Computer Methods and Programs in Biomedicine Update 7 (2025) 100176

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine Update

journal homepage:
www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine-update

MORIX: Machine learning-aided framework for lethality detection and MORtality inference with eXplainable artificial intelligence in MAFLD subjects

Domenico Lofù [a,*], Paolo Sorino [a,*], Tommaso Colafiglio [b], Caterina Bonfiglio [c], Rossella Donghia [c], Gianluigi Giannelli [d], Angela Lombardi [a], Tommaso Di Noia [a], Eugenio Di Sciascio [a], Fedelucio Narducci [a]

[a] Dept. of Electrical and Information Engineering (DEI), Politecnico di Bari, Bari, BA, 70125, Italy
[b] Dept. of Computer, Automatic and Management Engineering (DIAG), Sapienza Università di Roma, Roma, RM, 00185, Italy
[c] Unit of Data Science, National Institute of Gastroenterology "Saverio de Bellis", IRCCS Hospital, Castellana Grotte, BA, 70013, Italy
[d] Scientific Direction, National Institute of Gastroenterology "Saverio de Bellis", IRCCS Hospital, Castellana Grotte, BA, 70013, Italy

ARTICLE INFO

Keywords:
Machine learning techniques
MAFLD
Interpretability
Mortality
Epidemiology

ABSTRACT

Metabolic dysfunction-associated fatty liver disease (MAFLD) introduces new diagnostic criteria for fatty liver disease that are independent of alcohol consumption and viral hepatitis infection. Therefore, investigating how biochemical and anthropometric factors influence mortality in MAFLD subjects is of significant interest. In this work, we propose MORIX, an Artificial Intelligence-based framework capable of predicting fatal mortality outcomes in subjects with MAFLD. MORIX utilizes data from epidemiological datasets containing carefully selected anthropometric and biochemical information. This selection is achieved through Recursive Feature Elimination (RFE) using a Random Forest (RF) to train Machine Learning (ML) algorithms and provide a mortality risk (Yes/No) output. To provide physicians with a valuable tool, MORIX was trained and tested on a dataset of MAFLD subjects, comparing five different models: Random Forest (RF), eXtreme Gradient Boosting (XGB), Support Vector Machine (SVM), Multilayer Perceptron (MLP), and Light Gradient Boosting Model (LGBM) in a 5-fold cross-validation training strategy. Experimental results identified the RF as the best model, achieving a high accuracy for both mortality risks predicted. Additionally, an eXplainable Artificial Intelligence (XAI) analysis was conducted to clarify the diagnostic logic of the RF model and to assess the impact of each feature to the prediction. Moreover, a web application was developed to predict mortality risk and provide explanations of how the input features influenced the final prediction. In conclusion, the MORIX framework is easy to apply, and the required parameters are readily available in healthcare datasets, making it a practical tool for medical professionals.

1. Introduction

Non-alcoholic Fatty Liver Disease (NAFLD) is the predominant cause of chronic liver disease in Western countries, associated with heightened risks of cardiovascular diseases, type 2 diabetes mellitus, chronic renal disease, and increased mortality rates [1,2]. In severe cases, Non-alcoholic Fatty Liver Disease (NAFLD) may progress to fibrosis or hepatocarcinoma. The global prevalence of NAFLD is approximately 23.71% in Europe [3]. In particularly in our geographical area (Bari district of the Apulian Region in Italy) studies report a NAFLD prevalence of about 30%, predominantly among male and older individuals [4–7]. In 2020, a new term, Metabolic dysfunction-associated fatty liver disease (MAFLD), was proposed to replace NAFLD [8]. Differing from NAFLD, the MAFLD diagnosis does not necessitate the exclusion of other liver diseases, such as excessive alcohol consumption or viral hepatitis. MAFLD is diagnosed in patients with hepatic steatosis who also exhibit one of three metabolic conditions: overweight/obesity, metabolic dysfunction in lean subjects, or diabetes mellitus [9].

However, the utility of this new definition in improving the detection of critical clinical endpoints (*e.g.* mortality) remains debated.

* Corresponding authors.
E-mail addresses: domenico.lofu@poliba.it (D. Lofù), paolo.sorino@poliba.it (P. Sorino), tommaso.colafiglio@uniroma1.it (T. Colafiglio), catia.bonfiglio@irccsdebellis.it (C. Bonfiglio), rossella.donghia@irccsdebellis.it (R. Donghia), gianluigi.giannelli@irccsdebellis.it (G. Giannelli), angela.lombardi@poliba.it (A. Lombardi), tommaso.dinoia@poliba.it (T. Di Noia), eugenio.disciascio@poliba.it (E. Di Sciascio), fedelucio.narducci@poliba.it (F. Narducci).

https://doi.org/10.1016/j.cmpbup.2024.100176
Received 26 August 2024; Received in revised form 29 November 2024; Accepted 17 December 2024
Available online 6 January 2025

While substantial evidence connects MAFLD with cardiovascular diseases (CVD), malignancies, and liver-related outcomes, its impact on overall mortality remains a topic of debate [10]. In light of the increasing prevalence of NAFLD and the recent redefinition of MAFLD, there has been a surge in research focusing on cost-effective diagnostic techniques. Among these, Machine Learning (ML) has emerged as a particularly promising approach.

In clinical practice, numerous studies have explored how ML and e-health tools serve as viable alternatives to conventional diagnostic methods like Magnetic Resonance Imaging (MRI), Mild Cognitive Impairment (MCI), and ultrasounds [11–17]. ML techniques are already employed for NAFLD diagnosis [18,19]. Moreover, Saihood et al. [20] present a framework for diabetes diagnosis comprising three stages: data preprocessing, hyperparameter tuning, and ensemble methods. In particular, the preprocessing stage ensures data quality by addressing missing values, outliers, and feature selection. Subsequently, a grid search is applied to optimize algorithm hyperparameters. Finally, ensemble methods, including bagging, boosting, and stacking, are used to combine algorithms and enhance model robustness. Experimental results show that stacking with Random Forest and SVM achieved the highest accuracy of 97.50%, demonstrating that ensemble models can significantly outperform individual algorithms for reliable diagnosis.

Furthermore, from the perspective of developing models to predict heart attack, Al-Jamimi [21], presents a predictive model that combines eXtreme Gradient Boosting (XGBoost) with Bayesian optimization, demonstrating effectiveness in identifying risks for diseases such as heart attack, breast cancer, and diabetes. Specifically, the work adopt Support Vector Machine (SVM) using Recursive Feature Elimination (RFE) approach for feature selection, enabling the model to prioritize critical information and enhance prediction accuracy. The experimental results show that comparative analysis with established models (*e.g.* SVM, Random Forest (RF) and K-Nearest Neighbors (KNN)) underscores the model adaptability and potential for applications in precision healthcare.

According to Curci et al. [22], who demonstrated the benefits of physical activity and dietary interventions in NAFLD patients, this manuscript proposes a method to predict the NAFLD-patient mortality risk. This prediction allows physicians to recommend lifestyle modifications to patients with a high mortality risk.

As of 2023, the term Metabolic Dysfunction-Associated Steatotic Liver Disease (MASLD) was introduced [23], which differs from NAFLD because it refers to metabolic-associated liver disease with a focus potentially on steatohepatitis, emphasizing inflammation along with metabolic dysfunction. This paper, however, focuses on predicting mortality in NAFLD subjects as it is the accepted term in medical communities, reflecting a broader understanding of the metabolic factors involved in liver disease, and introducing a framework for such predictions.

Our experimental setup employs a population-based cohort with 15 years of follow-up data, comprising a dataset with 25 variables. We conduct a feature selection analysis to correlate features with the mortality outcome. The framework seamlessly integrates Machine Learning (ML) and eXplainable Artificial Intelligence (XAI), enabling both predictive accuracy and interpretability, thus making it a powerful tool for clinicians.

The integration of ML and XAI in our framework is a core of our work, providing clear, actionable insights alongside accurate predictions. To demonstrate the robustness of our approach, we compare 5 ML algorithms using common performance metrics like Accuracy, Precision, Recall, F1 score, Area Under the ROC Curve (AUC), and Confusion Matrix (CM) in order to identify the best one for the considered task.

