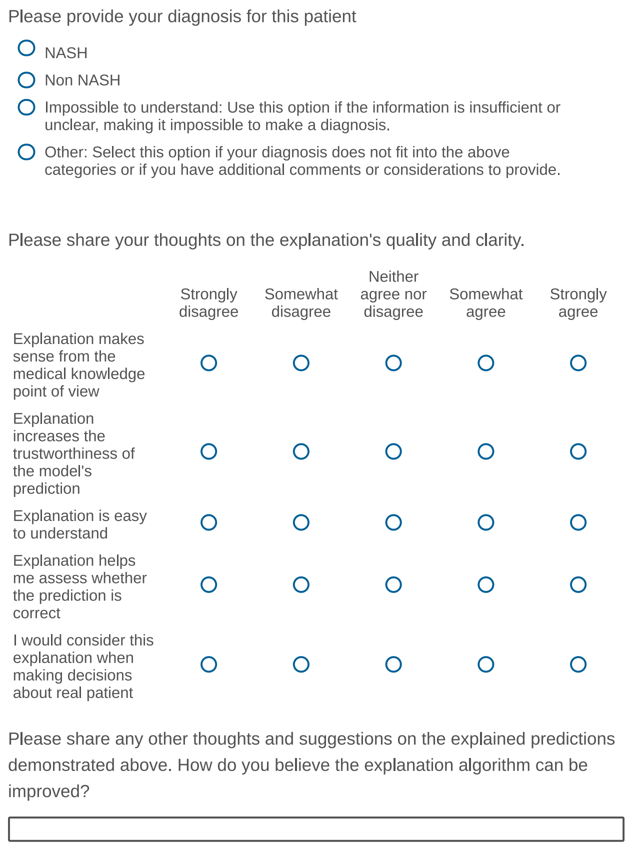
Active medical experience, years

How often do you work with patients diagnosed with NASH?

○ Never
○ Rarely
○ Sometimes
○ Often
○ Very often

Please indicate the diagnostic methods you usually use for NASH

Fig. 10 The interface of the primary questions asked to participants.



Please provide your diagnosis for this patient

○ NASH
○ Non NASH
○ Impossible to understand: Use this option if the information is insufficient or unclear, making it impossible to make a diagnosis.
○ Other: Select this option if your diagnosis does not fit into the above categories or if you have additional comments or considerations to provide.

Please share your thoughts on the explanation's quality and clarity.

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
Explanation makes sense from the medical knowledge point of view	○	○	○	○	○
Explanation increases the trustworthiness of the model's prediction	○	○	○	○	○
Explanation is easy to understand	○	○	○	○	○
Explanation helps me assess whether the prediction is correct	○	○	○	○	○
I would consider this explanation when making decisions about real patient	○	○	○	○	○

Please share any other thoughts and suggestions on the explained predictions demonstrated above. How do you believe the explanation algorithm can be improved?

Fig. 11 The example of an interface of the questions about the explained predictions.

Appendix E Best-performing models hyperparameters

Table 6 shows the data pre-processing and model-specific hyperparameters that resulted in the best performance of the corresponding model.

Data pre-processing

Recall the meaning of the engaged pre-processing steps

- **Missing drop.** Whether the features with too many (more than 50%) missing values were dropped.
- **Aug** Whether the augmentation was applied to the data.
- **KNN** Whether the KNN-imputation was used to impute the missing values.

See Section 3.3 for the details of the pre-preprocessing.

Model-specific hyperparameters.

We used sklearn [3] for the training of most of the models (RF, NB, SVM, KNN, DT, LR, MLP). For LR and NB the default sklearn parameters were used. For other models, the precise hyperparameters are specified.

We used xgboost [4] for XGB, catboost [5] for CatBoost, pgmpy [68] for BN training.

[3] https://github.com/scikit-learn/scikit-learn, retrieved on February 13th, 2024

[4] https://github.com/dmlc/xgboost, retrieved on February 13th, 2024

[5] https://catboost.ai/docs/en/, retrieved on April 1st, 2025

Table 6 Hyperparameters corresponding to the best-performing setup of the trained models.

