Machine Learning-Based Clinical Decision Support System for Hepatic Fibrosis Risk Prediction in General Practice

MICHELE BERNARDINI, Department of Information Engineering, Università Politecnica delle Marche, Ancona, Italy and Department of Theoretical and Applied Sciences, eCampus University, Novedrate, Italy
MARIACHIARA DI COSMO, Department of Information Engineering, Università Politecnica delle Marche, Ancona, Italy and Department of Innovative Technologies in Medicine and Dentistry, Università degli Studi "G. d'Annunzio" di Chieti-Pescara, Chieti, Italy
GAIA BARONE, National College of Ireland, Dublin, Ireland
LUCA ROMEO, Department of Economics and Law, Università di Macerata, Macerata, Italy
EMANUELE FRONTONI, Department of Political Sciences, Communication and International Relations, Università degli Studi di Macerata, Macerata, Italy

Hepatic steatosis, or non-alcoholic fatty liver disease (NAFLD), affects a significant portion of the global population and can lead to more severe liver conditions, including hepatic fibrosis. Early and accurate risk prediction of fibrosis is crucial for timely intervention. Traditional diagnostic methods are invasive and carry risks, while imaging techniques and blood-based biomarkers have limitations in routine general practice. This study presents a machine learning-based clinical decision support system designed to assess the risk of hepatic fibrosis in patients with NAFLD using routine laboratory tests. The framework is developed using electronic health record data collected over 15 years, initially encompassing 1,272,572 patients from general practice. After applying clinical selection criteria, two cohorts of 12,960 and 25,478 patients were used for model development and evaluation. The proposed approach provides a robust foundation for monitoring fibrosis risk by implementing a novel screening method, which preprocesses predictors by leveraging well-established clinical indicators (e.g., hepatic steatosis index, fibrosis-4 index), alongside a selected minimal number of predictors, making it practical and cost-effective for widespread clinical use. The study's findings indicate promising results for screening and monitoring fibrosis risk in NAFLD patients, achieving the best AUC of 92.97%, PRAUC of 75.44%, and Sensitivity of 79.63%.

CCS Concepts: • **Applied computing → Health informatics**; **Health care information systems**; • **Computing methodologies → Machine learning**;

Additional Key Words and Phrases: Hepatic Steatosis, Hepatic Fibrosis, Electronic Health Records, Predictive Medicine, General Practice

Authors' Contact Information: Michele Bernardini (corresponding author), Department of Information Engineering, Università Politecnica delle Marche, Ancona, Italy and Department of Theoretical and Applied Sciences, eCampus University, Novedrate, Italy; e-mail: m.bernardini@pm.univpm.it; Mariachiara di Cosmo, Department of Information Engineering, Università Politecnica delle Marche, Ancona, Italy and Department of Innovative Technologies in Medicine and Dentistry, Università degli Studi "G. d'Annunzio" di Chieti-Pescara, Chieti, Italy; e-mail: m.dicosmo@pm.univpm.it; Gaia Barone, National College of Ireland, Dublin, Ireland; e-mail: gaia.barone@ncirl.ie; Luca Romeo, Department of Economics and Law, Università di Macerata, Macerata, Italy; e-mail: luca.romeo@unimc.it; Emanuele Frontoni, Department of Political Sciences, Communication and International Relations, Università degli Studi di Macerata, Macerata, Italy; e-mail: emanuele.frontoni@unimc.it.

ACM 2637-8051/2026/3-ART19
https://doi.org/10.1145/3788673

ACM Reference format:
Michele Bernardini, Mariachiara di Cosmo, Gaia Barone, Luca Romeo, and Emanuele Frontoni. 2026. Machine Learning-Based Clinical Decision Support System for Hepatic Fibrosis Risk Prediction in General Practice. *ACM Trans. Comput. Healthcare* 7, 2, Article 19 (March 2026), 28 pages.
https://doi.org/10.1145/3788673

1 Introduction

Hepatic steatosis, also known as **Non-Alcoholic Fatty Liver Disease (NAFLD)**, is a condition characterized by the accumulation of fat in liver cells, which often precedes more severe liver diseases, including **Non-Alcoholic Steatohepatitis (NASH)** and fibrosis [4, 30]. NAFLD is one of the most common chronic liver diseases globally, estimated to affect up to 25% of the population [31]. Even though its critical progression to fibrosis is restricted to a minority (10–20% of them), it can lead to irreversible damage and potential cirrhosis [31]. Therefore, early and accurate prediction of this progression remains a pivotal challenge in hepatology, with significant implications for intervention and management.

The gold standard for evaluating liver fibrosis is liver biopsy. However, despite its diagnostic accuracy, it is invasive, costly, and associated with sampling variability and potential complications, making it unsuitable for population-level screening or routine use in general practice [38]. In this context, alternative non-invasive liver disease assessment methods are essential. Imaging methods, such as ultrasonography, computed tomography, or magnetic resonance imaging, can detect fatty liver's presence, but they are time-consuming, expensive, and often unavailable in daily routine [19, 28]. Blood-based biomarkers offer a widely accessible and cost-effective alternative, particularly useful for primary care and asymptomatic patients screening [1]. Current biomarkers are categorized into direct and indirect markers of fibrosis. Direct markers, such as the Enhanced Liver Fibrosis test, measure components directly involved in fibrogenesis or fibrinolysis, including substances like hyaluronic acid and the tissue inhibitor of metalloproteinases-1 [1]. Indirect markers assess fibrosis risk through surrogate measures that are not directly related to fibrogenesis but are altered in fibrotic conditions, including aminotransaminases and platelet count [1]. Standard biomarker-based indexes, mainly based on routine laboratory and anthropometric parameters, have been developed to evaluate NAFLD and fibrosis risk. Among them, the **Hepatic Steatosis Index (HSI)** and the **Fibrosis-4 Index (FIB-4)** have been reported to be closely associated with metabolic-related diseases such as insulin resistance, diabetes, and metabolic syndrome, among the principal risk factors for this condition. FIB-4, in particular, is recommended by international guidelines as a non-invasive tool to identify patients with suspected advanced fibrosis in NAFLD, especially in primary care settings [8].

Building upon the advantages of blood-based biomarkers, an automated system for fibrosis risk assessment in NAFLD patients could be incredibly beneficial in advancing early diagnosis and tailoring management strategies for liver fibrosis to reverse this condition. Integrating these biomarkers with advanced **Machine Learning (ML)** techniques into a **Clinical Decision Support System (CDSS)** for **General Practitioners (GPs)** could provide a tool to expedite patient screening and monitoring, and reduce costs and time associated with ongoing patient care.

This study aims to propose an ML-based CDSS, as depicted in Figure 1, for screening NAFLD patients and monitoring the risk of developing hepatic fibrosis leveraging routine laboratory test exams stored in **Electronic Health Records (EHRs)**.

Our contributions are multifaceted:

(i) We conducted a comprehensive retrospective study involving data from 595 EHR systems, comprising 1,272,572 patients whose clinical records were collected across 15 years (2008–2023) in daily general practice.
(ii) We introduced an innovative screening process for NAFLD patients using established and reliable clinical indices (HSI, FIB-4) from clinical practice, providing a solid foundation for monitoring fibrosis risk.

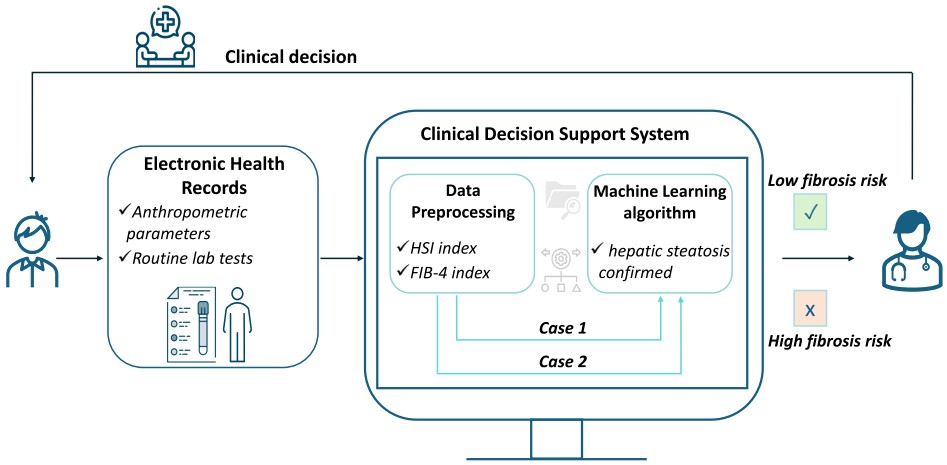


Fig. 1. Overview of the CDSS, which leverages routine patient exams to support hepatic steatosis (NAFLD) screening in the general population based on reliable, standardized indexes and monitoring their risk of hepatic fibrosis development through the integration of a ML algorithm.

(iii) We developed an ML-based CDSS to monitor fibrosis risk in NAFLD patients, employing varied assessment strategies based on patients' historical FIB-4 index values and allowing for more or less stringent evaluations of fibrosis progression risks.
(iv) We performed a detailed analysis to identify an optimal and minimal set of predictors, enhancing the CDSS's cost-effectiveness without compromising diagnostic precision.

The overall article is organized as follows: Section 2 presents the current state of the art; Section 3 details the proposed methodology, including dataset description, data preprocessing steps and ML experimental procedure; Section 4 provides a comprehensive analysis of the predictive performance achieved; Section 5 describes a use case for the proposed CDSS; Section 6 discusses the implications of our findings; finally, Section 7 concludes with reflections on the potential impact of incorporating ML into primary care for hepatological practice.

2 Related Work

EHRs facilitate the aggregation of large-scale, structured datasets, making them perfectly suitable for analysis through ML algorithms, which can reveal the complex hierarchical relationships between variables and offer flexible data modeling capabilities. Although ML potential is highly recognized for EHR data in various medical fields [2, 6, 20, 34, 36, 39], few studies, discussed below, leverage ML to predict patient risk for NAFLD and fibrosis.

Several studies have used ML to predict NAFLD within general populations. To assess NAFLD risk among 939 subjects, Perveen et al. [29] utilized a **Decision Tree (DT)** identifying seven risk factors based on clinical criteria defined by the Adult Treatment Panel III and achieving an AUC of 73.00%. Islam et al. [18] investigated the effectiveness of **Logistic Regression (LR)**, **Random Forest (RF)**, artificial neural network, and **Support Vector Machine (SVM)** in predicting NAFLD, using gender, age, and eight laboratory variables as predictors. LR obtained the best performance in a cohort of 994 subjects with an AUC of 76.00%. Yip et al. [40] assessed various ML models, including LR, ridge regression, AdaBoost, and DT, using six predictors from 23 routine clinical and laboratory variables across a cohort of 922 subjects, with ridge regression leading in performance metrics with an AUC of 88.00%. Despite their effectiveness, these studies primarily address the prediction of general NAFLD risk without distinguishing between isolated steatosis, NASH, or advanced fibrosis. This lack of differentiation could result in imprecise clinical guidance, as each stage of liver disease necessitates distinct management approaches.

Focusing on predicting fibrosis risk in a general population of 3,460 patients, the study by Blanes-Vidal et al. [9] employed ensemble learning models that processed 233 potential input variables, achieving an AUC of 94.00%. However, ML algorithms developed considering a general population to predict fibrosis risk may overgeneralize and reduce specificity, not capturing subtle changes specific to the progression from NAFLD to fibrosis. Hence, few studies have narrowed their focus to patients already diagnosed with NAFLD. Wu et al. [35] used an **eXtreme Gradient Boosting (XGB)** model to differentiate between simple steatosis, NASH, and fibrosis in 492 NAFLD patients, resulting in AUC scores of 90.00%, 82.00% and 83.00%, respectively. However, these studies [9, 35] do not detail their approach to handling the temporal dynamics of the data; thus, they may not adequately capture the evolving nature of liver disease progression over time. Similarly, Ghandian et al. [14] used an XGB model to predict NASH and fibrosis in 141,293 NAFLD patients, achieving AUC values of 79.00% and 87.00%, respectively. The study analyzed EHRs from 1 year before NAFLD diagnosis to 4 years after, based on the **International Classification of Disease (ICD)**-10 codes. The reliance on ICD codes, given the often silent progression of NAFLD and the need for liver biopsy confirmation, might limit the accuracy of disease onset detection.

Building upon the works discussed and their criticality, our study proposes an ML approach to fibrosis-risk prediction in NAFLD patients, identified from an extensive cohort of 1,272,572 subjects collected over 15 years of general practice. The study focuses on patients with confirmed NAFLD who are routinely followed in primary care, thus providing a clinically homogeneous population in which fibrosis progression is most relevant. To be reliable, compliance with well-established and clinically validated indices, such as HSI and FIB-4, is ensured and study design, including EHR selection and analysis criteria, aligns with established protocols already effective in other medical research fields [2, 7, 32], making our approach both innovative and grounded in proven research practices. Rather than providing a diagnostic tool to replace histology, the model is designed to highlight individuals whose profile from primary EHR data suggests a high probability of exceeding the guideline FIB-4 threshold, thus supporting timely referral for further diagnostic evaluation. Instead of modeling fibrosis progression as a multi-class or ordinal outcome, we adopt a binary classification strategy evaluated under two complementary configurations. Each configuration reflects distinct clinical screening scenarios on established FIB-4 thresholds. This design allows the system to align closely with real-world clinical practice and a realistic prospective-screening scenario while maintaining a clear and interpretable decision framework.