The key contributions of this work are summarized as follows:

- **Integrated ML-XAI Framework:** The development of MORIX, an end-to-end framework that seamlessly integrates ML and XAI to predict mortality in MAFLD subjects. This integration provides both accurate predictions and clear explanations to aid clinical decision-making, making it a powerful tool for clinicians.
- **Feature Selection and Optimization:** Implementation of a robust and sound feature selection process using RFE with a RF classifier to identify and optimize the most significant features associated with mortality, ensuring relevance and reducing redundancy in the dataset.
- **Explainability as a Core Strength:** Detailed post-hoc explainability analysis using SHAP values to interpret the model's predictions, elucidating the contribution of each feature and enhancing transparency. This approach provides physicians with actionable insights, fostering trust and facilitating informed clinical decisions.
- **Practical Application and User Interface:** Development of a user-friendly web application that integrates the trained ML model and XAI analysis. This allows healthcare professionals to input new data and receive both mortality risk predictions and explanatory insights in real-time, thus bridging the gap between advanced AI techniques and practical clinical use.

The remainder of the paper is organized as follows: in Section 2, we overview the most relevant related work; Section 3 describes the proposed approach; Section 4 provides an overview of the proposed method, and Section 5 reports the metrics used to measure the performance of our model, the explanation obtained from the XAI analysis, and the web application developed. Section 6 details the developed web user interface. Finally, Section 7 discusses the results and tightens the conclusion and future research directions.

2. Related work

This section summarizes relevant works and delineates the significant distinctions between this study and previous efforts, thereby highlighting our novel contributions.

The distribution of MAFLD and NAFLD among individuals with Chronic Kidney Disease (CKD) was investigated by Dan-Qin Sun et al. [24]. Their study sought to compare the prevalence of CKD in individuals with MAFLD or NAFLD and to examine the relationship between the presence and severity of MAFLD and CKD, as well as abnormal albuminuria. They found that renal dysfunction severity and CKD stage 1 prevalence escalated with increasing MAFLD severity, suggesting that MAFLD better identifies subjects with CKD compared to NAFLD. They further noted that elevated liver fibrosis scores in MAFLD are strongly and independently associated with CKD and abnormal albuminuria. However, their research did not incorporate mortality data to train an AI algorithm, nor did it perform any explainability analysis to assist physician decision-making processes.

Su Lin et al. [25] aimed to validate the diagnostic criteria of MAFLD by comparing it with NAFLD. They assessed clinical parameters of subjects diagnosed with MAFLD, NAFLD, and non-metabolic-dysfunction NAFLD (non-MD-NAFLD), identifying significant differences. Their findings indicated that MAFLD criteria could better identify at-risk subjects, with the MAFLD group exhibiting higher liver enzymes and more severe glucose and lipid metabolism disorders. However, their focus was solely on validating MAFLD diagnostic criteria without exploring other disease aspects or conducting explainability analysis, as addressed in our study.

Decraecker et al. [26] evaluated non-invasive methods to predict mortality in MAFLD subjects. They developed and compared several prognostic models, finding that non-invasive liver fibrosis assessment at baseline correlated well with all-cause and liver-related mortality (C-index $\approx 0.8 - 0.9$; AUC $\approx 0.72 - 0.74$). Unlike our work, they did not utilize variables readily available in healthcare databases to develop these models, nor did they incorporate AI algorithms or explainability analysis that could facilitate physician decision-making in MAFLD subjects.

Nguyen et al. [27] analyzed the initial characteristics and long-term outcomes of three groups of participants diagnosed with fatty

liver disease based on ultrasound findings: those meeting criteria for NAFLD but not MAFLD (non-MAFLD NAFLD), those meeting criteria for both NAFLD and MAFLD (NAFLD-MAFLD), and those meeting criteria for MAFLD but not NAFLD (MAFLD non-NAFLD). They discovered that non-NAFLD MAFLD was independently associated with higher all-cause, cardiovascular-related, and other-cause mortality rates. Importantly, MAFLD subjects without NAFLD had more than twice the mortality risk compared to those with NAFLD without MAFLD. Their study, however, did not explore using AI algorithms to predict mortality or conduct explainability analyses, areas our current research seeks to address.

Kim et al. [28] developed a Cox proportional hazards model to assess all-cause and cause-specific mortality in MAFLD versus NAFLD subjects, adjusting for known risk factors. They reported that MAFLD subjects with advanced fibrosis exhibited a 17% higher all-cause mortality risk, independent of metabolic risk factors, while NAFLD showed no significant association after adjustments. Unlike our study, they did not explore the potential of AI algorithms to predict mortality nor the utility of explainability analysis in clinical decision-making.

In [29], Drozdz et al. employed a machine learning approach to predict mortality risk in MAFLD patients, specifically from cardiovascular diseases (CVD). They trained a logistic regression classifier using either univariate feature ranking or principal component analysis (PCA) to identify high-risk individuals, achieving AUCs up to 0.87. However, their study was limited to CVD-related mortality and did not extend to other potential mortality risks or incorporate explainability analysis.

Huang et al. [30] investigated the differential impact of MAFLD and NAFLD on mortality, finding that MAFLD significantly increases the risk of all-cause mortality compared to NAFLD. Their study highlighted the importance of considering ethnic differences in drug development for MAFLD. However, like others, it lacked the integration of AI-driven tools for mortality prediction and did not offer explainability to aid physician decision-making.

Our research advances the current state-of-the-art by introducing a novel approach that employs AI algorithms and explainability techniques to predict mortality in MAFLD subjects, providing a ready-to-use framework to support physicians in their clinical decision-making process. This is the principal novelty of this study compared with the previous literature on this topic.

3. Proposed framework

In this section, we present the proposed framework as shown in Fig. 1. It consists of four key components: *Data Preparer*, *Predictor*, *Explainer*, and *User Interface*. Each component is designed to execute specific tasks, ensuring a streamlined and efficient workflow from data preparation to user interaction.

3.1. Data preparer

The *Data Preparer* is devoted to transforming raw data into a format suitable for analysis and modeling. This component involves several fundamental steps. Initially, data is acquired from various sources such as population studies and surveys, aiming to gather comprehensive and relevant datasets that can provide valuable insights. Following acquisition, the data undergoes preprocessing to ensure its quality and suitability for further analysis. Preprocessing includes data cleaning to remove noise and inconsistencies, normalization to standardize data ranges, and undersampling to address class imbalances, thereby enhancing the dataset's integrity and usability. The next step focuses on identifying and selecting relevant and informative features from the dataset. By removing non-representative features, the dataset is streamlined, improving the performance and interpretability of the predictive models. The final step in data preparation involves dividing the cleaned and processed dataset into training and test sets. This partitioning is crucial for validating the model's performance and ensuring that it can generalize to new, unseen data.

3.2. Predictor

The aim of the *Predictor* module is to learn the patterns from the prepared data and make predictions. It encompasses several important tasks. Initially, various ML algorithms are utilized to train the predictor on the prepared dataset. The choice of algorithms may vary based on the specific application and data characteristics, ensuring the best possible model performance. To optimize model performance, hyperparameters are fine-tuned through a process that involves adjusting the model parameters to enhance accuracy and efficiency, often employing techniques such as cross-validation. Once trained and optimized, the model is used to generate predictions. These predictions form the basis for subsequent analysis and decision-making, making this step crucial for the framework's overall effectiveness.

3.3. Explainer

This component handles the functions of making the predictions understandable and actionable. This involves a detailed analysis of the trained model to understand the contribution of each feature to the predictions. Techniques such as feature importance and partial dependence plots elucidate how different features influence the model's decisions. Additionally, post-hoc explanation methods provide transparency by explaining the model's predictions after they have been made. These explanations offer insights into the decision-making process, which is essential for validating the model's behavior and ensuring trust in its predictions.

3.4. User Interface

The purpose of the UI is to enhance interaction with the framework, specifically by allowing users to efficiently visualize and interpret the algorithmic results. This component includes a graphical representation of the results using interactive graphs and visualizations, which help users quickly grasp complex data patterns and model insights. Furthermore, the interface enables users to interact with the data, explore different aspects of the model's performance, and gain deeper insights. By providing an intuitive and responsive interface, users can make informed decisions based on the model's predictions.