Model	Missing drop	Aug	KNN	Model-specific parameters
RF	-	+	+	depth=10,estimators=40,max features=10, criterion=log-loss
XGB	-	+	+	estimators=150,dept=8,lr=0.2
NB	+	+	-	default
SVM	-	-	+	c=2,kernel=poly,gamma=auto
KNN	-	-	+	neighbours=35,algorithm=auto
BN	+	+	-	Struct.Learn.:Hill-Climbing,score=aicscore; Parameters: Max.Likelihood
DT	-	+	+	depth=10
LR	-	-	+	default
CatBoost	-	+	+	iterations=500, lr=0.05, depth=6, l2_leaf_reg=3
MLP	-	+	+	hidden layer = (64,32), activation = relu, lr=adaptive, solver=adam

Acknowledgements We thank Ekaterina Tkachenko for her help in engaging medical experts for collaboration, Natalia Motyleva for sharing her experience about NASH, and Viktoria Evdokimenko and Anna Badalyan for initial feedback about features commonly used for NASH diagnostics. The study was supported by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). The NIDDK Central Repository supplied the data reported here. This manuscript was not prepared in collaboration with Investigators of the NIDDK Central Repository, or NIDDK, and does not necessarily reflect their opinions or views.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This research was funded by the European Union's Horizon 2020 research and innovation program under the Marie Skodowska-Curie grant agreement No 860621. It also contributes to the project PID2020-112623GB-I00 funded by MCIN/AEI/10.13039/501100011033/ and by ERDF A way of making Europe. The support of the Galician Ministry for Education, Universities and Professional Training and the "ERDF A way of making Europe" is also acknowledged through grants "Centro de investigación de Galicia accreditation 2024-2027 ED431G-2023/04" and "Reference Competitive Group accreditation 2022-2025 ED431C 2022/19".

Code and data availability Data from the NIDDK NAFLD Adult Database used in this paper is available for request at the NIDDK Central Repository (NIDDK-CR) website(https://repository.niddk.nih.gov/studies/nafld_adult/, retrieved on February 13th, 2024) . The trained models and the code for explained predictions is available in our repository.(https://gitlab.nl4xai.eu/nikolay.babakov/explainable_prediction_NASH)

Declarations

Conflicts of Interest The authors declare that they are not involved in any financial and personal relationships with other people or organizations that could inappropriately influence this work.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Younossi, Z.M., Koenig, A.B., Abdelatif, D., Fazel, Y., Henry, L., Wymer, M.: Global epidemiology of nonalcoholic fatty liver disease-meta-analytic assessment of prevalence, incidence, and outcomes. Hepatology **64**(1), 73–84 (2016)
2. Marchesini, G., Marzocchi, R., Agostini, F., Bugianesi, E.: Nonalcoholic fatty liver disease and the metabolic syndrome. Curr. Opin. Lipidol. **16**(4), 421–427 (2005)
3. Ratziu, V., Bellentani, S., Cortez-Pinto, H., Day, C., Marchesini, G.: A position statement on NAFLD/NASH based on the EASL 2009 special conference. J. Hepatol. **53**(2), 372–384 (2010)
4. Drescher, H.K., Weiskirchen, S., Weiskirchen, R.: Current status in testing for nonalcoholic fatty liver disease (NAFLD) and nonalcoholic steatohepatitis (nash). Cells **8**(8), 845 (2019)
5. Schuppan, D., Afdhal, N.H.: Liver cirrhosis. The Lancet **371**(9615), 838–851 (2008)
6. Chalasani, N., Younossi, Z., Lavine, J.E., Diehl, A.M., Brunt, E.M., Cusi, K., Charlton, M., Sanyal, A.J.: The diagnosis and management of non-alcoholic fatty liver disease: practice guideline by the american association for the study of liver diseases, american college of gastroenterology, and the american gastroenterological association. Hepatology **55**(6), 2005–2023 (2012)
7. Perakakis, N., Stefanakis, K., Mantzoros, C.S.: The role of omics in the pathophysiology, diagnosis and treatment of non-alcoholic fatty liver disease. Metabolism **111**, 154320 (2020)
8. Yang, Y., Liu, J., Sun, C., Shi, Y., Hsing, J.C., Kamya, A., Keller, C.A., Antil, N., Rubin, D., Wang, H., et al.: Nonalcoholic fatty liver disease (NAFLD) detection and deep learning in a chinese community-based population. European Radiology, 1–13 (2023)
9. Zamanian, H., Mostaar, A., Azadeh, P., Ahmadi, M.: Implementation of combinational deep learning algorithm for non-alcoholic fatty liver classification in ultrasound images. Journal of Biomedical Physics & Engineering **11**(1), 73 (2021)
10. Pushpa, B., Baskaran, B., Vivekanandan, S., Gokul, P.: Liver fat analysis using optimized support vector machine with support vector regression. Technology and Health Care (Preprint), 1–20 (2023)
11. Docherty, M., Regnier, S.A., Capkun, G., Balp, M.-M., Ye, Q., Janssens, N., Tietz, A., Löffler, J., Cai, J., Pedrosa, M.C., Schat-

tenberg, J.M.: Development of a novel machine learning model to predict presence of nonalcoholic steatohepatitis. J. Am. Med. Inform. Assoc. **28**(6), 1235–1241 (2021). https://doi.org/10.1093/jamia/ocab003