3 Materials and Methods

Our study is based on a retrospective analysis of longitudinal EHRs obtained from GPs. The proposed CDSS is developed and validated exclusively on historical data, ensuring that model evaluation accurately reflects a realistic screening scenario. The fibrosis label, derived from the FIB-4 index, is used solely during training to distinguish between control and fibrosis-risk cases. In the testing phase, the model simulates real-world clinical conditions in which the fibrosis status remains unknown to the healthcare provider at the time of prediction. This section outlines the proposed methodology for screening and monitoring fibrosis risk in NAFLD patients intended for CDSS integration. Initially, the dataset used in this study is detailed, highlighting its composition and source. Following the dataset overview, we describe the data preprocessing strategies in depth: specifically, the patient and predictor selection criteria are central to our methodology. The patient selection process is designed to incorporate varied assessment strategies based on historical FIB-4 values, which allow for tailored evaluations of fibrosis progression risks, categorized into Case 1 and Case 2 configurations based on different levels of risk stringency. Additionally, we explore identifying a minimal and optimal number of predictors: this step is critical to enhance the CDSS's cost-effectiveness without sacrificing diagnostic accuracy. To assess the reliability of the proposed methodology and the robustness across different patient populations, we further introduce an independent dataset used for external evaluation. Subsequently, we present the ML technique employed for fibrosis-risk prediction, including its configuration and rationale. The proposed model is also compared to traditional classification algorithms to evaluate relative performance. Finally, we describe the experimental procedure and detail the

various experiments designed to assess model effectiveness, interpretability, and sensitivity to different clinical scenarios.

3.1 Dataset Description

This retrospective study uses a dataset of 595 Italian EHRs from general practice, initially including 1, 272, 572 patients (mean age = 59.27 years, with a SD = 15.89 years). Data were collected over 15 years (2008–2023), capturing routine clinical records with varying observation windows that reflect real-world, heterogeneous follow-up durations. The dataset is organized into three different structured fields.

(a) *Demographics* stores anonymized identifiers (patient ID), along with gender and year of birth for each patient.
(b) *Pathology* includes patient IDs, corresponding ICD-9 diagnostic codes, and their registration dates.
(c) *Lab Tests* includes patient IDs, lab test codes, outcomes, and the respective dates of these entries.

Despite the transition to ICD-10 codes in 2013, this study continues to utilize ICD-9 codes for several reasons: firstly, to maintain analytical consistency across the long-span dataset from 2008 to 2023, which began under the ICD-9 standard; to ensure compatibility with prior studies and facilitate comparative analyses; and lastly, to avoid the complexities and potential inaccuracies involved in retroactively mapping older diagnoses to the more detailed ICD-10 system.

3.2 Data Preprocessing

This section comprehensively describes the preprocessing steps to achieve the final dataset configuration, including the criteria for selecting patients, the establishment of experimental configurations for both Case 1 and Case 2, the predictor's selection process and the optimal predictor's subset selection (i.e., Case 1_subset, Case 2_subset).

To classify patients into low and high fibrosis-risk groups, subsequently labeled as "control" and "fibrosis" patients, respectively, HSI and FIB-4 indexes are employed, considering well-established thresholds from clinical guidelines [13, 15, 16]. As detailed in Figure 2, the calculations for HSI and FIB-4 incorporate key variables, including **Body Mass Index (BMI)**, **Alanine Aminotransferase (ALT)**, **Aspartate Aminotransferase (AST)**, **Platelet Count (PLT)**, age, gender and ICD-9 diagnosis codes.

3.2.1 Patients' Selection Criteria. Figure 2 outlines the key steps for patients' selection starting from the whole dataset of #1, 272, 572 subjects to the final eligible cohorts of control and fibrosis patients, based on specific criteria.

(a) *Hepatic ICD9*: Exclusion of patients with at least one ICD9 hepatic code recorded throughout the clinical history.
(b) *HSI*: Inclusion of patients whose HSI can be calculated; in other words, inclusion of patients whose ALT, AST and BMI are available at least once simultaneously (i.e., in terms of time) and both gender and ICD9 codes are available throughout the clinical history;
(c) $HSI > 36$: Inclusion of patients whose HSI > 36 at least once throughout the clinical history (i.e., NAFLD patients).
(d) $TWOI \in [1, 15]$: Exclusion of patients with a **Time Windows of Interest (TWOIs)** length less than 1 year, based on the clinical assumption that such a short clinical history may enclose little predictive information and, most importantly, could introduce bias.

Following this selection, depending on the FIB-4 values (as in Figure 2(d)), two different configurations (Case 1 and Case 2) are established by defining the observational TWOI for categorizing both control and fibrosis patients, as depicted in Figure 3, where the final cohort of each experimental configuration is obtained by executing an additional criteria:

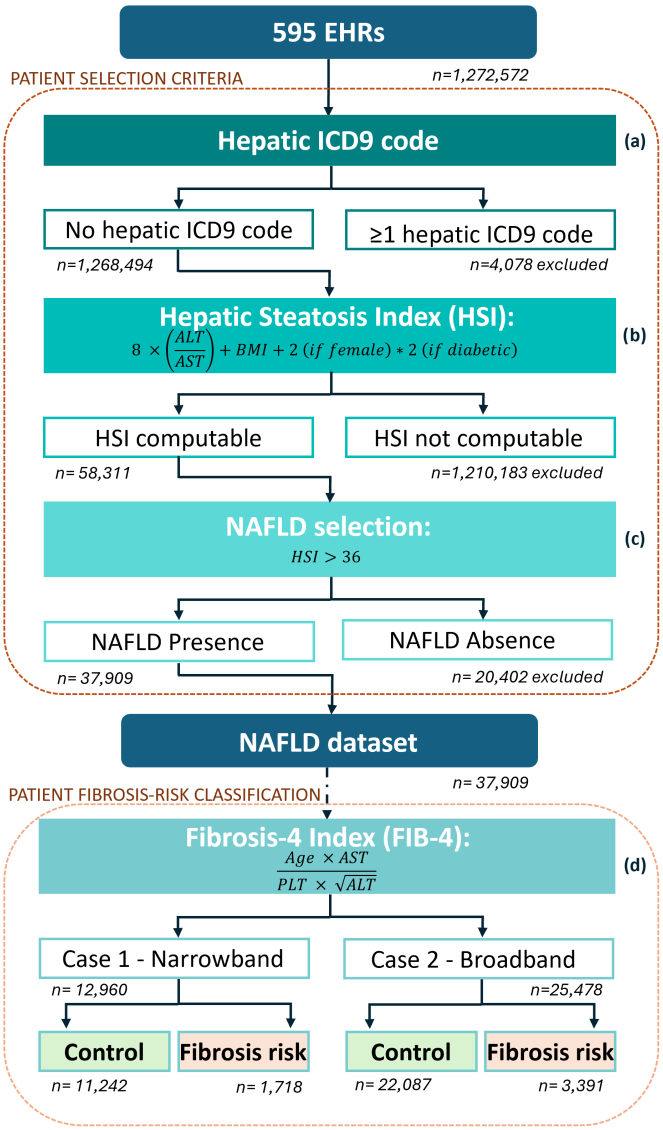
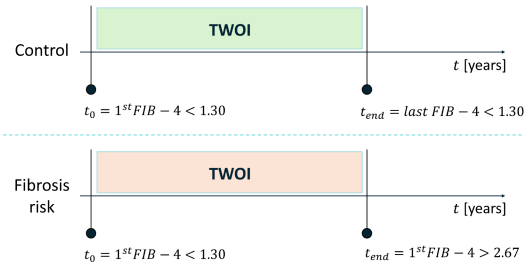
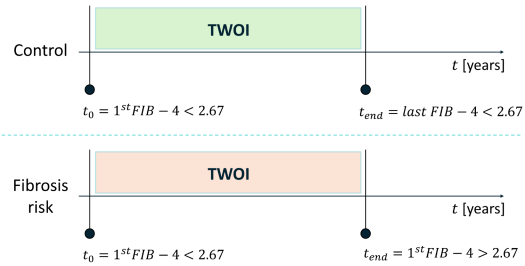


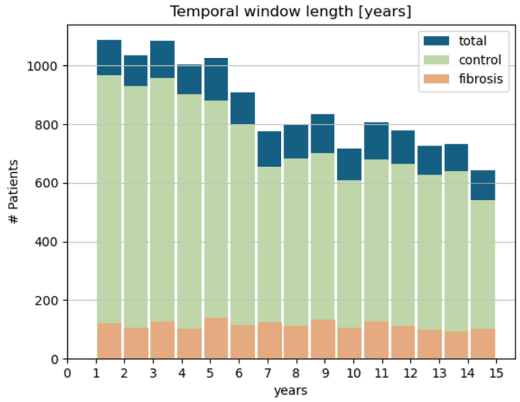
Fig. 2. Overview of patient selection process and fibrosis-risk stratification. Starting from the full dataset (n = 1,272,572), the selection criteria include: (a) exclusion of patients with at least one hepatic ICD9 code (a); (b) inclusion of patients with sufficient data to compute the HSI, which is based on ALT, AST, BMI, and PLT; (c) identification of the NAFLD cohort (n = 37,909) as those with HSI > 36; (d) fibrosis-risk stratification using the FIB-4, leading to the definition of control and at-risk groups under two temporal configurations: Case 1 (narrowband) and Case 2 (broadband). The HSI and FIB-4 formulas are based on established clinical criteria and incorporate age, sex, BMI, ALT, AST, PLT, and ICD9 codes. ALT, alanine aminotransferase; AST, aspartate aminotransferase; BMI, body mass index; PLT, platelet count



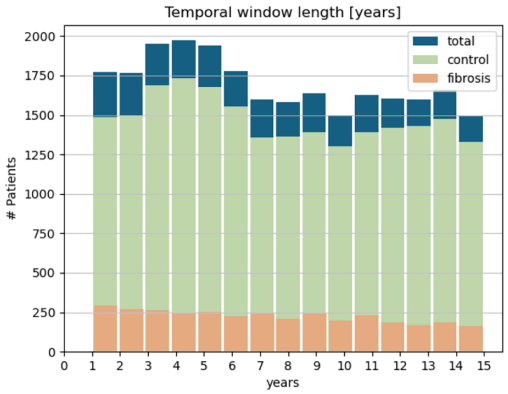
(a) Case 1 configuration. Control patients: TWOI ∈ [earliest FIB-4<1.30, latest FIB-4<1.30]; Fibrosis patients: TWOI ∈ [earliest FIB-4<1.30, earliest FIB-4>2.67].



(b) Case 2 configuration. Control patients: TWOI ∈ [earliest FIB-4<2.67, latest FIB-4<2.67]; Fibrosis patients: TWOI ∈ [earliest FIB-4<2.67, earliest FIB-4>2.67].



(c) Case 1 configuration. TWOI length distribution: The number of patients (i.e., control, fibrosis, total) per TWOI length in years. Patients with a TWOI length of less than 1 year are excluded from the study.



(d) Case 2 configuration. TWOI length distribution: The number of patients (i.e., control, fibrosis, total) per TWOI length in years. Patients with a TWOI length of less than 1 year are excluded from the study.

Fig. 3. Experimental configurations for fibrosis risk classification: Case 1—narrowband and Case 2—broadband, each defined by specific TWOIs.

Case 1—Narrowband Configuration. From the cohort of patients shown in Figure 2(c), in Case 1 (see Figure 3(a)), the TWOI for a control patient is defined as enclosed between the earliest FIB-4 < 1.30 and the latest FIB-4 < 1.30; while the TWOI of a fibrosis patient as enclosed between the earliest FIB-4 < 1.30 and the earliest FIB-4 > 2.67. This procedure returns 15,709 control patients and 2,486 patients with fibrosis. Patients with FIB-4 values between 1.30 and 2.67 are excluded from Case 1 (i.e., narrowband). Finally (see Figure 3(c)), only the patients with a TWOI length enclosed in the 1–15 year range are included in the study (i.e., # 11,242 control patients; # 1,718 fibrosis patients). The final cohort for the Case 1 configuration comprised 12,960 patients.

Case 2—Broadband Configuration. From the cohort of patients shown in Figure 2(c), in Case 2 (see Figure 3(b)), the TWOI of a control patient is defined as enclosed between the earliest FIB-4 < 2.67 and the latest FIB-4 < 2.67; while the TWOI of a fibrosis patient as enclosed between the earliest FIB-4 < 2.67 and the earliest FIB-4 > 2.67. This procedure returns 32,375 control patients and 4,511 patients with fibrosis. Patients with FIB-4 values between

1.30 and 2.67 are included in Case 2 (i.e., broadband). Finally (see Figure 3(d)), only the patients with a TWOI length enclosed in the 1–15 year range are included in the study (i.e., # 22,087 control patients; # 3,391 fibrosis patients). The final cohort for the Case 2 configuration comprised 25,478 patients.

In both cases, at the lower limit of TWOI, all patients (i.e., control and fibrosis patients) always start from a normal condition (i.e., FIB-4 < 1.30 (Case 1—narrowband) or FIB-4 < 2.67 (Case 2—broadband) and ICD-9 hepatic codes excluded), by simulating a realistic screening scenario. After that, moving forward in the clinical history, based on the FIB-4 index monitoring, some of them continue to be normal (i.e., control patients) until the end of the TWOI, while others (i.e., fibrosis patients) develop a high fibrosis risk (i.e., earliest FIB-4 > 2.67 coincided with the TWOI upper limit). It is worth noting that Case 1 represents a more specific subset of Case 2, based on stricter patient selection criteria defined by lower FIB-4 thresholds. Specifically, Case 1 includes control patients whose FIB-4 values remain consistently below 1.30, and fibrosis patients whose risk progresses from FIB-4 < 1.30 to FIB-4 > 2.67. In contrast, Case 2 adopts broader thresholds, including control patients with FIB-4 consistently below 2.67 and fibrosis patients whose risk progresses from FIB-4 < 2.67 to FIB-4 > 2.67.

3.2.2 Predictors' Selection Criteria. For both Case 1 and Case 2 configurations, the predictors are determined by computing the mean and the variation (Δ) of all laboratory test values within the defined TWOI for both control and fibrosis patients. Due to the irregular and sparse nature of lab test prescriptions by GPs, observations within the TWOI are often unevenly distributed. Consequently, using mean and variation offers a robust and generalizable strategy for feature extraction under these constraints. The mean and Δ values are computed by considering lab test values within the defined TWOI, thus using only data preceding the diagnostic event (i.e., the first FIB-4> 2.67 for fibrosis patients). This strategy ensures a realistic screening setting, where the model simulates early risk assessment based solely on past and routinely available clinical data.

Additionally, the following predictors' selection criteria are established:

(a) inclusion of all laboratory exams with missingness lower than 90% of the total patient count for both Case 1 (# predictors: 82) and Case 2 (# predictors: 86) configurations;
(b) inclusion of the variation (Δ) of each laboratory exam value, defined as the difference between the last and the first laboratory exam values within the TWOI;
(c) inclusion of both BMI and HSI values associated with their respective Δ and the presence of diabetes and gender as categorical predictors.