4. Proposed method

In this section, we summarize the rationale behind the dataset analysis from a given population study. Next, we detail the approach adopted to predict mortality in MAFLD subjects using a relational XAI method. Finally, we discuss the metrics, the results obtained, and the statistical analysis, along with the implementation details of the MORIX framework and its display in the web app. Moreover, for a clearer overview of the workflow, Fig. 2 reports a complete schematization of the MORIX pipeline.

In detail the proposed pipeline consists of several key phases:

- **Data Loading**: Importing data from its source into the pipeline.
- **Data Splitting**: Dividing the data into training and test sets.
- **Data Cleaning**: Removing or correcting erroneous data to improve quality.
- **Data Balancing**: Ensuring an even class distribution is even for better model performance.
- **Model Training**: Using the training data to fit the machine learning model.
- **Prediction**: Generating predictions on new or test data using the trained model.
- **SHAP Data**: Calculating SHAP (SHapley Additive exPlanations) values to interpret and explain model predictions.
- **Visualization**: Creating visual representations of the data and the results for easier interpretation.

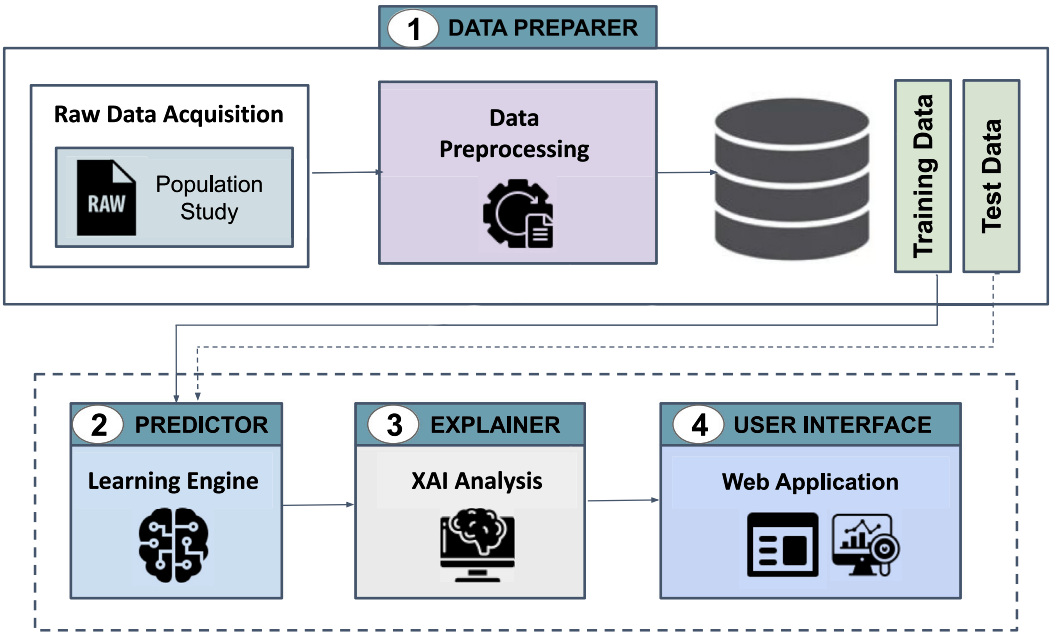


Fig. 1. MORIX classification framework architecture.

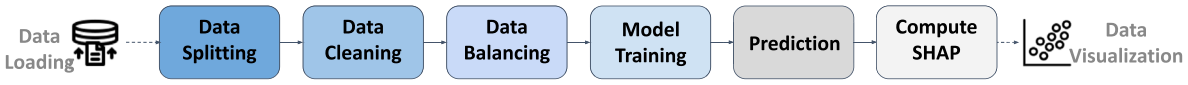


Fig. 2. Representation of the full MORIX pipeline.

4.1. Population study and clinical data collection

The cohort consisted of 1675 subjects aged over 30 years (543 women and 1131 men) diagnosed with MAFLD at recruitment. These participants were drawn from the second follow-up of the MICOL cohort and the NUTRIHEP cohort [31], recruited from January 2005 and monitored until December 31, 2020. The baseline characteristics of these subjects are provided in Supplementary Table 1.

The research was conducted at the National Institute of Gastroenterology and the "S. De Bellis" Research Hospital in Castellana Grotte, Italy. Informed consent was obtained from all participants. All procedures adhered to the ethical standards of the institutional research committee (MICOL study: DDG-CE-589/2004, DDG-CE-782/2013; NUTRIHEP study: DDG-CE-502/2005, DDG-CE-792/2014) and the 1964 Helsinki Declaration.

Participants provided information on sociodemographic characteristics, health status, medical history, tobacco use, dietary habits, education level, employment status, and marital status through interviews.

Weight was measured using an electronic SECA mechanical balance scale,[1] and height was measured with a wall-mounted SECA stadiometer.[2] Blood pressure and BMI were assessed in accordance with international guidelines [32]. Fasting venous blood samples were collected and processed by the central hospital's laboratory following standard procedures [33]. All participants underwent standardized ultrasonography, conducted by two radiologists using a Hitachi H21 Vision machine (Hitachi Medical Corporation, Tokyo, Japan) with a 3.5 MHz transducer. Hepatic steatosis was classified dichotomously as either absent or present.

Furthermore, it is important to highlight that the study population remains relevant and comparable to contemporary cohorts. Bonfiglio et al. [34] underscore the critical role of lifestyle factors (using the same population of our study), particularly dietary habits, in the onset and progression of MAFLD. They observe that the population in the studied geographical area exhibits highly conservative eating patterns, with consistent dietary behaviors persisting across different age groups over time. This stability in dietary habits is a noteworthy characteristic, as it likely contributed to the consistency of findings over the 15-year follow-up period, minimizing the confounding effects of significant dietary changes on MAFLD outcomes.

4.2. Dataset description

Subjects in the study had 30 characteristics, which included continuous and categorical values for biochemical, anthropometric and sociodemographic variables. Specifically, the 30 columns contained the values of: Aspartate Aminotransferase, Education, Job, Marital status, Weight, Hypertension, Blood Fats, Diabetes, Systolic blood pressure, Diastolic blood pressure, Cholesterol, Triglycerides, Blood Glucose, Alkaline Phosphatase, Total Bilirubin, HDL, LDL, Alanine Aminotransferase, Gamma-glutamyl transferase, NAFLD level, C-reactive protein, Age, Smoke, Hepatitis, Homa, MAFLD occurrence, BMI, Total alcohol consumed per day (in grams), Gender, and Status (*Mortality (Yes/No)*). The dataset is represented by tabular data, encompassing 1674 observations across these 30 features, including the target variable. Characteristics of the study population are detailed in the *Supplementary Table* 1 located in the *Appendix*.

4.3. Data pre-processing and features selection analysis

We processed the dataset using the Pandas library in Python,[3] which facilitated the removal of all missing values. Following the cleaning

[1] https://us.secashop.com/products/measuring-stations-and-column-scales/seca-700/

[2] https://uk.secashop.com/products/height-measuring-instruments/seca-206/

[3] https://pandas.pydata.org/

process, the dataset was reduced to 1459 subjects. Given the significant imbalance where the *Mortality (No)* class outnumbered the *Mortality (Yes)* class (1374 to 301), we employed a random undersampling strategy using the imbalanced-learn library,[4] achieving a balanced final dataset of 528 subjects. Indeed, the final number of 528 subjects is in line with the literature [35–37], where other studies use datasets of comparable size while still ensuring robustness and validity of the results.

The data were then standardized using the RobustScaler technique from scikit-learn,[5] which removes the median and scales the data according to the quantile range, thus minimizing the influence of outliers.

Feature selection was performed using the RFE method combined with a RF classifier, integrated within a grid search algorithm to optimize selection.[6] This was conducted with a 5-fold cross-validation strategy to determine the most predictive features while avoiding the problem of double dipping.

The dataset was subsequently split into training, testing, and validation sets. The latter was used to evaluate the performance of the model with the selected features iteratively, aiming to identify the optimal feature set for predicting mortality. By the end of this process, six key features were selected as most indicative of mortality outcomes: Weight, Cholesterol, Blood Glucose, Alkaline Phosphatase, LDL, and Age. The target variable, *Status*, was treated as a binary variable (*Mortality (No)* = 0, *Mortality (Yes)* = 1). Furthermore, all biochemical markers were treated as continuous variables in the models to reflect their natural variation and scale.