12. Perakakis, N., Polyzos, S.A., Yazdani, A., Sala-Vila, A., Kountouras, J., Anastasilakis, A.D., Mantzoros, C.S.: Non-invasive diagnosis of non-alcoholic steatohepatitis and fibrosis with the use of omics and supervised learning: a proof of concept study. Metabolism **101**, 154005 (2019)
13. Ma, H., Xu, C.-f., Shen, Z., Yu, C.-h., Li, Y.-m.: Application of machine learning techniques for clinical predictive modeling: a cross-sectional study on nonalcoholic fatty liver disease in china. BioMed research international **2018** (2018)
14. Sorino, P., Campanella, A., Bonfiglio, C., Mirizzi, A., Franco, I., Bianco, A., Caruso, M.G., Misciagna, G., Aballay, L.R., Buongiorno, C.: Development and validation of a neural network for NAFLD diagnosis. Sci. Rep. **11**(1), 20240 (2021)
15. Feng, G., Zheng, K.I., Li, Y.-Y., Rios, R.S., Zhu, P.-W., Pan, X.-Y., Li, G., Ma, H.-L., Tang, L.-J., Byrne, C.D.: Machine learning algorithm outperforms fibrosis markers in predicting significant fibrosis in biopsy-confirmed NAFLD. J. Hepatobiliary Pancreat. Sci. **28**(7), 593–603 (2021)
16. Deo, R., Panigrahi, S.: Explainability analysis of black box svm models for hepatic steatosis screening. In: 2022 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT), pp. 22–25 (2022). IEEE
17. Fialoke, S., Malarstig, A., Miller, M.R., Dumitriu, A.: Application of machine learning methods to predict non-alcoholic steatohepatitis (nash) in non-alcoholic fatty liver (nafl) patients. In: AMIA Annual Symposium Proceedings, vol. 2018, p. 430 (2018). American Medical Informatics Association
18. Canbay, A., Kälsch, J., Neumann, U., Rau, M., Hohenester, S., Baba, H.A., Rust, C., Geier, A., Heider, D., Sowa, J.-P.: Non-invasive assessment of NAFLD as systemic disease-a machine learning perspective. PLoS ONE **14**(3), 0214436 (2019)
19. Wu, C.-C., Yeh, W.-C., Hsu, W.-D., Islam, M.M., Nguyen, P.A.A., Poly, T.N., Wang, Y.-C., Yang, H.-C., Li, Y.-C.J.: Prediction of fatty liver disease using machine learning algorithms. Comput. Methods Programs Biomed. **170**, 23–29 (2019)
20. Njei, B., Osta, E., Njei, N., Al-Ajlouni, Y.A., Lim, J.K.: An explainable machine learning model for prediction of high-risk nonalcoholic steatohepatitis. Sci. Rep. **14**(1), 8589 (2024)
21. Ghandian, S., Thapa, R., Garikipati, A., Barnes, G., Green-Saxena, A., Calvert, J., Mao, Q., Das, R.: Machine learning to predict progression of non-alcoholic fatty liver to non-alcoholic steatohepatitis or fibrosis. JGH Open **6**(3), 196–204 (2022)
22. Payrovnaziri, S.N., Chen, Z., Rengifo-Moreno, P., Miller, T., Bian, J., Chen, J.H., Liu, X., He, Z.: Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. J. Am. Med. Inform. Assoc. **27**(7), 1173–1185 (2020). https://doi.org/10.1093/jamia/ocaa053
23. Kyrimi, E., Dube, K., Fenton, N., Fahmi, A., Neves, M.R., Marsh, W., McLachlan, S.: Bayesian networks in healthcare: What is preventing their adoption? Artif. Intell. Med. **116**, 102079 (2021). https://doi.org/10.1016/j.artmed.2021.102079
24. Salih, A., Galazzo, I.B., Gkontra, P., Rauseo, E., Lee, A.M., Lekadir, K., Radeva, P., Petersen, S., Menegaz, G.: A review of evaluation approaches for explainable ai with applications in cardiology. Authorea Preprints (2023)
25. Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., Keulen, M., Seifert, C.: From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. ACM Comput. Surv. **55**(13s), 1–42 (2023)