The total predictors collected consist of 86 predictors (i.e., 82 laboratory exams, BMI, HSI, diabetes, and gender) for Case 1 configuration and 90 predictors (86 laboratory exams, BMI, HSI, diabetes, and gender) for Case 2 configuration.

Table 1 shows the list of all predictors and their associated missingness percentage after the predictors' selection criteria. It is worth noting that the features (i.e., Age, ALT, AST, PLT) used to calculate FIB-4 (see Figure 2) are not included among the predictors to evaluate the fibrosis-risk profile.

To manage missingness, we adopted a conservative imputation strategy. We assigned a fixed out-of-range value to missing entries, which was also explicitly passed to the model via its missing hyperparameter. This approach leverages XGB's native mechanism to learn optimal split directions for missing values, treating missingness as a potentially informative structural pattern rather than as noise to be imputed. Unlike omics or ICU datasets, where missingness often follows a non-random mechanism and may carry clinical meaning, EHR data collected in general practice are typically affected by irregular test prescriptions driven by heterogeneous and often unrelated factors such as clinical discretion, cost considerations, or administrative lapses. Consequently, most imputation methods rely on the assumption that EHR data are either Missing at Random or **Missing Completely at Random (MCAR)**. It is important to recognize, however, that these are idealized mechanisms and that real-world EHR data generally fall somewhere in between [3]. Although the MCAR assumption rarely holds perfectly, several studies [17, 41] have demonstrated that, under typical clinical acquisition settings, MCAR can serve as a reasonable and

Table 1. List of Predictors and Associated Missingness (Expressed in Percentage [%]) for Case 1 and Case 2 Configurations

#	Predictors	Case 1 [%]	Case 2 [%]	#	Predictors	Case 1 [%]	Case 2 [%]
1	Albumin/Globulin Ratio (A/G)	75.08	73.18	46	International Normalized Ratio (INR)	80.91	78.39
2	Albumin (%)	71.27	69.16	47	Iron [S]	53.32	53.33
3	Alkaline Phosphatase [P]	67.28	63.93	48	Ketones	67.25	65.26
4	Alpha-1 Globulin (%)	71.37	69.33	49	Lactate Dehydrogenase (LDH)	85.82	83.52
5	Alpha-2 Globulin (%)	71.01	68.93	50	Low-Density Lipoprotein (LDL Cholesterol)	16.52	14.06
6	Amylase	88.63	88.00	51	Leukocyte Esterase	89.68	88.44
7	Anti-Thyroid Peroxidase Antibodies (AbTPO)	88.99	89.97	52	Leukocytes	65.74	64.62
8	Glomerular Filtrate	54.26	50.28	53	Leukocyte Count (n°)	4.24	4.47
9	Basophils (%)	43.19	43.54	54	Lymphocytes (%)	36.63	36.89
10	Basophils Count (n°)	47.75	47.17	55	Lymphocyte Count (n°)	46.80	45.85
11	Beta-1 Globulin (%)	75.78	73.71	56	Magnesium [S]	88.48	87.00
12	Beta-2 Globulin (%)	81.47	79.65	57	Mean Corpuscular Hemoglobin (MCH)	29.31	29.21
13	Bilirubin	65.49	63.65	58	Mean Corpuscular Hemoglobin Concentration (MCHC)	31.76	32.06
14	BMI	0.00	0.00	59	Mean Corpuscular Volume (MCV)	11.17	11.35
15	Calcium [S]	62.45	57.15	60	Microalbumin [U]	≥90	88.36
16	Chloride [S]	77.77	74.41	61	Microalbuminuria 24 h [U]	88.67	84.76
17	Vitamin B12 (Cobalamin) [S]	82.46	80.14	62	Monocytes (%)	39.68	40.12
18	C-Reactive Protein (CRP) [S]	53.84	51.83	63	Monocyte Count (n°)	46.98	46.34
19	Creatine Kinase (CK or CPK)	73.56	66.83	64	Neutrophils (%)	25.58	25.31
20	Creatinine [S]	3.91	2.89	65	Nitrites	≥90	89.03
21	Diabetes Diagnosis	0.00	0.00	66	Parathyroid Hormone (PTH) [S]	≥90	87.70
22	Direct Bilirubin [S]	69.65	67.89	67	pH (U)	41.44	40.00
23	Carcinoembryonic Antigen (CEA)	≥90	88.48	68	Phosphorus [S]	87.82	85.36
24	Eosinophil Count (n°)	47.19	46.72	69	Potassium [S]	30.12	24.04
25	Eosinophils (%)	41.49	41.80	70	Proteins	56.45	54.62
26	Erythrocytes	6.10	6.57	71	Psychosocial Status (PS)	42.82	41.88
27	Erythrocyte Sedimentation Rate (ESR)	49.92	48.41	72	Prostate-Specific Antigen (PSA) [S]	73.98	66.45
28	Ferritin [P]	62.31	62.61	73	PSA Free [S]	≥90	89.47
29	Folates [S]	81.47	79.04	74	Prothrombin Time (PT)	88.26	86.51
30	Free Triiodothyronine (FT3) [P]	72.79	73.62	75	Partial Thromboplastin Time (PTT)	85.79	85.10
31	Free Thyroxine (FT4) [P]	56.92	57.22	76	Red Cell Distribution Width (RDW)	51.30	50.74
32	Functional Fibrinogen	85.29	84.41	77	Oxygen Saturation	88.23	86.95
33	Gamma Globulin (%)	71.83	69.71	78	Sodium [S]	33.39	27.32
34	GGT [S]	14.00	12.73	79	Total and Fractionated Bilirubin [S]	85.93	85.69
35	Glucose [S]	3.88	3.07	80	Total Bilirubin [S]	60.92	59.04
36	Glucose [U]	61.38	59.55	81	Total Cholesterol	4.93	3.63
37	Glycated Hemoglobin (HbA1c) [B]	56.92	49.42	82	Total Proteinemia [S]	68.83	67.66
38	Glycated Hemoglobin (HbA1c) [mmol/mol]	84.13	80.65	83	Total Proteins	81.81	80.92
39	Hemoglobin (Hb)	0.93	1.35	84	Triglycerides [S]	5.39	4.00
40	Hematocrit (HCT)	9.98	10.66	85	Thyroid Stimulating Hormone (TSH) [P]	34.01	34.72
41	High-Density Lipoprotein (HDL Cholesterol)	6.95	5.20	86	TSH Reflex	84.95	85.45
42	Hb Total	67.68	64.80	87	Urea (Azotemia)	49.95	46.14
43	Homocysteine [S]	89.69	≥90	88	Uric Acid (Uricemia)	34.49	27.98
44	HSI	0.00	0.00	89	Urobilinogen	64.33	62.62
45	Indirect Bilirubin [S]	87.92	87.21	90	Vitamin D [B]	64.07	61.07
				91	Gender	0.00	0.00

The variation Δ of each predictor is omitted. [B], [Blood]; GGT, Gamma-Glutamyl Transferase; [P], [Plasma]; [S], [Serum]; [U], [Urine].

practically valid working hypothesis for modeling missingness. In light of this unpredictability, we adopted the simplifying assumption of MCAR to avoid introducing unverifiable statistical assumptions about the missingness mechanism. This choice preserves the integrity of the missingness pattern under MCAR and aligns with prior work on predictive modeling in sparse EHR contexts [5, 7, 26].

Optimal Predictors Subset Selection Criteria: Case 1_subset and Case 2_subset Configurations. We extracted global feature-importance scores using XGB's built-in mechanism from only the training folds within each iteration of the external 10-fold **Cross-Validation (CV)**. In particular, within each outer-training partition, feature importances were estimated on the nested training folds of the inner CV and averaged across these inner folds to obtain a stable ranking representative of the outer-training data. As a result, the subset of selected predictors could differ across outer folds. A predictor (either its mean value or its Δ) was included in the fold-specific optimal subset if its average importance exceeded 1% of the total cumulative feature importance computed in the inner CV, in either the Case 1 or Case 2 configuration. In such cases, the full pair (mean and Δ) was selected. This strategy ensures consistency with real-world clinical settings, where both the exam value and its variability

can be retrieved simultaneously from available laboratory data. Using a relative importance threshold (i.e., 1%), we selected only features contributing meaningfully to predictive performance. This choice avoids including low-importance features that may introduce noise or redundancy. Moreover, including a whole pair (mean and Δ), if at least one exceeds 1% of the total cumulative importance across all predictors, ensures clinical interpretability and consistency.

Table 2 presents the most discriminative and stable predictors identified for Case 1_subset and Case 2_subset configurations. The table reports those predictors whose average feature-importance values—computed by first averaging within each inner CV loop and subsequently across outer folds—exceeded 1% of the total cumulative importance.

Table 3 summarizes the whole statistics of the NAFLD dataset after the preprocessing stage for Case 1, Case 1_subset, Case 2, Case 2_subset configurations.

3.3 External Validation Dataset: FIMMG Dataset

We introduce a new independent and external dataset (FIMMG dataset) extracted from the standardized FIMMG Netmedica Cloud computing infrastructure [6]. It contains 10 years (2010-2019) of clinical history collected by 6 GPs in a specific regional area in Italy. The FIMMG dataset originally consists of 14,175 patients and six main fields.

- —The demographic field is composed of age and gender.
- —The monitoring field (i.e., diastolic and systolic blood pressure, height, weight, and waist) contains only continuous predictors and the lab tests field, where all the laboratory outcomes are stored.
- —The remaining three fields comprise pathologies (ICD-9 codes), drugs, and exam prescriptions.

We followed the same preprocessing criteria (see Sections 3.2.1 and 3.2.2) to build the final FIMMG dataset configuration used for fair and robust external validation. Table 4 shows the descriptive statistics of the FIMMG dataset after the preprocessing stage for two different configurations (i.e., Case 1_subset, Case and Case 2_subset).

It is worth noting that, to perform an external validation of the trained XGB model over the FIMMG dataset, it was not possible to replicate the full Case 1 and Case 2 configurations, as the FIMMG dataset contains only a part of the predictors listed in Table 1. However, the FIMMG dataset includes all the optimal subsets of predictors listed in Table 2. Thus, to ensure a fair and consistent external validation of the trained XGB model, the evaluation was conducted exclusively within the configurations employing the optimal set of predictors (i.e., Case 1_subset and Case 2_subset).

3.4 ML Approach

The ML core of the framework is represented by the XGB model, an ensemble learning method that combines the predictions of multiple weak DT classifiers to produce a more robust prediction. The experimental procedure includes an external 10-fold CV with a nested 5-fold CV for hyperparameter tuning. The overall setup is available for review and replication through our publicly accessible repository.[1] The model's hyperparameters are tuned in the nested 5-fold CV through a grid search procedure ($n\_estimators : [75, 100, 150, 200]$; $max\_depth : [6, 25, 50, 75]$; $eta : [0.05, 0.1, 0.2, 0.3]$) by maximizing the macro-Recall score. Unlike standard Recall, macro-Recall averages the recall of each class, thereby avoiding bias toward the majority class and promoting a fair tradeoff in Sensitivity for both fibrosis detection and control discrimination. This choice aligns with the state-of-the-art [26] and the clinical goal of minimizing false negatives while ensuring that model performance is not skewed by class imbalance.

During model training, the **Synthetic Minority Oversampling TEchnique (SMOTE)** was used to mitigate the high-class imbalance ratio, also evident in Table 3. Standard SMOTE operates over the entire feature space using

[1] https://github.com/michelebernardini/fibrosis-risk-prediction.

Table 2. Stable and Discriminative Predictors Identified through Nested CV

#	Predictors	Importance [%]			
		Case 1_subset		Case 2_subset	
		Mean	Δ	Mean	Δ
1	Glomerular Filtrate	1.68	<1.00	1.13	<1.00
2	BMI	1.87	1.12	1.95	1.05
3	Creatinine [S]	1.62	<1.00	1.58	1.08
4	Erythrocytes	1.35	1.05	1.42	1.22
5	ESR	1.20	<1.00	1.01	<1.00
6	GGT [S]	2.03	2.05	2.19	1.76
7	Glucose [S]	1.90	1.11	1.66	1.13
8	HbA1c [B]	<1.00	<1.00	1.17	<1.00
9	Hb	1.47	1.02	1.38	1.07
10	HCT	1.22	1.27	1.35	1.30
11	HSI	3.27	2.58	2.44	2.04
12	HDL Cholesterol	1.20	<1.00	1.37	1.15
13	LDL Cholesterol	1.69	1.12	1.70	1.18
14	Leukocytes n°	2.59	1.60	2.14	1.48
15	MCV	1.62	1.23	1.52	1.32
16	Neutrophils %	1.04	<1.00	1.02	<1.00
17	Potassium [S]	1.24	1.06	1.40	1.03
18	Sodium [S]	<1.00	<1.00	1.18	<1.00
19	Total Cholesterol	1.27	1.29	1.34	1.13
20	Triglycerides [S]	1.17	1.00	1.42	1.17
21	TSH [P]	1.24	<1.00	1.24	<1.00
22	Urea	1.18	<1.00	1.19	<1.00
23	Uricemia	1.49	<1.00	1.45	1.12

Optimal subset of predictors (i.e., mean values and corresponding variations (Δ) expressed in percentage [%]) along with their relative feature-importance scores, as determined by XGB's built-in feature-importance mechanism. Results are reported for both the Case 1_Subset and Case 2_Subset configurations. A predictor (i.e., either the mean value or its Δ) was included in the optimal subset if its average feature-importance value—computed by first averaging within each inner CV loop and subsequently across outer folds—Exceeded 1% of the total cumulative feature importance in either the Case 1 or Case 2 configuration. In such cases, the full pair (mean and Δ) was selected to ensure consistency and clinical interpretability. Δ is defined as the difference between the last and the first laboratory exam values within the observational TWOI. [B], [Blood]; [P], [Plasma]; [S], [Serum]; [U], [Urine].

the extra-value–imputed dataset, where the fixed out-of-range constant enables uniform distance computation despite the presence of missing values. However, applying SMOTE after extra-value imputing does not fully preserve the original missingness structure in the generated synthetic samples and may slightly distort distances in the feature space. For this reason, we also evaluated the **SMOTE for Nominal and Continuous (SMOTE-NC)**,