4.4. Predictive models

To identify the most effective classifier for predicting mortality in MAFLD patients, we employed the following models: Multilayer Perceptron (MLP) [38]; RF [39]; SVM [40]; Extreme Gradient Boosting (XGBoost) [41]; Light Gradient Boosting Machine (LGBM) [42]. These models were selected in alignment with recommendations in the literature, particularly those by Abdullah Alanazi, who emphasizes their applicability and efficacy in healthcare contexts [43–47].

The predictive models were developed using the Python programming language, specifically utilizing the Scikit-learn library, version 1.4.2.[7] For each algorithm under study, a hyperparameter tuning is performed using a grid search with *5-fold* cross-validation strategy. Upon the end of the process, the framework is capable of predicting the patient's outcome as either *Mortality (No)* or *Mortality (Yes)*. The effectiveness of the models is evaluated based on several metrics, described further forward. These include percentage accuracy, Area Under the Receiver Operating Characteristic Curve (Receiver Operating Characteristics (ROC)), precision, recall, and F1 score.

These metrics serve as the basis for comparing five different algorithms to identify the most effective one for predicting mortality in MAFLD subjects. The comparative analysis and results of these algorithms are presented in Section 5.

4.5. Explainability machine learning

A post-hoc XAI approach [48] is adopted To clarify how features contribute to predicting the target variable, it is essential to recognize the inherent trade-off between model accuracy and interpretability when combining prediction and explanation models. This challenge is addressed through the concept of XAI. Tools like Local Interpretable Model-Agnostic Explanations (LIME) [49] and Shapley Additive exPlanations (SHAP) [50] are capable of providing explanations independent of the predictive model.

In this study, SHAP is utilized to interpret the contribution of features to the prediction. As demonstrated by Lombardi et al. [51], consider D as a dataset of samples $D = \left[\left(x_1, y_1\right), \left(x_2, y_2\right), \ldots, \left(x_z, y_z\right)\right]$, where x_i denotes the feature vector for sample i and y_i denotes the label for sample i. Let f be a classifier, and f_{x_i} the prediction for the test instance i, corresponding to the predicted label. The aim is to elucidate the contribution of each feature j among the S features as the average marginal contribution of the feature value across all possible coalitions, *i.e.*, all possible subsets of feature values with and without the feature j. A coalition, F, is defined as a subset of S, such that $F \subseteq S$.

Given that $f_{x_i}(F)$ is the prediction for f_{x_i} given F, the marginal contribution of adding the jth feature value to F is expressed as:

$$\left[f_{x_i}(F \cup j) - f_{x_i}(F)\right] \tag{1}$$

To calculate the exact Shapley value, all possible subsets of feature values excluding the jth feature value, $F \subseteq S - \{j\}$, must be considered:

$$\phi(f, x) = \sum_{F \subseteq S \setminus \{j\}} \frac{|F|!\left(|S| - |F| - 1\right)!}{|S|!} \left[f_{x_i}(F \cup \{j\}) - f_{x_i}(F)\right] \tag{2}$$

Here, $|F|!$ represents the number of permutations of feature values that precede the jth feature, $(|S| - |F| - 1)!$ represents the number of permutations of feature values that follow the jth feature, and $|S|!$ is the total number of permutations [50]. The term $\phi(f, x)$, known as the *SHAP value of a single feature*, corresponds to the Shapley value in game theory [50].

The Shapley value quantifies each player's contribution in a cooperative game with multiple players. SHAP determines the Shapley value for each feature as if it were a player in the learned model. SHAP values are computed across all feature combinations, which requires exponential time. The results of the post-hoc explanation analysis conducted in this study are presented in Section 5.4.

4.6. MORIX pipeline formalization

A description of the operations performed is provided in the Algorithm 1. It details all the previously described steps useful for the development of the MORIX approach. Specifically, the algorithm follows a systematic process to prepare the dataset, train the model, and analyze the results. Initially, the dataset D is split into training and test sets during the **Setup Phase**. This is followed by a preprocessing step where records with missing values are removed from both the training and test sets. To ensure that the model is trained on a well-balanced dataset, the training set is then balanced to equalize the number of records for each class. This balancing step is crucial for improving model performance and fairness. Before training begins, an undersampling approach is applied to the dataset. Indeed, this step helps to mitigate class imbalance by reducing the number of samples from the majority class, which is typically performed if the dataset is skewed towards one class. The balanced dataset is then used to train the model. After training, the **Summarization Phase** involves computing SHAP (SHapley Additive exPlanations) values for each test instance. These values help to interpret and explain the model's predictions. The algorithm then stores the computed SHAP values along with the test instances and their corresponding predictions. Finally, in the **Finalization Phase**, the results from the predictions and SHAP values are compiled and summarized. This ensures that the summarized data is made available for further use or analysis, providing insights into the model's behavior and its decision-making process.

[4] https://imbalanced-learn.org/stable/index.html
[5] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html
[6] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
[7] http://scikit-learn.org

Algorithm 1 Algorithm of MORIX framework

Functions:
splitting_dataset(D): Function to split the dataset D.
clean_data(D): Function to remove records with missing values from D.
balance_datasets(D): Function to equalize the number of records for each class in D.
train_model(D): Function to train a model on the dataset D.
predict_outcomes($model, D$): Function to predict outcomes using the model on the dataset D.
compute_shap($model, instance$): Function to calculate SHAP values for the instance using the model.

Inputs:
D: Original dataset.

Outputs:
predictions: Predicted outcomes for the test set.
shap_values: SHAP values for the test instances.

procedure SETUP PHASE
    $D_{train}, D_{test} \leftarrow splitting\_dataset(D)$
    $D_{train} \leftarrow clean\_data(D_{train})$
    $D_{test} \leftarrow clean\_data(D_{test})$
    $D_{train} \leftarrow balance\_datasets(D_{train})$
    $D_{test} \leftarrow balance\_datasets(D_{test})$
    $model \leftarrow train\_model(D_{train})$
    $predictions \leftarrow predict\_outcomes(model, D_{test})$
end procedure

procedure SUMMARIZATION PHASE
    for each instance in D_{test} **do**
        $shap\_value \leftarrow compute\_shap(model, instance)$
        Store instance, prediction, and $shap\_value$
    end for
end procedure

procedure FINALIZATION PHASE
    Make the summarized data available for further use or analysis.
end procedure

5. Experimental results

This section presents and discusses the experimental evaluation results of our model for predicting mortality in MAFLD subjects.

Dataset Splitting. The dataset is split into training and testing sets using the standard 80/20 method. For reproducibility, the Scikit-learn library's train-test split function is employed with a random seed set to 42.[8]

Decision-Maker Hyperparameter Tuning and Optimization. The classifiers employed in this study include MLP, RF, SVM, XGBoost, and LGBM. Details on these algorithms can be found in Section 4.4. Hyperparameter tuning for these classifiers is performed using a grid search strategy, which is documented for reproducibility.[9] This tuning process utilizes a 5-fold cross-validation method, with accuracy as the evaluation metric. The specific values of the hyperparameters explored are listed in Table 1 for clarity and reproducibility.

To determine the most effective predictive model, we assessed the performance of various algorithms using the metrics outlined in Section 5.1. The results are summarized in Table 2, which presents the performance of each algorithm in predicting mortality.

Additionally, Confusion Matrices (CMs). are utilized to illustrate the classification accuracy during the testing phase, particularly focusing on the number of subjects misclassified by each algorithm. These matrices are depicted in Fig. 3, providing a visual representation of the model's performance in terms of true positives, true negatives, false positives, and false negatives.

Finally, to determine whether the differences in the accuracy of the various models were statistically significant, we performed a statistical analysis.

5.1. Evaluation metrics

In this section, metrics based on accuracy are presented. These metrics, which primarily rely on the CM, are essential for evaluating the performance of classification models.

Accuracy measures the overall proportion of correct predictions across the entire dataset:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

where TP, TN, FP, and FN denote the numbers of true positive, true negative, false positive, and false negative predictions, respectively.