26. Quinn, T.P., Jacobs, S., Senadeera, M., Le, V., Coghlan, S.: The three ghosts of medical ai: Can the black-box present deliver? Artificial Intelligence in Medicine 124, 102158 (2022) https://doi.org/10.1016/j.artmed.2021.102158
27. Loh, H.W., Ooi, C.P., Seoni, S., Barua, P.D., Molinari, F., Acharya, U.R.: Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). Computer Methods and Programs in Biomedicine, 107161 (2022)
28. Osheroff, J.A., Teich, J.M., Middleton, B., Steen, E.B., Wright, A., Detmer, D.E.: A roadmap for national action on clinical decision support. J. Am. Med. Inform. Assoc. **14**(2), 141–145 (2007)
29. Khairat, S., Marc, D., Crosby, W., Al Sanousi, A.: Reasons for physicians not adopting clinical decision support systems: critical analysis. JMIR Med. Inform. **6**(2), 8912 (2018)
30. Yang, Q., Zimmerman, J., Steinfeld, A., Carey, L., Antaki, J.F.: Investigating the heart pump implant decision process: opportunities for decision support tools to help. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 4477–4488 (2016)
31. Tonekaboni, S., Joshi, S., McCradden, M.D., Goldenberg, A.: What clinicians want: contextualizing explainable machine learning for clinical end use. In: Machine Learning for Healthcare Conference, pp. 359–380 (2019). PMLR
32. Yu, K.-H., Beam, A.L., Kohane, I.S.: Artificial intelligence in healthcare. Nature biomedical engineering **2**(10), 719–731 (2018)
33. Musen, M.A., Middleton, B., Greenes, R.A.: Clinical decision-support systems. In: Biomedical Informatics: Computer Applications in Health Care and Biomedicine, pp. 795–840. Springer, (2021)
34. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17, pp. 4768–4777. Curran Associates Inc., Red Hook, NY, USA (2017)
35. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
36. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
37. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) **51**(5), 1–42 (2018)
38. Chen, H., Lundberg, S.M., Erion, G., Kim, J.H., Lee, S.-I.: Forecasting adverse surgical events using self-supervised transfer learning for physiological signals. NPJ Digital Medicine **4**(1), 167 (2021)
39. Duckworth, C., Chmiel, F.P., Burns, D.K., Zlatev, Z.D., White, N.M., Daniels, T.W., Kiuber, M., Boniface, M.J.: Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during covid-19. Sci. Rep. **11**(1), 23017 (2021)
40. Zeng, X., Hu, Y., Shu, L., Li, J., Duan, H., Shu, Q., Li, H.: Explainable machine-learning predictions for complications after pediatric congenital heart surgery. Sci. Rep. **11**(1), 17244 (2021)
41. Koo, B.S., Eun, S., Shin, K., Yoon, H., Hong, C., Kim, D.-H., Hong, S., Kim, Y.-G., Lee, C.-K., Yoo, B.: Machine learning model for identifying important clinical features for predicting remission in patients with rheumatoid arthritis treated with biologics. Arthritis Research & Therapy **23**(1), 1–10 (2021)
42. El-Sappagh, S., Alonso, J.M., Islam, S.R., Sultan, A.M., Kwak, K.S.: A multilayer multimodal detection and prediction model based on explainable artificial intelligence for alzheimer's disease. Sci. Rep. **11**(1), 2660 (2021)
43. Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for explainable ai: Challenges and prospects. arXiv preprint arXiv:1812.04608 (2018)