Table 3. NAFLD Dataset Statistics after the Preprocessing Stage for the Different Configurations: Case 1, Case 1_Subset, Case 2, and Case 2_Subset

NAFLD dataset	Case 1	Case 1_subset	Case 2	Case 2_subset
# Patients:	12,960	as Case 1	25,478	as Case 2
# Control:	11,242	as Case 1	22,087	as Case 2
Gender (#M/#F)	0.80	as Case 1	0.90	as Case 2
Diabetes %	10.02	as Case 1	14.66	as Case 2
Mean BMI ± SD; mean Δ ± SD	29.15 ± 5.33; −0.05 ± 2.40	as Case 1	29.25 ± 5.07; −0.06 ± 2.35	as Case 2
Mean HSI ± SD; mean Δ ± SD	41.49 ± 6.94; −0.10 ± 11.50	as Case 1	40.70 ± 6.34; −0.61 ± 10.26	as Case 2
#Fibrosis:	1,718	as Case 1	3,391	as Case 2
Gender (#M/#F)	1.27	as Case 1	1.31	as Case 2
Diabetes %	24.91	as Case 1	25.51	as Case 2
Mean BMI ± SD; mean Δ ± SD	28.35 ± 4.68; −0.57 ± 2.68	as Case 1	28.81 ± 4.62; −0.65 ± 2.68	as Case 2
Mean HSI ± SD; mean Δ ± SD	38.79 ± 5.73; −3.53 ± 8.31	as Case 1	38.53 ± 5.44; −2.69 ± 7.34	as Case 2
Prevalence %	13.26	as Case 1	13.31	as Case 2
# Predictors	86	21	90	23
# Predictors + Δ	170	42	178	46

The statistics, breakdown by condition (i.e., Control, Fibrosis), include the total number of patients, the ratio between the number of male and female patients, the prevalence [%] of fibrosis patients, the presence [%] of diabetes diagnosis, the mean BMI ± SD and the mean associated Δ ± SD, the mean HSI ± SD and the mean associated Δ ± SD. The count of predictors and extended predictors (Predictors + Δ) is reported for each experimental configuration. Δ is defined as the difference between the last and the first laboratory exam values within the observational TWOI.

Table 4. Descriptive Statistics of the External Validation Dataset, Named FIMMG Dataset, Used in Case 1_Subset and Case 2_Subset Configurations

FIMMG dataset	Case 1_subset	Case 2_subset
# Patients:	1,066	2,060
# Control:	927	1,790
Gender (#M/#F)	0.83	0.89
Diabetes %	8.63	13.46
Mean BMI ± SD; mean Δ ± SD	29.12 ± 5.50; 0.05 ± 2.35	29.17 ± 5.10; −0.02 ± 2.25
Mean HSI ± SD; mean Δ ± SD	41.40 ± 6.60; −0.20 ± 6.43	40.64 ± 6.14; −0.66 ± 6.01
# Fibrosis:	139	270
Gender (#M/#F)	1.20	1.29
Diabetes %	23.02	24.07
Mean BMI ± SD; mean Δ ± SD	28.86 ± 4.49; −0.68 ± 2.74	29.43 ± 4.54; −0.77 ± 2.72
Mean HSI ± SD; mean Δ ± SD	39.31 ± 5.24; −2.82 ± 8.58	39.29 ± 5.56; −2.40 ± 7.54
Prevalence %	13.04	13.11
# Predictors	21	23
# Predictors + Δ	42	46

which is specifically designed to handle mixed-type data by also restricting distance computations to features without missing values.

Additionally, the feature-importance strategy used by the XGB model takes into account the frequency of a feature's involvement in data splits across all trees, emphasizing its relevance to the model's decision-making process. This metric provides a global view of the relevance of each feature in the model's prediction. The higher

the frequency, the more the model relied on that feature when constructing the boosted DTs. While model-agnostic interpretability methods (e.g., **Shapley Additive Explanations (SHAP)**) provide instance-level insights, we prioritized a global feature analysis to identify predictors most consistently associated with fibrosis risk across the entire dataset. This decision aligns with the clinical objectives of our study, which emphasize the identification of globally discriminative features over individual-level explanations.

3.5 Experimental Procedure

To evaluate the performance of the proposed ML approach, compare it with other traditional ML models, and assess all associated experiments, we employed standard metrics such as Accuracy, macro-Precision (Precision), macro-Recall (Recall), macro-F1 score (F1), AUC, the area under the **Precision-Recall Curve (PRAUC)**, Specificity, and Sensitivity. Experimental comparisons of the proposed XGB approach for both Case 1 and Case 2 configurations were conducted to evaluate predictive performance and model interpretability through feature-importance analysis. Specifically, we compared:

- —several traditional ML models, including LR, DT, RF, **K-Nearest Neighbor (KNN)**, and SVM (see Section 4.1);
- —a reduced set of predictors and their associated Δ values (i.e., Case 1_subset and Case 2_subset), selected based on the most informative features identified by the XGB model (see Section 4.2).

Additionally, to conduct an external validation, the XGB model was trained over the NAFLD dataset (see Table 3), and then, the exported trained XGB model was finally evaluated over the FIMMG dataset (see Table 4) for both Case 1_subset and Case 2_subset configurations. This external validation is reported in Section 4.3.

To further assess the fairness and robustness of the proposed model, we performed a stratified analysis by demographic subgroups. Specifically, we evaluated performance across three age categories (Age < 65, $65 \leq$ Age ≤ 74, and Age ≥ 75) and by gender, following prior literature [13, 15, 16]. This analysis helps ensure that predictive accuracy is consistent across clinically relevant populations and is not biased by age or sex. The fairness analysis is reported in Section 4.4.

We also conducted a sensitivity analysis to evaluate how different imputation methods, oversampling techniques, and feature importance criteria affect model performance and interpretability (see Section 4.5). To further assess the effectiveness of different imputation strategies to manage the MCAR mechanism, we conducted internal experiments to compare our extra-value imputation approach against three state-of-the-art strategies (see Section 4.5.1): (i) mean imputation, (ii) KNN imputation [24], and (iii) **Multivariate Imputation by Chained Equations (MICE)** [22]. We also evaluated an alternative oversampling technique, such as SMOTE-NC, to assess its impact on model performance in comparison to standard SMOTE (see Section 4.5.2). Although not the focus of our feature-importance analysis, we additionally report, for comparison (see Section 4.5.3): (i) permutation importance, a global model-agnostic metric that quantifies performance degradation when feature values are shuffled; and (ii) SHAP values, a local interpretability method that attributes feature contributions to individual predictions.

4 Results

This section presents the results of a comprehensive set of experiments to evaluate the predictive performance and interpretability of the proposed XGB approach. We compare XGB with several ML models (see Section 4.1) and assess its performance using reduced predictor subsets derived from XGB feature importance (see Section 4.2). External validation on the FIMMG dataset is reported in Section 4.3). To evaluate fairness and robustness, we perform a stratified analysis across demographic subgroups by age and gender (see Section 4.4). Finally, a sensitivity analysis explores the impact of different imputation methods, oversampling strategies, and feature importance criteria (see Section 4.5).

Table 5. Results for Case 1 Configuration: Performance of the Proposed Model and Comparisons with Other Traditional ML Models

Case 1	Accuracy	F1	Precision	Recall$_{CI}$	AUC$_{CI}$	PRAUC$_{CI}$	Specificity	Sensitivity
LR	82.70	70.85	68.16	$78.00_{75.54-80.28}$	$85.64_{82.50-88.33}$	$56.40_{49.29-62.37}$	84.42	71.54
DT	82.40	65.33	64.20	$67.10_{65.33-68.85}$	$67.10_{64.26-69.80}$	$45.17_{41.07-49.69}$	87.92	46.21
RF	88.04	76.56	75.50	$79.60_{78.14-81.03}$	$89.55_{88.22-90.96}$	$64.40_{61.14-67.90}$	91.14	67.99
KNN	76.50	61.23	60.16	$66.00_{64.55-67.61}$	$66.00_{63.32-68.63}$	$43.43_{39.91-46.58}$	80.31	51.75
SVM	80.95	69.98	70.38	$69.25_{67.79-71.05}$	$79.92_{78.85-80.99}$	$53.65_{50.79-56.51}$	79.25	63.56
XGB	90.23	80.76	78.40	$84.10_{82.98-85.20}$	$\mathbf{92.97_{91.81-93.93}}$	$75.44_{71.66-78.34}$	92.49	75.67

Predictive performances are reported in Accuracy, F1, Precision, Recall, AUC, PRAUC, Specificity, and Sensitivity. Each metric, expressed in percentage, represents the average value obtained from a 10-fold CV. The Recall, AUC, and PRAUC are reported along with their corresponding CIs. The method achieving the highest AUC is highlighted in bold.

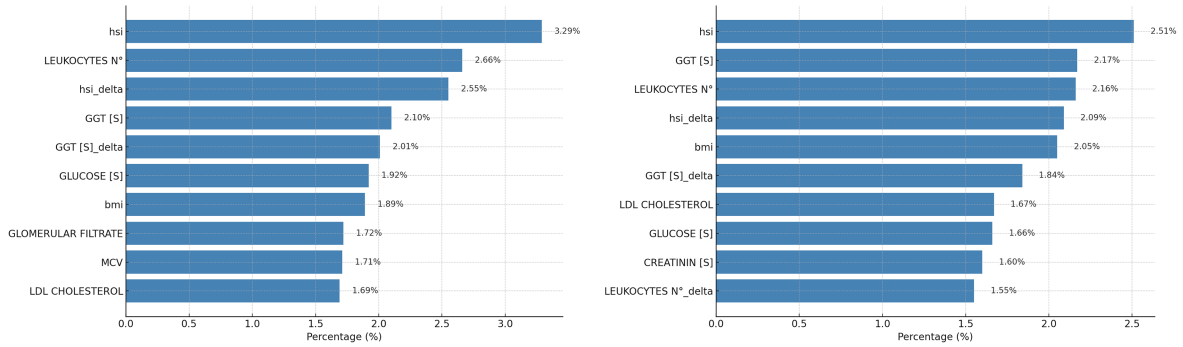
Table 6. Results for Case 2 Configuration: Performance of the Proposed Model and Comparisons with All Other Traditional ML Models

Case 2	Accuracy	F1	Precision	Recall$_{CI}$	AUC$_{CI}$	PRAUC$_{CI}$	Specificity	Sensitivity
LR	78.00	65.33	63.60	$72.80_{71.79-73.95}$	$80.20_{78.55-81.95}$	$45.46_{42.14-48.97}$	79.84	59.08
DT	78.10	59.70	58.84	$61.80_{60.79-62.91}$	$61.80_{59.71-63.57}$	$37.65_{34.79-40}$	83.97	39.69
RF	85.40	70.20	69.20	$71.60_{70.68-72.44}$	$82.96_{81.85-84.06}$	$50.54_{48.38-52.88}$	90.38	52.79
KNN	73.73	58.15	57.70	$62.74_{61.33-64.22}$	$62.74_{60.57-64.97}$	$39.72_{36.95-42.49}$	77.74	47.71
SVM	76.01	61.67	62.38	$60.32_{58.63-61.99}$	$74.14_{73.07-75.21}$	$48.41_{45.55-51.27}$	77.85	59.96
XGB	86.80	73.54	72.00	$75.50_{74.44-76.41}$	$\mathbf{86.96_{86.43-88.00}}$	$59.30_{56.67-61.03}$	90.93	60.10

Predictive performances are reported in Accuracy, F1, Precision, Recall, AUC, PRAUC, Specificity, and Sensitivity. Each metric, expressed in percentage, represents the average value obtained from a 10-fold CV. The Recall, AUC, and PRAUC are reported along with their corresponding CIs. The method achieving the highest AUC is highlighted in bold.

4.1 Traditional ML Models

Tables 5 and 6 present comprehensive results obtained through Case 1 and Case 2 configurations, respectively. Across both experimental scenarios, the XGB model's predictive performance always remains superior to that of the other traditional ML models. Specifically, while the overall predictive performance of the XGB model in Case 1 surpasses that of Case 2, there are nuanced differences between the cases. While Accuracy and Specificity are comparable (Case 1: Accuracy = 90.23%, Specificity = 92.49%; Case 2: Accuracy = 86.80%, Specificity = 90.93%), there is a pronounced disparity in terms of Recall and Sensitivity (Case 1: Recall = 84.10%, Sensitivity = 75.67%; Case 2: Recall = 75.50%, Sensitivity = 59.98%). Notably, the highest AUC = 92.97% is achieved by Case 1, while Case 2 achieves AUC = 86.96%. The RF model emerges as the most robust alternative among competitors, with AUC = 89.55% for Case 1 and AUC = 82.96% for Case 2. Statistical tests are used to benchmark the performance of the XGB model against chance levels. Moreover, we focused on analyzing the statistical gain regarding Recall and AUC between the XGB and other traditional ML competitors. Since all metrics had a distribution departing from normality, we used non-parametric Wilcoxon signed-rank tests on the per-fold performance metrics (i.e., paired metric values across the 10-fold CV) between the compared models. Given five pairwise comparisons (five competitors), we applied a Bonferroni correction, setting the adjusted significance threshold to $\alpha = 0.01$. The performance in terms of Accuracy, F1, Precision, Recall, Specificity, Sensitivity, AUC, and PRAUC is significantly better ($p < 0.01$) compared to chance levels (0.5) for both Case 1 and Case 2. In both Case 1 and Case 2, the AUC and Recall of the XGB model are significantly greater ($p < 0.01$) than all other competitors.



(a) Case 1 configuration.

(b) Case 2 configuration.

Fig. 4. Top-10 features importance.

Figure 4 shows the top-10 features importance results for both Case 1 (Figure 4(a)) and Case 2 (Figure 4(b)) configurations.