Recall indicates the fraction of actual positive cases that are correctly classified:

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

Precision assesses the ratio of correct positive predictions to the total predicted positives:

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

F1 Score calculates the harmonic mean of precision and recall, providing a balance between the two when the dataset is imbalanced:

$$F1 = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

The F1 score integrates both precision and recall into a unified metric, making it particularly valuable when handling imbalanced datasets.

Area Under the Receiver Operating Characteristic Curve (AUC) assesses the model's capability to differentiate between classes. It is particularly useful for binary classification problems:

$$AUC = \frac{\sum_{x^- \in X^-} \sum_{x^+ \in X^+} \mathbb{1}(f(x^-) < f(x^+))}{|X^-| \, |X^+|} \tag{7}$$

where

$$\mathbb{1}(\cdot) = \begin{cases} 1 & \text{if } f(x^-) < f(x^+) \\ 0 & \text{otherwise} \end{cases}$$

X^+ is the set of positive samples, X^- is the set of negative samples, $f(\cdot)$ is the result of the model prediction, and $\mathbb{1}(\cdot)$ is an indicator function [52].

5.2. Best model performance evaluation

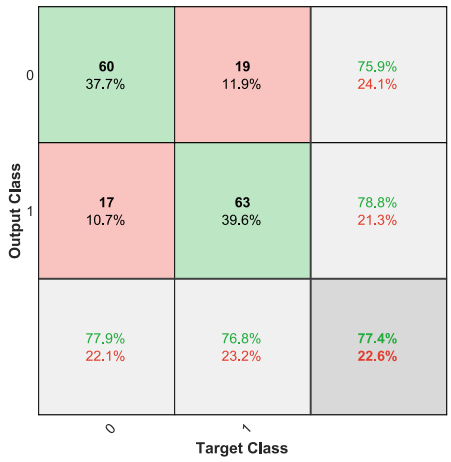
In this section, a performance comparison of the considered classifiers was carried out. The results are illustrated using CMs and performance metrics as shown in Fig. 3 and Table 2.

In particular, the RF model demonstrates superior performance. As detailed in Fig. 3(b) and Table 2, RF achieves the highest recall (0.84) and F1 score (0.83) for the *Mortality (Yes)* class, alongside the highest precision (0.83) and F1 score (0.83) for the *Mortality (No)* class. It also registers the fewest misclassifications for the *Mortality (Yes)* class and boasts the best accuracy on the test set (0.83), highlighting its effectiveness for this task.

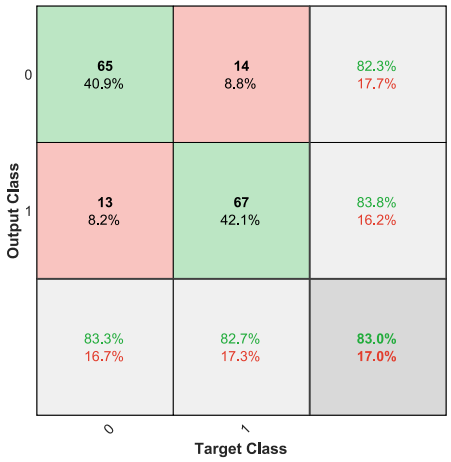
Conversely, the XGBoost (XGB) model records the best precision (0.84) for predicting *Mortality (Yes)* and the best recall (0.85) for *Mortality (No)*, with Fig. 3(d) showing the lowest number of misclassifications for the *Mortality (No)* class.

[8] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

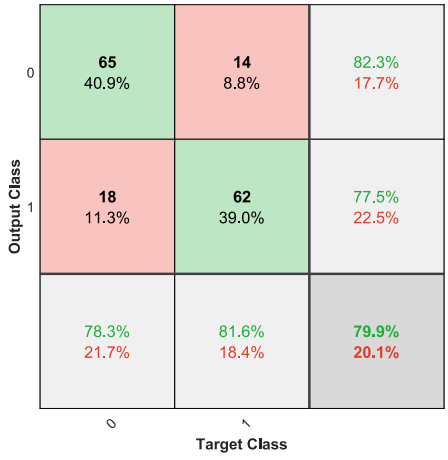
[9] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html



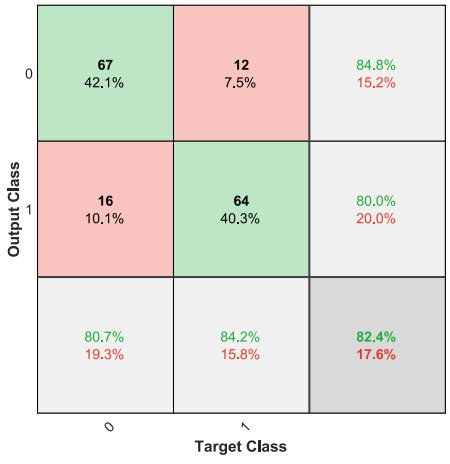
(a) Confusion Matrix of MLP Classifier



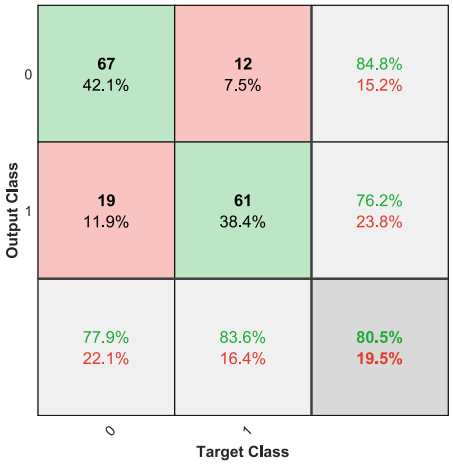
(b) Confusion Matrix of RF Classifier



(c) Confusion Matrix of SVM Classifier



(d) Confusion Matrix of XGB Classifier



(e) Confusion Matrix of LGBM Classifier

Fig. 3. Depiction of confusion matrix values during the evaluation of LGBM, MLP, RF, SVM, and XGB classifiers. 0 represents *Mortality (No)*, while 1 represents *Mortality (Yes)*.

Table 1
List of hyperparameters, their corresponding values, and types for the classification models presented in this study.

Algorithm	Hyperparameter	Values	Type
Multilayer Perceptron	random_state	{42}	Integer
	hidden_layer_sizes	{(50,), (100,), (50, 50), (100, 150), (100, 100)}	Tuple of Integers
	activation	{relu, tanh, logistic}	String
	solver	{sgd, adam}	String
	learning_rate	{constant, adaptive}	String
Random Forest	random_state	{42}	Integer
	n_estimators	{50, 100, 200}	Integer
	max_depth	{10, 20, 30, None}	Integer or None
	min_samples_split	{2, 5, 10}	Integer
	min_samples_leaf	{1, 2, 4}	Integer
Support Vector Machines	random_state	{42}	Integer
	C	{0.1, 1, 10, 100}	Float
	gamma	{0.1, 0.01, 0.001}	Float
	kernel	{linear, rbf, poly}	String
eXtreme Gradient Boosting	random_state	{42}	Integer
	max_depth	{3, 4, 5}	Integer
	learning_rate	{0.1, 0.01, 0.001}	Float
	n_estimators	{100, 200, 300}	Integer
Light Gradient Boosting	random_state	{42}	Integer
	learning_rate	{0.1, 0.01, 0.001}	Float
	num_leaves	{31, 50, 100}	Integer
	max_depth	{-1, 10, 20, 30, 300}	Integer

Table 2
Mortality prediction results for the evaluated algorithms. "No/Yes" indicates mortality status. For each value, the lower and upper bounds of the 95% CI for each considered metric are provided in brackets.