44. Buçinca, Z., Lin, P., Gajos, K.Z., Glassman, E.L.: Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In: Proceedings of the 25th International Conference on Intelligent User Interfaces, pp. 454–464 (2020)
45. Rosenfeld, A.: Better metrics for evaluating explainable artificial intelligence. In: Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems, pp. 45–50 (2021)
46. Yang, G., Ye, Q., Xia, J.: Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. Information Fusion **77**, 29–52 (2022)
47. Ghassemi, M., Oakden-Rayner, L., Beam, A.L.: The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health **3**(11), 745–750 (2021)
48. Breiman, L.: Random forests. Machine learning **45**, 5–32 (2001)
49. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT press, (2009)
50. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16, pp. 785–794. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939785
51. Dorogush, A.V., Ershov, V., Gulin, A.: Catboost: gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363 (2018)
52. Popescu, M.-C., Balas, V.E., Perescu-Popescu, L., Mastorakis, N.: Multilayer perceptron and neural networks. WSEAS Transactions on Circuits and Systems **8**(7), 579–588 (2009)
53. Izza, Y., Ignatiev, A., Marques-Silva, J.: On explaining decision trees. arXiv preprint arXiv:2010.11034 (2020)
54. Audemard, G., Bellart, S., Bounia, L., Koriche, F., Lagniez, J.-M., Marquis, P.: On the explanatory power of decision trees. arXiv preprint arXiv:2108.05266 (2021)
55. Alves, M.A., Castro, G.Z., Oliveira, B.A.S., Ferreira, L.A., Ramírez, J.A., Silva, R., Guimarães, F.G.: Explaining machine learning based diagnosis of covid-19 from routine blood tests with decision trees and criteria graphs. Comput. Biol. Med. **132**, 104335 (2021). https://doi.org/10.1016/j.compbiomed.2021.104335
56. Rahimibashar, F., Miller, A.C., Salesi, M., Bagheri, M., Vahedian-Azimi, A., Ashtari, S., Moghadam, K.G., Sahebkar, A.: Risk factors, time to onset and recurrence of delirium in a mixed medical-surgical icu population: A secondary analysis using cox and chaid decision tree modeling. EXCLI J. **21**, 30 (2022)
57. Ghiasi, M.M., Zendehboudi, S., Mohsenipour, A.A.: Decision tree-based diagnosis of coronary artery disease: Cart model. Comput. Methods Programs Biomed. **192**, 105400 (2020)
58. Sivaprasad, A., Reiter, E., Tintarev, N., Oren, N.: Evaluation of Human-Understandability of Global Model Explanations using Decision Tree (2023)
59. Maruf, S., Zukerman, I., Reiter, E., Haffari, G.: Explaining Decision-Tree predictions by addressing potential conflicts between predictions and plausible expectations. In: Belz, A., Fan, A., Reiter, E., Sripada, Y. (eds.) Proceedings of the 14th International Conference on Natural Language Generation, pp. 114–127. Association for Computational Linguistics, Aberdeen, Scotland, UK (2021). https://aclanthology.org/2021.inlg-1.12
60. Mariotti, E., Alonso, J.M., Gatt, A.: Towards harnessing natural language generation to explain black-box models. In: 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence, pp. 22–27. Association for Computational Linguistics, Dublin, Ireland (2020). https://aclanthology.org/2020.nl4xai-1.6
61. Kyrimi, E., Mossadegh, S., Tai, N., Marsh, W.: An incremental explanation of inference in Bayesian Networks for increasing model trustworthiness and supporting clinical decision making. Artif. Intell. Med. **103**, 101812 (2020)
62. Timmer, S.T., Meyer, J.-J.C., Prakken, H., Renooij, S., Verheij, B.: A two-phase method for extracting explanatory arguments from Bayesian Networks. Int. J. Approximate Reasoning **80**, 475–494 (2017). https://doi.org/10.1016/j.ijar.2016.09.002
63. Kitson, N.K., Constantinou, A.C., Guo, Z., Liu, Y., Chobtham, K.: A survey of bayesian network structure learning. Artificial Intelligence Review, 1–94 (2023)
64. Herm, L.-V., Heinrich, K., Wanner, J., Janiesch, C.: Stop ordering machine learning algorithms by their explainability! a user-centered investigation of performance and explainability. Int. J. Inf. Manage. **69**, 102538 (2023)
65. Rosenfeld, A.: Better metrics for evaluating explainable artificial intelligence. In: Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems. AAMAS '21, pp. 45–50. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2021)
66. Crook, B., Schlüter, M., Speith, T.: Revisiting the performance-explainability trade-off in explainable artificial intelligence (xai). In: 2023 IEEE 31st International Requirements Engineering Conference Workshops (REW), pp. 316–324 (2023). IEEE
67. Nanfack, G., Temple, P., Frénay, B.: Constraint enforcement on decision trees: A survey. ACM Comput. Surv. **54**(10s) (2022) https://doi.org/10.1145/3506734
68. Ankan, A., Panda, A.: pgmpy: Probabilistic graphical models using python. In: Proceedings of the 14th Python in Science Conference (SCIPY 2015) (2015). Citeseer

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.