From this analysis, the HSI values emerge as the most crucial predictors for both experimental configurations, though without overwhelmingly dominating over other features. The number of leukocytes and **Gamma-GlutamylTransferase (GGT [S])** consistently appear among the top-ranking positions, highlighting their significance in assessing fibrosis risk. Moreover, it is particularly insightful to observe that the changes in these variables over time (Δ values for HSI, leukocyte number, and GGT [S]) are also prominent among the top predictors. These temporal variations are critical for the XGB model's ability to accurately discern variations in the fibrosis-risk profile, indicating that dynamic changes in these markers highly indicate evolving clinical conditions. Furthermore, the consistent appearance of these critical predictors across both Case 1 and Case 2 configurations underscores the reliability of the XGB model in discovering and utilizing clinically relevant features to predict fibrosis risk effectively.

4.2 Optimal Subset of Predictors: Case 1_Subset and Case 2_Subset Configurations

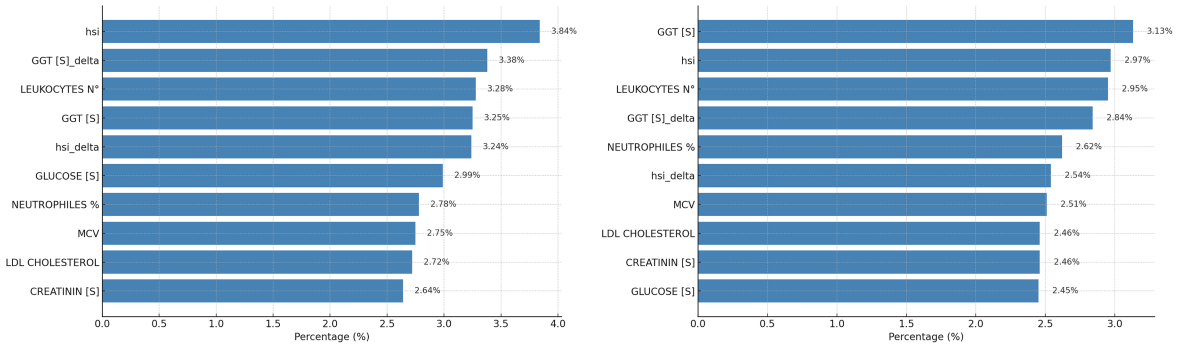
Moreover, we investigate the generalization performance of the optimal subset of predictors (see Section 3.5 and Table 2). This step is crucial for enhancing the CDSS's cost-effectiveness without compromising diagnostic accuracy, highlighting the potential of XGB with fewer features and reducing the burden of annotating all examinations for clinicians. Table 7 compares the XGB predictive performance results for both Case 1 and Case 2 configurations (see Tables 5 and 6) with respect to the predictive performance results obtained from both Case 1_subset and Case 2_subset configurations. In Case 1_subset, AUC decreases from 92.97% to 91.00%, while Sensitivity increases from 75.67% to 79.63%. In Case 2_subset, AUC decreases from 86.96% to 83.84%, while Sensitivity rises from 60.10% to 66.82%. The predictive performance variations in terms of AUC and Recall appear larger in Case 2 than in Case 1. There is no statistical difference in Recall between Case 2_subset and Case 2 ($p = 0.278$), thus highlighting a good tradeoff between the model simplicity and generalization performance.

The top-10 feature importance results for both Case 1_subset and Case 2_subset configurations are listed in Figure 5. In Case 1_subset, the HSI value remains the most important predictor, but now the GGT [s]_delta appears more critical than its associated GGT [S] value. BMI and glomerular filtrate exit from top-10 features importance ranking, while neutrophils % (# 7) and creatinine [S] (# 10) enter. In Case 2_subset, the GGT [S] value becomes the most important predictor, followed by HSI and leukocytes n°. BMI and leukocytes n°_delta exit from the top-10 features importance ranking, while neutrophils % (#5) and mcv (#7) enter in.

Table 7. Results for Case 1 and Case 2 Configurations along with Their Optimal Predictor Subsets, Case 1_Subset and Case 2_Subset, Respectively

XGB model	Accuracy	F1	Precision	Recall_{CI}	AUC_{CI}	PRAUC_{CI}	Specificity	Sensitivity
Case 1	90.23	80.76	78.40	$84.10_{82.98-85.20}$	$\mathbf{92.97_{91.81-93.93}}$	$75.44_{71.66-78.34}$	92.49	75.67
Case 1_subset	85.64	75.40	72.00	$83.10_{82.31-83.81}$	$91.00_{89.71-91.95}$	$70.20_{67.96-71.27}$	86.55	79.63
Case 2	86.80	73.54	72.00	$75.50_{74.44-76.41}$	$\mathbf{86.96_{86.43-88.00}}$	$59.30_{56.67-61.03}$	90.93	60.10
Case 2_subset	81.25	68.60	66.26	$75.10_{74.27-75.99}$	$83.84_{82.42-84.89}$	$51.60_{50.60-52.82}$	83.45	66.82

Predictive performances are reported in Accuracy, F1, Precision, Recall, AUC, PRAUC, Specificity, and Sensitivity. Each metric, expressed in percentage, represents the average value obtained from a 10-fold CV. The Recall, AUC, and PRAUC are reported along with their corresponding CIs. The configuration achieving the highest AUC is highlighted in bold.



(a) Case 1_subset configuration. (b) Case 2_subset configuration.

Fig. 5. Top-10 features importance.

Table 8. External Validation Experimental Results over the FIMMG Dataset Using the Optimal Subset of Predictors (i.e., *Case 1_Subset** and *Case 2_Subset**)

XGB model	Accuracy	F1	Precision	Recall	AUC	PRAUC	Specificity	Sensitivity
Case 1_subset	85.64	75.40	72.00	83.10	91.00	70.20	86.55	79.63
*Case 1_subset**	87.99	77.82	74.54	83.62	91.09	69.58	89.54	77.70
Case 2_subset	81.25	68.60	66.26	75.10	83.84	51.60	83.45	66.82
*Case 2_subset**	79.42	63.52	62.00	67.40	77.71	34.89	83.69	51.11

Predictive performances are reported in Accuracy, F1, Precision, Recall, AUC, PRAUC, Specificity, and Sensitivity. For the NAFLD dataset, each metric, expressed in percentage, represents the average value obtained from a 10-fold CV. The Recall, AUC, and PRAUC are reported along with their corresponding CIs.

4.3 External Validation: *Case 1_Subset** and *Case 2_Subset** Configurations

Table 8 compares the predictive performance of the XGB model using the optimal subset of predictors under both 10-fold CV and external validation experimental procedures. From now on, to distinguish the external validation experimental results from those obtained through the 10-fold CV using the optimal subset of predictors, we will refer to the external validation experimental results in italics and with an asterisk (i.e., *Case 1_subset**, *Case 2_subset**).

Table 9. Performance Metrics Stratified by Age [Years] and Gender: Case 1 and Case 2 Configurations for XGB Model

Case 1	Accuracy	F1	Precision	Recall	AUC	PRAUC	Specificity	Sensitivity	Count [#]	Count
Age < 65	93.35	71.27	67.65	77.61	89.41	50.65	94.99	60.23	9,220	71.14
65 ≤ Age ≤ 74	84.95	79.38	77.68	81.93	90.83	77.21	87.24	76.62	2,226	17.18
Age ≥ 75	79.26	79.08	79.33	78.98	87.71	88.69	74.51	83.46	1,514	11.68
Male	89.50	82.52	79.94	86.26	93.89	80.43	91.04	81.48	5,950	45.91
Female	90.91	78.36	76.31	80.97	91.75	68.28	93.65	68.30	7,010	54.09
Case 2	**Accuracy**	**F1**	**Precision**	**Recall**	**AUC**	**PRAUC**	**Specificity**	**Sensitivity**	**Count [#]**	**Count**
Age < 65	93.17	71.48	68.78	75.45	88.16	50.64	95.30	55.59	12,350	48.47
65 ≤ Age ≤ 74	86.16	75.08	74.24	76.06	86.43	63.64	90.87	61.24	6,372	25.01
Age ≥ 75	75.80	69.70	68.94	70.89	79.25	60.81	80.86	60.93	6,756	26.52
Male	84.88	74.21	72.22	77.26	87.39	63.35	88.32	66.20	12,391	48.63
Female	88.64	72.08	71.63	72.56	86.10	53.43	93.29	51.84	13,087	51.37

All Metrics Are Expressed in Percentage [%].

*Case 1_subset** results closely comparable with the Case 1_subset, indicating strong generalizability of the trained XGB. AUC is nearly unchanged (91.00% vs. 91.09%), while PRAUC shows only a modest reduction (70.20% to 69.58%). Specificity improves (from 86.55% to 89.54%), whereas Sensitivity slightly decreases (from 79.63% to 77.70%).

On the contrary, *Case 2_subset** results are less comparable with the Case 2_subset, by observing a more evident degradation in predictive performance when transitioning to externally validate the trained XGB model on a new cohort of patients. AUC decreases from 83.84% to 77.71%, while PRAUC exhibits a marked decline from 51.60% to 34.89%, as well as the Sensitivity from 66.82% to 51.11%, indicating reduced ability to identify fibrosis cases correctly. Notably, Specificity remains almost unchanged (83.45% vs. 83.69%).

4.4 Fairness Analysis

We have stratified the experimental results by three ranges of age (i.e., Age < 65; 65 ≤ Age ≤ 74; Age ≥ 75) [13, 15, 16] and by gender. Table 9 refers to Case 1 and Case 2 configurations, while Table 10 refers to Case 1_subset and Case 2_subset configurations. The stratified analysis shows that model performance remained consistently high (AUC > 79.25 for Case 1 and Case 2 and AUC > 75.15 for Case 1_subset and Case 2_subset) and relatively uniform across age groups and gender, with only minor variations confirming the robustness and demographic fairness of the proposed approach.

4.5 Sensitive Analysis

In this section, we present the results of a sensitive analysis assessing how different imputation methods, missingness thresholds, oversampling techniques and feature importance criteria affect model performance and interpretability.

4.5.1 Comparison of Imputation Strategies and Missingness Thresholds. The comparative results with state-of-the-art data imputation strategies are summarized in Table 11. Our extra-value imputation consistently outperforms all alternative strategies across all configurations, achieving the highest AUC in each setting. This outcome confirms our approach's robustness and practical effectiveness in handling missing laboratory data within real-world EHRs. Aligned with the MCAR assumption, our data imputation strategy preserves the original

Table 10. Performance Metrics Stratified by Age and Gender: Case 1_Subset and Case 2_Subset Configurations for XGB Model

Case 1_subset	Accuracy	F1	Precision	Recall	AUC	PRAUC	Specificity	Sensitivity	Count [#]	Count
Age < 65	89.08	65.40	61.77	78.75	86.48	43.36	90.15	67.36	9220	71.14
65 ≤ Age ≤ 74	78.35	73.25	71.70	79.16	87.69	71.82	77.73	80.58	2226	17.18
Age ≥ 75	75.36	74.74	76.24	74.68	85.30	86.31	63.66	85.70	1514	11.68
Male	84.91	77.34	74.27	84.53	92.21	76.20	85.09	83.98	5950	45.91
Female	86.25	72.85	69.38	80.91	89.29	61.82	87.72	74.11	7010	54.09
Case 2_subset	Accuracy	F1	Precision	Recall	AUC	PRAUC	Specificity	Sensitivity	Count [#]	Count
Age < 65	89.43	65.94	62.54	75.32	85.61	42.65	91.13	59.52	12,350	48.47
65 ≤ Age ≤ 74	79.61	68.86	66.91	73.89	82.07	55.22	82.29	65.48	6372	25.01
Age ≥ 75	67.78	64.10	64.45	68.65	75.15	53.85	66.87	70.44	6756	26.52
Male	79.21	69.22	67.18	76.13	84.42	56.22	80.60	71.66	12,391	48.63
Female	83.15	67.34	64.93	73.25	82.67	45.08	86.01	60.49	13,087	51.37

All Metrics Are Expressed in Percentage [%].

Table 11. Comparison of Different Data Imputation Strategies: Experimental Results Obtained by the XGB Model in Case 1, Case 1_Subset, Case 2, and Case 2_Subset Configurations

Case 1	Accuracy	F1	Precision	Recall_{CI}	AUC_{CI}	PRAUC_{CI}	Specificity	Sensitivity
Extra-Value (Ours)	90.23	80.76	78.40	$84.10_{82.98-85.20}$	$\mathbf{92.97_{91.81-93.93}}$	$75.44_{71.66-78.34}$	92.49	75.67
Mean	88.46	78.61	75.48	$83.91_{82.39-85.58}$	$92.51_{91.75-93.27}$	$73.85_{72.10-75.61}$	90.11	77.71
KNN	86.58	75.87	72.80	$81.76_{80.20-83.28}$	$89.91_{89.14-90.67}$	$67.93_{66.19-69.67}$	88.32	75.20
MICE	85.75	75.37	72.06	$82.74_{81.24-84.30}$	$90.91_{90.26-91.57}$	$69.08_{68.10-70.06}$	86.83	78.64
Case 1_subset	Accuracy	F1	Precision	Recall_{CI}	AUC_{CI}	PRAUC_{CI}	Specificity	Sensitivity
Extra-value (Ours)	85.64	75.40	72.00	$83.10_{82.31-83.81}$	$\mathbf{91.00_{89.71-91.95}}$	$70.20_{67.96-71.27}$	86.55	79.63
Mean	85.49	75.06	71.80	$82.54_{80.97-84.06}$	$90.61_{89.99-91.32}$	$69.59_{68.44-70.65}$	86.55	78.52
KNN	83.51	72.90	69.83	$81.60_{80.23-82.87}$	$88.86_{88.07-89.66}$	$65.45_{64.22-66.71}$	84.20	78.99
MICE	83.22	72.71	69.67	$81.80_{80.34-83.12}$	$89.49_{88.70-90.29}$	$65.44_{63.76-67.07}$	83.73	79.86
Case 2	Accuracy	F1	Precision	Recall_{CI}	AUC_{CI}	PRAUC_{CI}	Specificity	Sensitivity
Extra-value (Ours)	86.80	73.54	72.00	$75.50_{74.44-76.41}$	$\mathbf{86.96_{86.43-88.00}}$	$59.30_{56.67-61.03}$	90.93	60.10
Mean	84.59	71.72	69.35	$76.01_{75.21-76.86}$	$86.31_{85.89-86.72}$	$56.93_{55.71-58.14}$	87.71	64.32
KNN	82.77	68.72	66.65	$72.79_{71.03-74.49}$	$82.59_{81.73-83.44}$	$47.16_{45.03-49.28}$	86.39	59.19
MICE	82.18	69.13	66.79	$74.70_{73.84-75.47}$	$84.02_{83.43-84.62}$	$50.59_{48.96-55.22}$	84.90	64.49
Case 2_subset	Accuracy	F1	Precision	Recall_{CI}	AUC_{CI}	PRAUC_{CI}	Specificity	Sensitivity
Extra-value (Ours)	81.25	68.60	66.26	$75.10_{74.27-75.99}$	$\mathbf{83.84_{82.42-84.89}}$	$51.60_{50.60-52.82}$	83.45	66.82
Mean	81.12	68.10	65.87	$74.19_{73.63-74.75}$	$83.27_{82.90-83.64}$	$49.60_{48.73-50.47}$	83.63	64.76
KNN	78.54	65.46	63.71	$72.33_{71.34-73.34}$	$80.84_{80.38-81.30}$	$43.87_{42.29-45.46}$	80.79	63.87
MICE	77.77	65.17	63.54	$72.86_{72.12-73.55}$	$81.49_{81.04-81.94}$	$45.51_{44.45-46.57}$	79.55	66.17

The Recall, AUC, and PRAUC are reported along with their corresponding CIs. The configuration achieving the highest AUC is highlighted in bold.

missingness pattern without imputing from other features. It allows the model to learn whether absence carries predictive value, crucial in general practice settings with irregular and heterogeneous test prescriptions.