		Classifier				
		RF	XGB	MLP	LGBM	SVM
Precision	No	0.83 (0.74, 0.91)	0.81 (0.70, 0.85)	0.78 (0.68, 0.84)	0.78 (0.69, 0.83)	0.78 (0.68, 0.82)
	Yes	0.83 (0.74, 0.90)	0.84 (0.72, 0.87)	0.77 (0.66, 0.81)	0.84 (0.71, 0.86)	0.82 (0.70, 0.85)
Recall	No	0.82 (0.74, 0.90)	0.85 (0.73, 0.89)	0.76 (0.65, 0.82)	0.85 (0.72, 0.87)	0.82 (0.71, 0.85)
	Yes	0.84 (0.75, 0.91)	0.80 (0.68, 0.86)	0.79 (0.67, 0.83)	0.76 (0.65, 0.81)	0.78 (0.67, 0.82)
F1 score	No	0.83 (0.76, 0.88)	0.83 (0.72, 0.86)	0.77 (0.66, 0.82)	0.81 (0.70, 0.84)	0.80 (0.69, 0.83)
	Yes	0.83 (0.76, 0.89)	0.82 (0.71, 0.85)	0.78 (0.67, 0.82)	0.80 (0.69, 0.84)	0.79 (0.68, 0.82)
Accuracy		0.83 (0.77, 0.88)	0.82 (0.72, 0.85)	0.77 (0.67, 0.81)	0.81 (0.70, 0.84)	0.80 (0.69, 0.83)
AUC		0.88 (0.82, 0.93)	0.88 (0.80, 0.91)	0.88 (0.78, 0.90)	0.88 (0.78, 0.89)	0.86 (0.77, 0.88)

While the Multilayer Perceptron (MLP), Light Gradient Boosting Machine (LGBM), and Support Vector Machine (SVM) models perform well on average, they do not exhibit outstanding metrics compared to RF and XGB, except for SVM, which achieves the highest Area Under the Receiver Operating Characteristic Curve (AUC) of 0.89. However, SVM is noted for its balanced performance and minimal misclassifications across both classes, making it the most reliable among the three. In contrast, LGBM is identified as the least effective model, with a high number of misclassifications, particularly for the *Mortality (Yes)* class. Given the primary goal of this work — to accurately identify subjects at the highest risk of mortality for early physician intervention — the model with the fewest errors in predicting *Mortality (Yes)* is considered the most suitable. Therefore, RF, with only 13 misclassified samples as seen in Fig. 3(b), is proposed as the best model.

Specifically, the best model utilized the following hyperparameters: `max_depth` = 10, `min_samples_leaf` = 4, `min_samples_split` = 10, `n_estimators` = 100, and `random_state` = 42.

Fig. 4 presents the ROC curves for each classifier. Although RF does not have the highest AUC when compared with SVM, its AUC score confirms its reliable performance in correctly classifying both *Mortality (No)* and *Mortality (Yes)* classes.

5.3. Statistical analysis

To assess and compare the performance of the different ML models, a statistical analysis is carried out. The primary objective was to determine whether the observed model accuracy differences were statistically significant. The models evaluated included Random Forest (RF), Support Vector Machine (SVM), XGBoost, Multi-Layer Perceptron (MLP), and LightGBM.

To assess the robustness and significance of differences in model accuracies, a *bootstrap analysis* [53] was conducted. Therefore, for each model considered, 1000 bootstrap samples were generated from the test set, and the accuracy of each model was computed for each bootstrap sample.

Moreover, based on previous evaluations, the Random Forest (RF) model was identified as the best-performing model. For each of the other models (SVM, XGBoost, MLP, LightGBM), a *paired t-test* was conducted to compare the bootstrap accuracies against those of the RF model. The following metrics were computed for each comparison:

- **T-statistic**: A measure of the difference in accuracies between the models.
- **P-value**: Indicates the statistical significance of the observed differences.
- **Mean Difference**: The average difference in accuracies between the Random Forest model and the compared model.
- **95% Confidence Interval (CI)**: The lower and upper bounds of the 95% CI for the mean difference.

As shown in Table 3, the statistical analysis confirmed that the differences in accuracies between the RF model and the others were

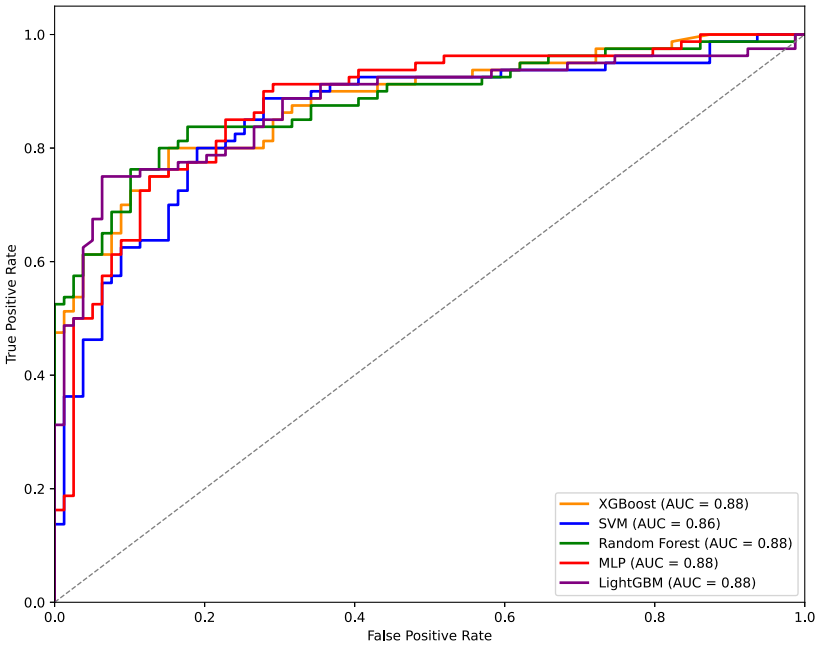


Fig. 4. Representation of ROC curve and AUC for each model under study.

Table 3
Results of the t-test on bootstrapped datasets comparing the best model (Random Forest) vs the best configurations of all other considered models.

Model	T-statistic	*P*-value	Mean difference (95% CI)
XGBoost	2.62	<0.01	0.004 (−0.082, 0.082)
MLP	20.03	<0.01	0.028 (−0.057, 0.113)
LightGBM	17.20	<0.01	0.023 (−0.063, 0.101)
SVM	19.94	<0.01	0.028 (−0.057, 0.113)

statistically significant (P-value < 0.01), as determined by the statistical analysis. Figs. 5 and 6 show the distributions of bootstrap accuracies and their relative differences. Therefore, it demonstrates the superior performance of the RF model compared to other considered models.

5.4. Explainability results

To elucidate the decision-making process of the chosen best model, an XAI analysis is conducted. The goal of XAI analysis is to enhance the transparency of the model by analyzing and comparing its decision strategy. This facilitates a deeper understanding of the algorithmic reasoning and the specific contributions of each feature to the prediction outcome.

XAI analysis is performed using the SHAP (SHapley Additive exPlanations) library in Python,[10] which is applied post-hoc to the RF model detailed in Section 5.2. Fig. 7 displays a global synthesis diagram generated from SHAP values, illustrating the influence of anthropometric and biochemical parameters on model predictions.

The diagram portrays the feature importance in descending order (from top to bottom). The horizontal position of each dot indicates whether the feature influences the prediction towards the positive or negative class. The color coding (red for positive and blue for negative) further helps in visualizing the direction of each feature's impact on the prediction outcome.

Shapley values are pivotal as they provide insights into the prediction direction from a clinical perspective. As depicted in Fig. 7, the features Age and $Weight$ emerge as the most and least significant, respectively, indicating that Age is the predominant variable influencing mortality risk prediction.

This analysis confirms that the model has accurately identified Age as the most critical variable for the task at hand, aligning with clinical expectations and enhancing trust in the model's predictive capabilities.

To deepen our understanding of how individual features impact the prediction of outcomes, an extended XAI analysis was conducted. This analysis focuses on the contribution of each feature during the prediction phase and evaluates their explanatory power within the model.

Fig. 8 illustrates the incremental impact of key variables on the model's explanatory power. Notably, the variable Age alone accounts for 60% of the explanation provided by the model. When the variable $Blood\ Glucose$ is added, the combined features explain over 70% of the model. Incorporating the remaining features, including LDL, $Alkaline\ Phosphatase$, $Cholesterol$, and $Weight$, brings the total explained variance to 100%.

This analysis underscores that the selected set of features is optimal for the task at hand. It demonstrates that these features collectively provide a complete and nuanced understanding of the model, affirming their appropriateness and efficacy in capturing the dynamics of the predicted outcomes.

To further assess the performance of the RF model, we examined its predictive behavior on two randomly selected subjects from the test set, representing the *Mortality (No)* and *Mortality (Yes)* classes, respectively. Fig. 9 illustrates the SHAP value analysis for a prediction of *Mortality (Yes)*. The figure displays positive and negative SHAP values on the left and right sides, respectively, highlighting the competing influences on the model's decision.

[10] https://shap.readthedocs.io/en/latest/index.html

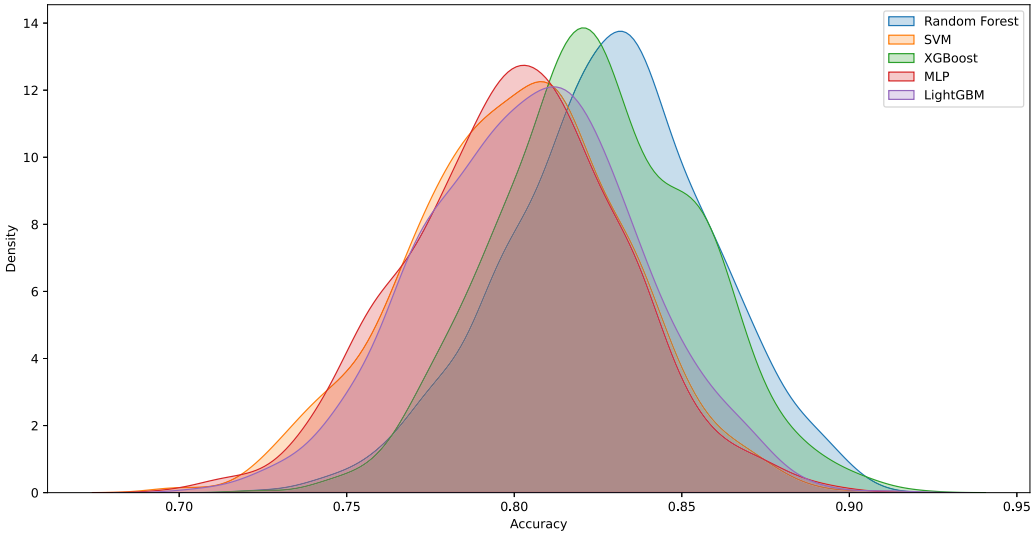


Fig. 5. Distribution of bootstrapped accuracies for the compared models.

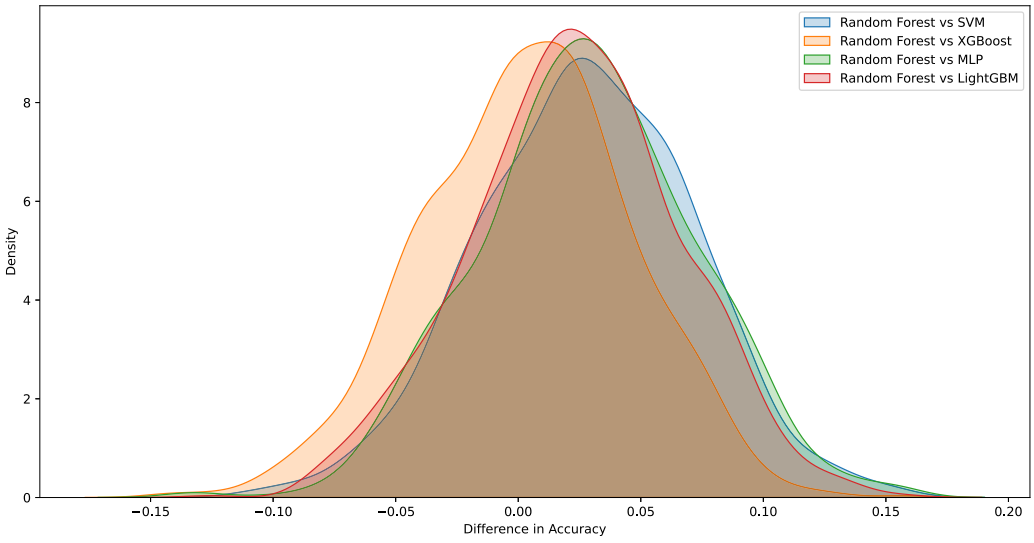


Fig. 6. Distribution of differences in bootstrapped accuracies comparing the best model (Random Forest) against other models.

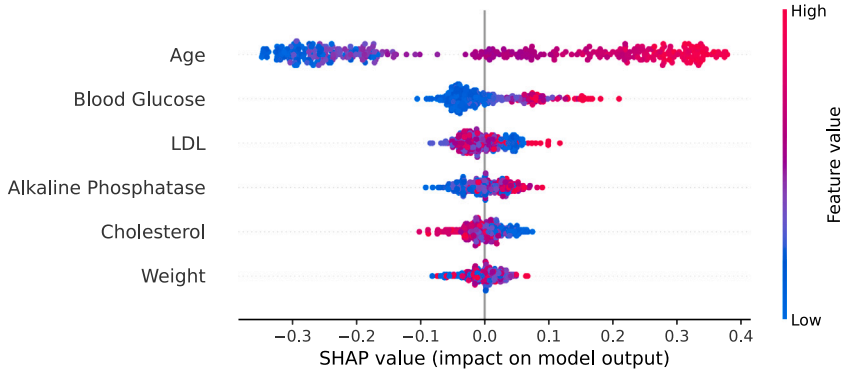


Fig. 7. Visualization of Shapley values for the anthropometric and biochemical parameters analyzed.

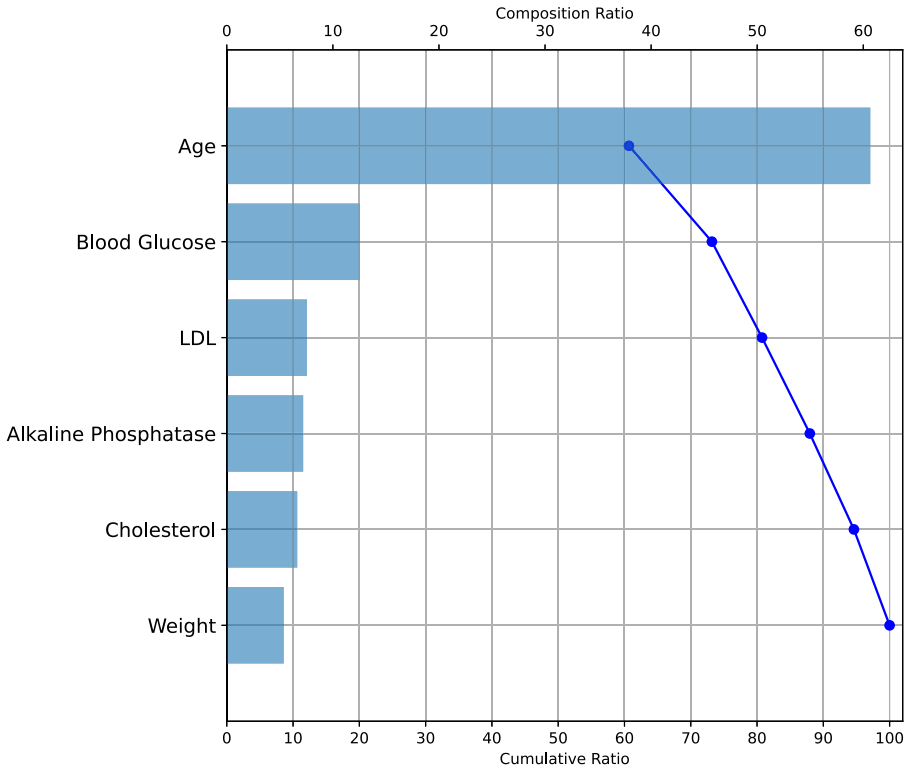


Fig. 8. Percentage of the incremental impact of key variables on the model's explanatory power.