Table 12. Comparison of Different Missingness Thresholds (Expressed in Percentage [%]): Experimental Results Obtained by the XGB Model in Case 1 and Case 2 Configurations

Case 1	Accuracy	F1	Precision	Recall_{CI}	AUC_{CI}	PRAUC_{CI}	Specificity	Sensitivity	# Predictors
< 90% (Ours)	90.23	80.76	78.40	$84.10_{82.98-85.20}$	$\mathbf{92.97_{91.81-93.93}}$	$75.44_{71.66-78.34}$	92.49	75.67	86
< 80%	89.83	80.28	77.68	$84.17_{82.46-85.77}$	$92.56_{91.88-93.22}$	$74.45_{73.20-75.70}$	91.89	76.30	63
< 70%	89.58	79.87	77.22	$83.84_{81.98-85.70}$	$92.46_{91.69-93.23}$	$74.87_{73.48-76.25}$	91.65	75.96	53
< 60%	88.35	78.55	75.37	$84.10_{82.47-85.74}$	$92.10_{91.42-92.78}$	$73.63_{72.32-74.94}$	89.88	78.28	40
< 50%	87.42	77.37	74.07	$83.54_{81.96-85.12}$	$91.19_{90.18-92.20}$	$71.17_{69.65-72.69}$	88.80	78.34	33
Case 2	Accuracy	F1	Precision	Recall_{CI}	AUC_{CI}	PRAUC_{CI}	Specificity	Sensitivity	# Predictors
< 90% (Ours)	86.80	73.54	72.00	$75.50_{74.44-76.41}$	$\mathbf{86.96_{86.43-88.00}}$	$59.30_{56.67-61.03}$	90.93	60.10	90
< 80%	86.38	72.89	71.34	$75.39_{74.37-76.41}$	$86.69_{86.17-87.21}$	$59.15_{57.37-60.93}$	90.52	59.45	66
< 70%	86.62	73.24	71.72	$75.24_{73.86-76.62}$	$86.81_{86.06-87.56}$	$58.10_{56.83-59.37}$	90.73	59.80	59
< 60%	85.23	71.92	69.86	$75.27_{73.96-76.58}$	$85.96_{85.23-86.69}$	$56.67_{55.38-57.96}$	88.86	61.54	43
< 50%	82.76	69.50	67.19	$74.59_{73.40-75.78}$	$84.05_{83.10-85.00}$	$51.83_{50.39-53.27}$	85.75	63.22	34

The total number of selected predictors (# Predictors) is reported, but the variation Δ of each predictor is omitted from the counting. The Recall, AUC, and PRAUC are reported along with their corresponding CIs. The configuration achieving the highest AUC is highlighted in bold.

Table 13. SMOTE and SMOTE-NC Comparison Experimental Results in Case 1 and Case 2 Configurations

XGB model	Accuracy	F1	Precision	Recall_{CI}	AUC_{CI}	PRAUC_{CI}	Specificity	Sensitivity
Case 1 (SMOTE) (Ours)	90.23	80.76	78.40	$84.10_{82.98-85.20}$	$\mathbf{92.97_{91.81-93.93}}$	$75.44_{71.66-78.34}$	92.49	75.67
Case 1 (SMOTE-NC)	88.78	79.14	75.99	$84.44_{83.07-85.74}$	$92.23_{91.67-92.79}$	$71.98_{70.72-73.24}$	90.95	78.52
Case 2 (SMOTE) (Ours)	86.80	73.54	72.00	$75.50_{74.44-76.41}$	$\mathbf{86.96_{86.43-88.00}}$	$59.30_{56.67-61.03}$	90.93	60.10
Case 2 (SMOTE-NC)	85.80	72.66	70.68	$75.66_{75.13-76.22}$	$86.38_{85.88-86.89}$	$57.50_{56.27-58.73}$	89.48	61.84

The Recall, AUC, and PRAUC are reported along with their corresponding CIs. The configuration achieving the highest AUC is highlighted in bold.

The comparative results with different missingness thresholds, ranging from 90% to 50%, for both Case 1 and Case 2 configurations, are summarized in Table 12. Overall, the best results in both configurations are achieved from our operative choice by selecting a missingness threshold of 90%. The results decrease smoothly by gradually tightening the missingness threshold until 50%, but always maintaining a reasonable comparative performance. The consistent pattern observed across Case 1 and Case 2 demonstrates the robustness and stability of the XGB model in handling high to moderate levels of missingness.

4.5.2 Comparison of Oversampling Techniques. Standard SMOTE is compared against SMOTE-NC across Case 1 and Case 2 configurations. SMOTE and SMOTE-NC yield comparable performance (see Table 13), indicating that the potential distortions introduced by standard SMOTE after extra-value imputation have only a minimal effect and can be reasonably neglected in this context. These comparable results can also be explained by the fact that, in our dataset, nearly all predictors contain at least one missing value, even after conservative filtering. As a consequence, the effective feature space available to SMOTE-NC becomes minimal and potentially uninformative, leading to unreliable neighbor selection and degraded sample quality. Figure 6 shows the consistency and the coherence of the XGB's built-in importance mechanism to select the most important features by adopting SMOTE and SMOTE-NC oversampling strategies for Case 1 and Case 2 configurations. The analysis highlights that the most discriminative features emerge at the top across both oversampling strategies and configurations, confirming the robustness and effectiveness of SMOTE.

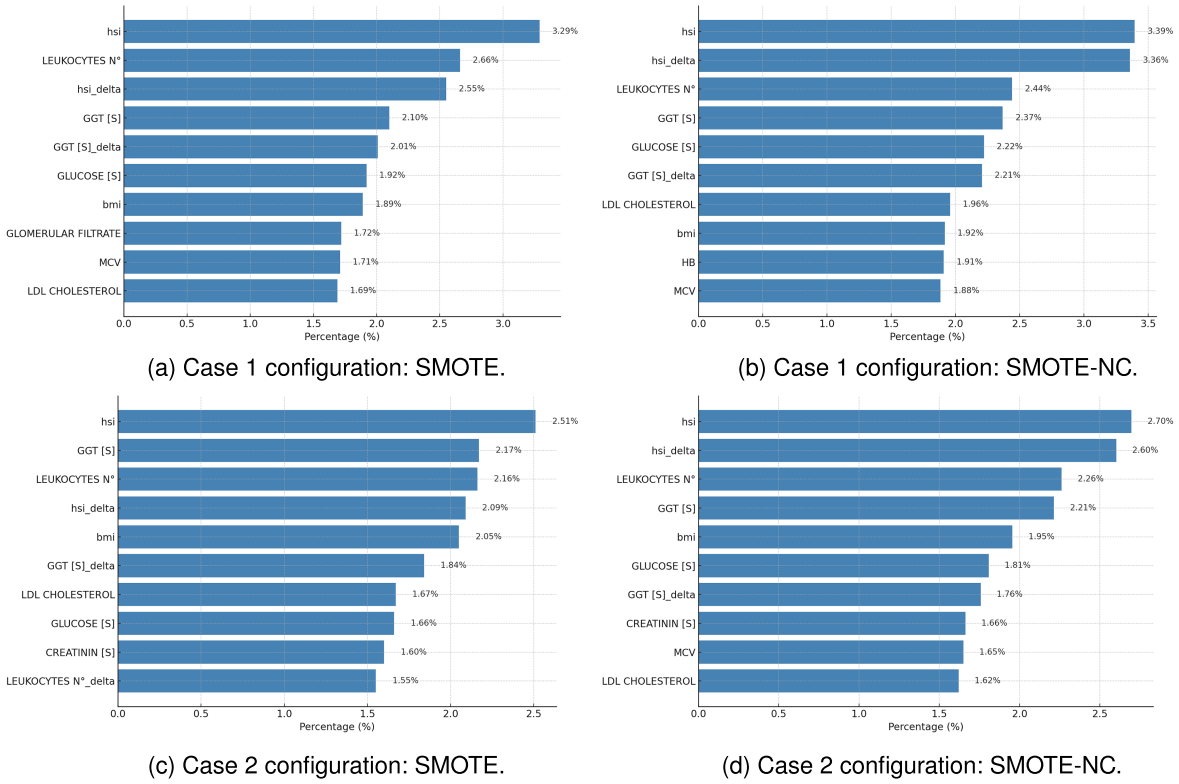


Fig. 6. Top-10 feature importance using SMOTE and SMOTE-NC for both Case 1 and Case 2 configurations.

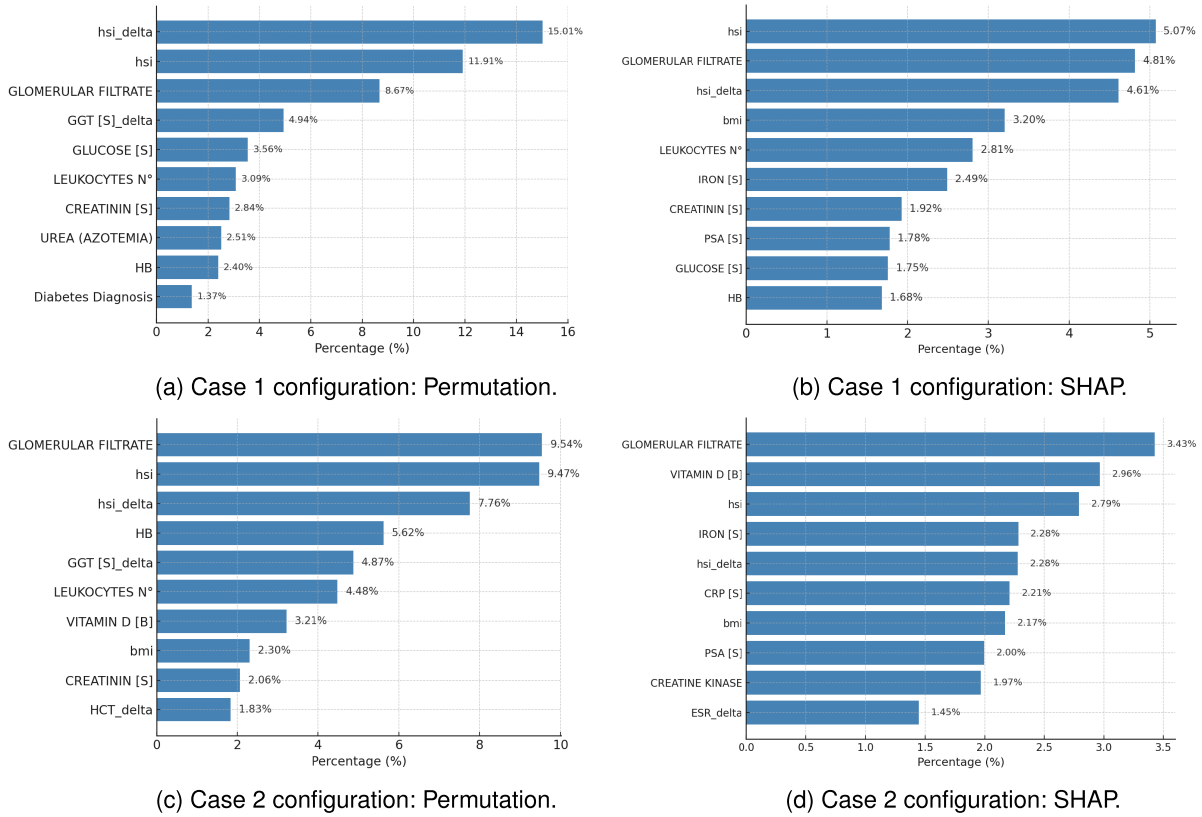
4.5.3 Comparison of Feature Importance Criteria. To better contextualize the feature importance results shown in Figure 4, which rely on XGB's built-in frequency-based importance, we also computed model-agnostic importance scores using two complementary strategies: permutation importance and SHAP. These complementary analyses (see Figure 7) support and reinforce the findings computed using XGB's built-in mechanism based on the frequency criterion. The most influential predictors identified by XGB (e.g., HSI, glomerular filtrate, GGT-S, glucose, leukocytes) for Case 1 and Case 2 configurations consistently appear among the top-ranked features in at least one of the model-agnostic strategies—permutation importance or SHAP analysis. This agreement across interpretability methods enhances confidence in the clinical relevance of the selected predictors and demonstrates the robustness of our model in terms of transparency and interpretability.

5 CDSS: A Use Case

This section presents a practical application of the CDSS ready to be embedded within EHRs for assessing fibrosis risk in general practice, as illustrated in Figure 1.

The CDSS is structured to serve both newly registered and long-term patients within the EHR system. It becomes operational once a patient's HSI exceeds the threshold of 36, alerting the GP to potential health risks.

The CDSS continuously updates the values of the predictors and their variations i.e., Δ values) throughout the patient's clinical history. To ensure consistency, data entered into the system is automatically converted to a standardized unit of measurement and normalization techniques are integrated to manage the varying scales of different laboratory exams. This standardization is crucial for maintaining the integrity of longitudinal data analysis, particularly when comparing current results with previous entries in a patient's clinical history.



(a) Case 1 configuration: Permutation.

(b) Case 1 configuration: SHAP.

(c) Case 2 configuration: Permutation.

(d) Case 2 configuration: SHAP.

Fig. 7. Top-10 feature importance through permutation importance and SHAP analysis for both Case 1 and Case 2 configurations.

The operational mechanism of the CDSS involves two different but complementary processes that follow in a cascade.

— *Broadband screening*: Initially, a broad screening, i.e., Case 2 (see Figure 3(b)), using a TWOI delimited by all possible FIB-4 values, classifies the patient condition as either "control" or "fibrosis," accompanied by a predictive probability.
— *Narrowband confirmation*: Subsequently, a more focused screening, i.e., Case 1 (see Figure 3(a)), excluding FIB-4 values in the range $1.30 < FIB - 4 < 2.67$, seeks to confirm or contrast the previous prediction of the fibrosis-risk profile.

If both stages classify a patient as "control," the CDSS reassures the GP with a reliable confidence level of a low fibrosis risk in the NAFLD patient. Conversely, if either stage classifies a risk of fibrosis, the CDSS indicates a moderate to high risk for the NAFLD patient. The physician then formulates a treatment plan based on (i) the probability of the prediction outputs, (ii) input lab test completeness, and (iii) the patient's overall clinical condition.

The deployment of the CDSS is planned in several phases to ensure robust integration within GPs' EHR systems. Initially, the CDSS will undergo rigorous pilot testing in several GPs' offices to address any operational issues and refine the predictive algorithms based on real-world data feedback. Then, the CDSS will be integrated into the EHR systems of selected general practices, with continuous monitoring and support to ensure seamless operation.

After a period of assessment and adjustment, the CDSS will be fully deployed across additional GP practices, with ongoing updates and improvements based on user feedback and evolving clinical needs.

6 Discussion

This study focuses on developing and researching predictive algorithms to assess the risk of hepatic fibrosis, leveraging ML to analyze clinical data derived from EHR routine lab tests. Unlike many existing EHR systems that rely on hospital data, the proposed approach sources data directly from GPs. The data includes non-hospitalized patients, allowing a more complete understanding of subjects' health over time. The data collection comprises longitudinal observations, enabling an analysis of the progression of individuals' health conditions. This longitudinal perspective is essential for understanding the risk of hepatic fibrosis, evaluating treatment efficacy, and determining the impact of lifestyle changes. However, these data also encapsulate the complexities inherent in GP scenarios, where sparse observations and irregular exams characterize data over time. This distinctive data-sourcing strategy not only facilitates a holistic view of chronic diseases and lifestyle factors but also underscores the necessity for the specialized ML approach developed in this study to navigate the challenges of such a complex data environment.

This study proposes an ML core integrated into a CDSS to help clinicians identify patients at risk of hepatic fibrosis and provide timely personalized care interventions. The developed CDSS aligns with EU and global strategic policy objectives related to AI in healthcare [27]. It contributes to developing trustworthy AI systems, empowers clinicians in their healthcare decision-making, and supports data for disease prevention and personalized care—objectives that align with the European Commission's White Paper on Artificial Intelligence.[2] The overall approach strongly emphasizes data privacy and governance, adhering to the General Data Protection Regulation (GDPR)[3] and ensuring secure and ethical handling of patient data. Moreover, the proposed CDSS ensures transparency and interpretability by exploiting the peculiarity of ensemble-based white-box models (i.e., XGB), which provide a feature importance that is model-intrinsic and can support clinical decision-making.

The developed CDSS conducts clinical validation in real-world scenarios involving individuals and their caregivers. This real-world validation approach ensures that the developed CDSS effectively assesses the risk of hepatic fibrosis in NAFLD patients.

6.1 ML Impact

As reported in Tables 5 and 6, the experimental results showcase the potential of ML in predicting hepatic fibrosis risk, with all explored models achieving acceptable AUC values ($AUC > 60\%$). Notably, the XGB model effectively uses GP-sourced data in making predictions: it achieved the highest performance metrics, recording an AUC of 92.97% for Case 1 and 86.96% for Case 2.

GP data, characterized by its sparsity and irregular timing of observations, poses significant challenges for traditional data analysis techniques. The preprocessing stage, including patients' and predictors' selection criteria, lays the foundation for designing a suitable ML model that is specifically tailored for dealing with heterogeneous and sparse tabular data. The XGB model efficiently utilizes the extracted features, compensating for gaps in the data and reducing the noise from irregular observations. The employed dataset includes a broad spectrum of patients who disclose NAFLD, reflecting the diverse conditions managed in primary care settings. The XGB model's strength lies in its ability to discern complex patterns within this diversity, facilitating reliable risk predictions for hepatic fibrosis across varied patient populations. The ensemble nature of the XGB model, which integrates insights from multiple DTs, enhances the model's generalizability and robustness against overfitting. This peculiarity ensures that the model's predictions remain reliable despite the variability in individual patient data.

[2] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.
[3] https://eur-lex.europa.eu/EN/legal-content/summary/general-data-protection-regulation-gdpr.html.

The XGB model performance remains consistently robust, even when utilizing a reduced set of predictors chosen based on the XGB model's feature importance from Case 1 and Case 1 configurations. As demonstrated in both Case 1_subset and Case 2_subset configurations, the AUC remains stable, achieving 91.00% and 83.84%, respectively, compared to when using all predictors (as shown in Table 7). Moreover, the Recall remains consistently high across all configurations, achieving 84.10% for Case 1 and 83.10% for Case 1_subset and 75.50% for Case 2 and 75.10% for Case 2_subset. These stable Recall rates underscore the model's ability to deliver reliable predictions, demonstrating its robust performance regardless of the predictor set size.

The external validation experimental results over the FIMMG dataset highlight important insights regarding the generalizability of the proposed trained XGB model. These findings confirm the robustness and reliability of the trained XGB model under the narrowband Case 1 configuration, suggesting that a well-defined and clinically constrained case definition, combined with an optimized predictor subset, can guarantee model generalization across different and unseen patient populations. On the other hand, the less stringent criteria used in broadband Case 2 configuration may have introduced greater heterogeneity within classes, negatively affecting the model's ability to generalize over fibrosis condition. Overall, these findings may highlight the critical importance of carefully selecting both the experimental configurations and an optimal subset of predictors when designing clinically deployable predictive models within a CDSS. Such operative choices can significantly influence the generalizability and scalability of the resulting models, ultimately impacting their effectiveness in real-world clinical settings.

A critical requirement for CDSSs is interpretability, mainly when dealing with complex clinical tasks. XGB provides intrinsic global interpretability by exposing feature importance through split frequencies and gain-based metrics within its constituent DTs [23]. In our work, we explicitly focused on the model-intrinsic interpretability. The model's structure allows for the extraction and examination of the feature-importance scores, which indicate how significantly each variable (e.g., lab test) influences fibrosis-risk prediction. This approach offers transparent and clinically meaningful insights into which variables most influence the model's predictions, without relying on *post hoc*, model-agnostic tools such as SHAP or LIME. By doing so, we aim to preserve interpretability directly aligned with the model's internal logic, making it more accessible and trustworthy for clinical decision-making. Such transparency is invaluable in clinical settings, as it aids GPs and other healthcare providers to understand the rationale behind predictions and guide informed treatment decisions, follow-up protocols, and lifestyle recommendations. This level of insight also fosters trust among clinicians and patients in the system's recommendations [37].

6.2 Clinical Impact

The proposed CDSS offers a multifaceted and proactive approach to managing NAFLD in primary care. Patients with this condition and other chronic diseases require continuous monitoring of disease progression [2]. Thus, the role of GPs is fundamental in controlling health conditions and referring to specialized care when needed. Therefore, the proposed CDSS provides automatic and continual support for GPs to promptly identify patients seeking medical attention and enable timely implementation of target interventions, which can reverse disease progression before it advances further [33].

While the FIB-4 index remains a valuable, low-cost tool for initial liver fibrosis screening based on AST and ALT levels, its diagnostic capacity is inherently limited to a small subset of available biomarkers and predefined thresholds. In contrast, our proposed ML-based CDSS is designed to go beyond FIB-4 by leveraging the predictive power of a broader set of routinely collected lab tests and clinical features, thus enabling the discovery of additional biomarkers and hidden multivariate patterns that are not directly apparent to clinicians. Our approach supports the clinical intuition that fibrosis risk arises from the simultaneous effect of multiple interdependent factors—such as metabolic, inflammatory, and hematological signals—whose interactions are often too complex to be captured through traditional scoring systems or clinician reasoning alone. The developed CDSS framework is therefore

essential in uncovering these subtle, non-linear relationships, offering a more comprehensive and individualized assessment of fibrosis risk.

By employing two screening scenarios in a cascade fashion, the CDSS aims at aiding GPs also in cases in which the risk of fibrosis is more ambiguous: Case 1 configuration, excluding borderline risk i.e., when $1.30 < FIB-4 < 2.67$), provides support for evident fibrosis assessment, which possibly requires an urgent treatment; at the same time, Case 2 configuration, encompassing a boarder range of FIB-4 values, ensures that no potential risk is overlooked, thus making it particularly valuable for patient with a short clinical history.

A distinct feature of this CDSS is the importance given to the patient's clinical history and how predictors change over time. Incorporating a 15-year clinical history window allows for detecting subtle critical changes in a patient's condition that shorter time frames might miss and for the comprehensive analysis of long-term disease development. In addition, the system emphasizes the importance of temporal variations (Δ) in predictive factors. Capturing the dynamic changes introduces a significant novelty, considering that such temporal variations are often underutilized in clinical practice. Moreover, this temporal sensitivity is crucial, as the top-10 features importance ranking for both screening scenarios (see Figure 4) includes several Δ changes, underscoring their value in distinguishing a fibrosis-risk profile. Figures 4 and 5 show that the most impactful predictors are consistently identified, underscoring not only the model's effectiveness and robustness but also its clinical reliability. This consistency is crucial in clinical settings, particularly in screening phases in which high Sensitivity is essential for the early detection of fibrosis cases.

Regarding specificity, the model consistently achieves high performances across all configurations (see Table 7): 92.49% and 86.55% for Case 1 and Case 1_subset, respectively; and 90.93% and 83.45% for Case 2 and Case 2_subset, respectively. These performances indicate a model's ability to correctly identify non-fibrotic cases, which is paramount in preventing false positives, especially in broad screening applications. The sensitivity of the model also demonstrates its adeptness at correctly identifying fibrotic cases, which has improved notably when relying on a subset of the most impactful predictors. As reported in Table 7, sensitivity increases from 75.67% in Case 1 to 79.63% in Case 1_subset, and from 60.10% in Case 2 to 66.82% in Case 2_subset. This enhancement suggests that even with fewer predictors, the model can achieve comparable, if not superior, performance in accurately identifying fibrosis risk, underscoring its critical capability for effective clinical screening and proactive healthcare interventions.

Overall, the clinical significance of this CDSS is rooted in its ability to provide an accurate, nuanced and comprehensive evaluation of fibrosis risk, thereby facilitating preventative measures. By leveraging the combination of Case 1 and Case 2 screenings, alongside the innovative use of Δ variations in predictive factors over a broad clinical history and a reduced optimal number of predictors, the proposed system has the potential to empower GPs in supporting their timely and preventive actions aimed at reversing patient conditions.

6.3 Socio-Economic Impact

The proposed CDSS holds the promise of significant socio-economic benefits. The early detection and preventive care facilitated by the CDSS could notably reduce the healthcare costs associated with hepatic fibrosis, translating into savings for healthcare systems and patients [10]. Moreover, the requirement for fewer diagnostic tests to achieve effective and reliable screening represents a substantial benefit for patients, clinicians, and diagnostic centers. This reduction in necessary tests contributes to cost savings and decreases the time required for medical processing, thereby enhancing the efficiency of healthcare delivery. By focusing on patient-centric care, the CDSS empowers patients to manage their condition actively, likely leading to improved health outcomes and reduced healthcare burdens [21]. This reduction will translate into care and hospitalization services in terms of reduced beds occupied by patients with avoidable critical conditions, improved preventive actions that could avoid complications and hospitalizations, better allocation of health and care personnel, and better functioning of territorial praesidium structures of care [25]. Given the prevalence of NAFLD globally, the widespread implementation of such a CDSS

could have a profound public health impact, contributing to healthier populations and reducing the prevalence of hepatic fibrosis [25]. This impact, in turn, enhances societal well-being by improving healthcare access, quality, and efficiency, fostering well-being and quality of life. Additionally, the CDSS drives digital health innovation, encouraging the development of new technologies and solutions that can be integrated into EHRs, further revolutionizing healthcare delivery and patient outcomes. The effectiveness of the proposed CDSS can incentivize healthcare professionals, potentially leading to the adoption of more ML-driven, incentive-based healthcare systems.

6.4 Ethical Aspects

The proposed framework is also compliant with the ethics guidelines of the European Commission (Human Agency and Oversight [12]) and is currently designed for screening purposes. In line with current best practices for clinical prediction model development and validation, the study follows the TRIPOD-AI guidelines [11], ensuring transparency, reproducibility, and ethical rigor in the reporting of ML-based CDSS. The completed TRIPOD-AI checklist can be found in Supplementary Material, with section references indicating where each reporting item is addressed in the article.

The key steps for patients' selection (see Figure 2) demonstrate robust generalization performance across diverse patient demographics, ensuring fair and equitable predictive accuracy for both male and female patients. This fairness is essential for maintaining trust and efficacy in clinical practice, ensuring the system provides reliable patient risk assessments. Moreover, we leverage the white-box nature of XGB. This model excels at pattern discrimination and offers valuable characterizations of these patterns. The CDSS allows clinicians to understand the underlying factors influencing each prediction by extracting and examining feature-importance scores. This level of transparency is crucial for clinical acceptance, as it helps healthcare providers make informed decisions based on transparent clinical insights into the decision-making process. The extract results in an optimal subset of predictors (see Section 3.5) confirmed a saving of the CDSS's cost-effectiveness without compromising diagnostic accuracy. This approach reduces the burden on clinicians to annotate all examinations, simplifying the diagnostic process and making it more accessible and practical for widespread clinical use.

6.5 Limitations and Future Work

A first limitation of this study concerns the assumption of MCAR adopted to handle missing data. Potential deviations from this assumption could influence the distribution of certain laboratory variables with missing values and, consequently, the stability of their imputations. However, it is worth noting that the primary objective of this study was not to introduce new imputation methodologies but to develop and validate an ML-based CDSS for monitoring fibrosis risk in NAFLD patients. Nonetheless, the imputation strategies adopted in this work (extra-value, mean, KNN, and MICE) are widely used in the literature even in contexts where the MCAR assumption is not strictly verified, and the consistent performance observed across these methods suggests that the overall predictive accuracy of the proposed models remains robust to this limitation.

A further limitation of the current study is the lack of temporal stratification in predicting the onset of hepatic fibrosis, whether in the short-term or long-term, as it has been explored in some previous studies for other diseases (e.g., [7]). Understanding the timing of fibrosis onset is crucial for tailoring interventions and managing patient care more effectively. Future work will refine the XGB model to distinguish between short-term and long-term risks of progression to hepatic fibrosis in patients diagnosed with NAFLD. This approach may also encapsulate the trajectory of laboratory markers and clinical indices over time, enhancing the prediction of fibrosis onset with temporal precision. An additional interesting future direction involves finding an optimal balance between the minimum number of required lab tests, lab test types, and an acceptable level of missing data, which could significantly streamline the diagnostic process. This approach aims to reduce the burden on healthcare systems and make the prediction model more accessible and practical for widespread clinical applications. Exploring

these aspects could mitigate the current system's limitations and increase its effectiveness and applicability in predicting hepatic fibrosis risk.

7 Conclusions

The study introduces a CDSS built on an XGB model that leverages GP-sourced data for predicting hepatic fibrosis risk in NAFLD patients. The system demonstrates notable predictive performance and high transparency by incorporating extensive EHR data, clinically meaningful preprocessing, and employing advanced ML techniques. The present findings highlight the substantial potential of the proposed approach in enhancing early detection and personalized management of hepatic fibrosis.

Privacy Statement. The Ethical Committees of the University approved the experimental study and its guidelines as a clinical non-interventional (observational) study. EHR data are anonymous and their use, detention and conservation are regulated by an agreement between the authors and data owners. The entire process is under the EU GDPR regulation.

Acknowledgments

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] Leon A. Adams. 2011. Biomarkers of liver fibrosis. *Journal of Gastroenterology and Hepatology* 26, 5 (2011), 802–809.

[2] Gopi Battineni, Getu Gamo Sagaro, Nalini Chinatalapudi, and Francesco Amenta. 2020. Applications of machine learning predictive models in the chronic disease diagnosis. *Journal of Personalized Medicine* 10, 2 (2020), 21.

[3] Brett K. Beaulieu-Jones, Daniel R. Lavage, John W. Snyder, Jason H. Moore, Sarah A. Pendergrass, and Christopher R. Bauer. 2018. Characterizing and managing missing structured data in electronic health records: Data analysis. *JMIR Medical Informatics* 6, 1 (Feb. 2018), e11.

[4] Mark Benedict and Xuchen Zhang. 2017. Non-alcoholic fatty liver disease: An expanded review. *World Journal of Hepatology* 9, 16 (2017), 715–732.

[5] Michele Bernardini, Anastasiia Doinychko, Luca Romeo, Emanuele Frontoni, and Massih-Reza Amini. 2023. A novel missing data imputation approach based on clinical conditional generative adversarial networks applied to EHR datasets. *Computers in Biology and Medicine* 163 (2023), 107188.

[6] Michele Bernardini, Luca Romeo, Emanuele Frontoni, and Massih-Reza Amini. 2021. A semi-supervised multi-task learning approach for predicting short-term kidney disease evolution. *IEEE Journal of Biomedical and Health Informatics* 25, 10 (2021), 3983–3994.

[7] Michele Bernardini, Luca Romeo, Adriano Mancini, and Emanuele Frontoni. 2021. A clinical decision support system to stratify the temporal risk of diabetic retinopathy. *IEEE Access* 9 (2021), 151864–151872.

[8] Annalisa Berzigotti, Emmanouil Tsochatzis, Jerome Boursier, Laurent Castera, Nora Cazzagon, Mireen Friedrich-Rust, Salvatore Petta, Maja Thiele, and European Association for the Study of the Liver. 2021. EASL clinical practice guidelines on non-invasive tests for evaluation of liver disease severity and prognosis—2021 update. *Journal of Hepatology* 75, 3 (2021), 659–689.

[9] Victoria Blanes-Vidal, Katrine P. Lindvig, Maja Thiele, Esmaeil S. Nadimi, and Aleksander Krag. 2022. Artificial intelligence outperforms standard blood-based scores in identifying liver fibrosis patients in primary care. *Scientific Reports* 12, 1 (2022), 2914.

[10] Insook Cho and David W. Bates. 2018. Behavioral economics interventions in clinical decision support systems. *Yearbook of Medical Informatics* 27, 01 (2018), 114–121.

[11] Gary S. Collins, Karel G. M. Moons, Paula Dhiman, Richard D. Riley, Andrew L. Beam, Ben Van Calster, Marzyeh Ghassemi, Xiaoxuan Liu, Johannes B. Reitsma, Maarten Van Smeden, et al. 2024. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *British Medical Journal* 385 (2024), q902.

[12] EU. 2019. Ethics Guidelines for Trustworthy AI. Retrieved March 31, 2025 from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[13] Gabriele Forlani, Carlo Giorda, Roberta Manti, Natalia Mazzella, Salvatore De Cosmo, Maria Chiara Rossi, Antonio Nicolucci, Paolo Di Bartolo, Antonio Ceriello, and Pietro Guida. 2016. The burden of NAFLD and its characteristics in a nationwide population with type 2 diabetes. *Journal of Diabetes Research* 2016, 1 (2016), 2931985.

[14] Sina Ghandian, Rahul Thapa, Anurag Garikipati, Gina Barnes, Abigail Green-Saxena, Jacob Calvert, Qingqing Mao, and Ritankar Das. 2022. Machine learning to predict progression of non-alcoholic fatty liver to non-alcoholic steatohepatitis or fibrosis. *JGH Open* 6, 3 (2022), 196–204.

[15] Carlo Giorda, Gabriele Forlani, Roberta Manti, Natalia Mazzella, Salvatore De Cosmo, Maria Chiara Rossi, Antonio Nicolucci, Giuseppina Russo, Paolo Di Bartolo, Antonio Ceriello, et al. 2017. Occurrence over time and regression of nonalcoholic fatty liver disease in type 2 diabetes. *Diabetes/Metabolism Research and Reviews* 33, 4 (2017), e2878.

[16] Carlo Bruno Giorda, Gabriele Forlani, Roberta Manti, Arianna Mazzotti, Salvatore De Cosmo, Maria Chiara Rossi, Antonio Nicolucci, Paolo Di Bartolo, Antonio Ceriello, Pietro Guida, et al. 2018. Trend over time in hepatic fibrosis score in a cohort of type 2 diabetes patients. *Diabetes Research and Clinical Practice* 135 (2018), 65–72.

[17] Sebastien Haneuse, David Arterburn, and Michael J. Daniels. 2021. Assessing missing data assumptions in EHR-based studies: A complex and underappreciated task. *JAMA Network Open* 4, 2, (2021), e210184–e210184.

[18] Md Mohaimenul Islam, Chieh-Chen Wu, Tahmina Nasrin Poly, Hsuan-Chia Yang, and Yu-Chuan Jack Li. 2018. Applications of machine learning in fatty live disease prediction. In *Building Continents of Knowledge in Oceans of Data: The Future of co-Created eHealth*. IOS Press, 166–170.

[19] Rustam N. Karanjia, Mary M. E. Crossey, I. Jane Cox, Haddy K. S. Fye, Ramou Njie, Robert D. Goldin, and Simon D. Taylor-Robinson. 2016. Hepatic steatosis and fibrosis: Non-invasive assessment. *World Journal of Gastroenterology* 22, 45 (2016), 9880–9897.

[20] Georgios Katsimpras, Fotis Aisopos, Peter Garrard, Maria-Esther Vidal, and Georgios Paliouras. 2022. Improving early prognosis of dementia using machine learning methods. *ACM Transactions on Computing for Healthcare* 3, 3 (2022), 1–16.

[21] Clemens Scott Kruse and Nolan Ehrbar. 2020. Effects of computerized decision support systems on practitioner performance and patient outcomes: Systematic review. *JMIR Medical Informatics* 8, 8 (2020), e17283.

[22] Roderick J. A. Little and Donald B. Rubin. 2019. *Statistical Analysis with Missing Data*. John Wiley & Sons.

[23] Octavio Loyola-Gonzalez. 2019. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access* 7 (2019), 154096–154113.

[24] Ramasamy Malarvizhi and Antony Selvadoss Thanamani. 2012. K-nearest neighbor in missing data imputation. *International Journal of Engineering Research and Development* 5, 1 (2012), 5–7.

[25] Raghad Muhiyaddin, Alaa A. Abd-Alrazaq, Mowafa Househ, Tanvir Alam, and Zubair Shah. 2020. The impact of clinical decision support systems (CDSS) on physicians: A scoping review. *Studies in Health Technology and Informatics* 272 (2020), 470–473.

[26] Antonio Nicolucci, Luca Romeo, Michele Bernardini, Marco Vespasiani, Maria Chiara Rossi, Massimiliano Petrelli, Antonio Ceriello, Paolo Di Bartolo, Emanuele Frontoni, and Giacomo Vespasiani. 2022. Prediction of complications of type 2 diabetes: A machine learning approach. *Diabetes Research and Clinical Practice* 190 (2022), 110013.

[27] Nikos Th. Nikolinakos. 2023. *EU Policy and Legal Framework for Artificial Intelligence, Robotics and Related Technologies-The AI Act*. Springer.

[28] Sonali Paul and Andrew M. Davis. 2018. Diagnosis and management of nonalcoholic fatty liver disease. *JAMA* 320, 23 (2018), 2474–2475.

[29] Sajida Perveen, Muhammad Shahbaz, Karim Keshavjee, and Aziz Guergachi. 2018. A systematic machine learning based approach for the diagnosis of non-alcoholic fatty liver disease risk and progression. *Scientific Reports* 8, 1 (2018), 2112.

[30] Mary E. Rinella and Arun J. Sanyal. 2015. Genetics, diagnostics and therapeutic advances in NAFLD. *Nature Reviews: Gastroenterology & Hepatology* 12, 2 (2015), 65–66.

[31] Arun J. Sanyal. 2019. Past, present and future perspectives in nonalcoholic fatty liver disease. *Nature Reviews: Gastroenterology & Hepatology* 16, 6 (2019), 377–386.

[32] Yuqi Si, Jingcheng Du, Zhao Li, Xiaoqian Jiang, Timothy Miller, Fei Wang, W. Jim Zheng, and Kirk Roberts. 2021. Deep representation learning of patient data from electronic health records (EHR): A systematic review. *Journal of Biomedical Informatics* 115 (2021), 103671.

[33] Mengxi Sun and Tatiana Kisseleva. 2015. Reversibility of liver fibrosis. *Clinics and Research in Hepatology and Gastroenterology* 39 (2015), S60–S63.

[34] Jiaqi Wang, Junyu Luo, Muchao Ye, Xiaochen Wang, Yuan Zhong, Aofei Chang, Guanjie Huang, Ziyi Yin, Cao Xiao, Jimeng Sun, et al. 2024. Recent advances in predictive modeling with electronic health records. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI '24)*, 8272–8280.

[35] Yonghui Wu, Xi Yang, Heather L. Morris, Matthew J. Gurka, Elizabeth A. Shenkman, Kenneth Cusi, Fernando Bril, and William T. Donahoo. 2022. Noninvasive diagnosis of nonalcoholic steatohepatitis and advanced liver fibrosis using machine learning methods: Comparative study with existing quantitative risk scores. *JMIR Medical Informatics* 10, 6 (2022), e36997.

[36] Jiabao Xu, Xuefeng Xi, Jie Chen, Victor S. Sheng, Jieming Ma, and Zhiming Cui. 2022. A survey of deep learning for electronic health records. *Applied Sciences* 12, 22 (2022), 11709.

[37] Qian Xu, Wenzhao Xie, Bolin Liao, Chao Hu, Lu Qin, Zhengzijin Yang, Huan Xiong, Yi Lyu, Yue Zhou, and Aijing Luo. 2023. Interpretability of clinical decision support systems based on artificial intelligence from technological and medical perspective: A systematic review. *Journal of Healthcare Engineering* 2023, 1 (2023), 9919269.

[38] Nobutake Yamamichi, Takeshi Shimamoto, Kazuya Okushin, Takako Nishikawa, Hirotaka Matsuzaki, Seiichi Yakabi, Mami Takahashi, Ryoichi Wada, Kazuhiko Koike, and Mitsuhiro Fujishiro. 2022. Fibrosis-4 index efficiently predicts chronic hepatitis and liver cirrhosis development based on a large-scale data of general population in Japan. *Scientific Reports* 12, 1 (2022), 20357.

[39] Siyue Yang, Paul Varghese, Ellen Stephenson, Karen Tu, and Jessica Gronsbell. 2023. Machine learning approaches for electronic health records phenotyping: A methodical review. *Journal of the American Medical Informatics Association* 30, 2 (2023), 367–381.

[40] Terry Cheuk-Fung Yip, Andy J. Ma, Vincent Wai-Sun Wong, Yee-Kit Tse, Hung L.-Y. Chan, Pong C. Yuen, and Grace Lai-Hung Wong. 2017. Laboratory parameter-based machine learning model for excluding non-alcoholic fatty liver disease (NAFLD) in the general population. *Alimentary Pharmacology & Therapeutics* 46, 4 (2017), 447–456.

[41] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. 2018. GAIN: Missing data imputation using generative adversarial nets. In *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80, PMLR, 5689–5698.

Received 21 June 2024; revised 19 October 2025; accepted 25 December 2025