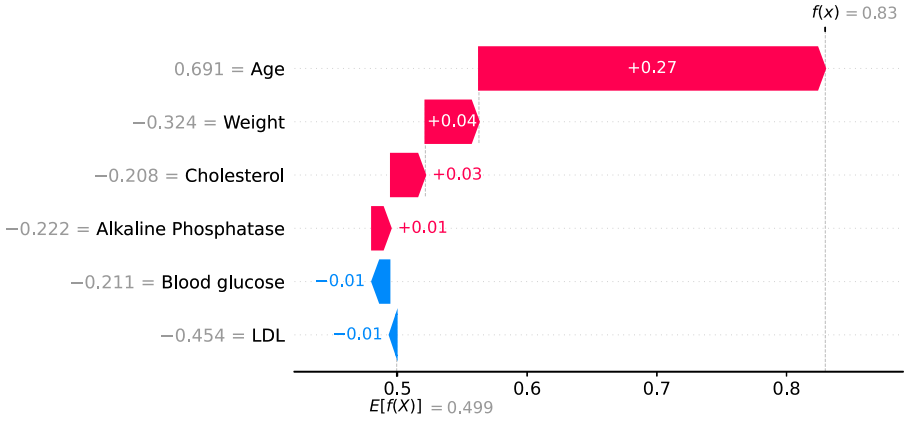


Fig. 9. SHAP value analysis demonstrating the influence of various features on the prediction of *Mortality (Yes)* for a test subject using the RF model.

As depicted in Fig. 9, features such as *Age*, *Weight*, *Cholesterol*, and *Alkaline Phosphatase* positively influence the prediction, implying a higher likelihood of mortality. Conversely, *Blood Glucose* and *LDL* negatively impact the prediction, suggesting a lower likelihood of mortality.

Similarly, Fig. 10 shows the SHAP value analysis for a prediction of *Mortality (No)*, indicating which features predominantly influence this outcome.

In Fig. 10, the features *Age*, *Blood Glucose*, *LDL*, and *Cholesterol* negatively influence the prediction of *Mortality (No)*, whereas *Weight* and *Alkaline Phosphatase* exert a positive influence, suggesting a deviation from the expected mortality risk. This dual analysis using SHAP values provides critical insights into how specific features drive the predictions made by the RF model, offering valuable explanations for each predicted class.

6. MORIX Web User Interface

After identifying the best RF model and conducting XAI analysis, the RF model was exported through Python's pickle[11] library. This approach ensures that the model does not require retraining for each new prediction, thereby enhancing efficiency. Furthermore, we developed a user-friendly web application using HTML and CSS [54] (Fig. 11), which allows physicians to interact seamlessly with the model loaded in the backend using Python's joblib[12] library. This setup ensures robust performance during real-time predictions.

The interface of the user web application is designed to be intuitive, guiding physicians through the process of entering data and receiving predictions. Indeed, it not only provides mortality risk predictions and their probability but also visualizes the impact of each variable on the

[11] https://docs.python.org/3/library/pickle.html
[12] https://joblib.readthedocs.io/en/stable/

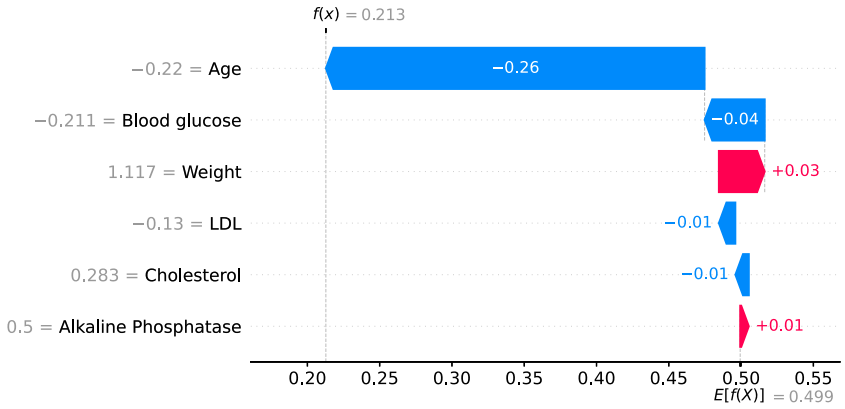


Fig. 10. SHAP value analysis demonstrating the influence of various features on the prediction of *Mortality (No)* for a test subject using the RF model.

prediction outcome through SHAP Waterfall plots, thereby promoting transparency and trust in the model's decisions. Example scenarios are outlined below:

1. **High-Risk Prediction:** On the one hand when the model predicts a high risk of mortality, as indicated by a 97% probability due to elevated factors like high blood glucose levels (Fig. 12), physicians receive a detailed breakdown. The SHAP plot identifies high glucose as a critical factor. This prompts the physician to prioritize glucose management, potentially adjusting medications, recommending lifestyle changes, and scheduling frequent follow-ups to closely monitor the patient's condition.
2. **Low-Risk Prediction:** On the other hand, if the model indicates a low risk of mortality with a 77% probability, showing stable indicators such as lower age and balanced cholesterol levels (Fig. 13), the physician might opt for a less aggressive approach. This scenario suggests continuing with routine monitoring and perhaps focusing on preventive measures to maintain current health levels. The SHAP plot serves as a reassurance that the current health management strategies are effective, reducing the need for immediate interventions.

As previously described, both scenarios illustrate the application's potential to aid in clinical decision-making by providing physicians with a data-driven tool that complements their expertise and enhances patient care strategies. This approach not only improves efficiency but also ensures personalized care by tailoring medical interventions based on individual risk profiles generated by the model.

The integration of the web application with the trained RF model was implemented using python's flask[13] library. A flask object handles requests from the web, serving an HTML file that facilitates interaction with the RF algorithm.

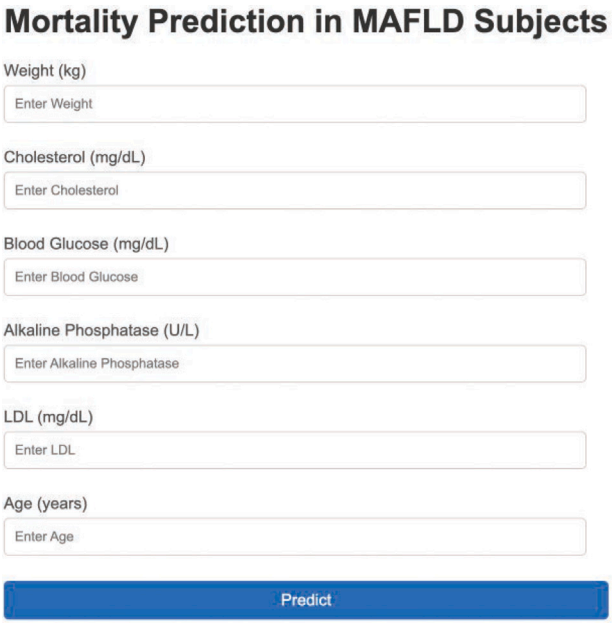


Fig. 11. Home screen of MORIX web user interface able to represent the mortality risk Mortality (No/Yes) for a subject with MAFLD.

7. Conclusion and further directions

In this paper, we proposed MORIX, a framework that seamlessly integrates ML and XAI to predict mortality in MAFLD subjects.

The framework is composed of four main components: (1) a *Data preparer* that handles data preprocessing operations; an (2) *Predictor* that executes predictions of Mortality Risk using ML techniques; (3) a *Explainer* that interprets the best-model results and provides explanations for the model's predictions, and (4) a *User Interface* that provides a graphical representation of training and explanation results via a web app. Therefore, it allows physicians to interact with the AI algorithm.

We conducted a comparison between five ML models to ascertain the optimal model for predicting the *Mortality (Yes)* or *Mortality (No)* classes and implemented an explainability analysis to delineate how each feature contributes to the predictions, aiding physicians in decision-making.

Our approach included experiments to confirm that the chosen features accurately represent mortality risk, ensuring that significant predictors are used without redundancy to avoid the double-dipping problem.

The experimental results highlight the RF model's superior performance, achieving precision, recall, and F1 scores of (0.83, 0.82, 0.83) for the *Mortality (No)* class, and (0.83, 0.84, 0.83) for the *Mortality (Yes)* class, respectively. Additionally, the best model attained an accuracy of 0.83 and an AUC of 0.88.

This method enables physicians to proactively recommend lifestyle changes, such as diet and physical activity, in alignment with the findings of Curci et al. [22], while also relying on the model explanation.

However, our current approach has limitations. The dataset used was initially unbalanced, which hindered the classification performance. We addressed this issue by applying an undersampling technique to balance the dataset, significantly improving the model's performance. Furthermore, despite performing undersampling, all 528 samples are representative of the population. This population is homogeneous, as all residents follow the Mediterranean diet and reside in

[13] https://flask.palletsprojects.com/en/3.0.x/

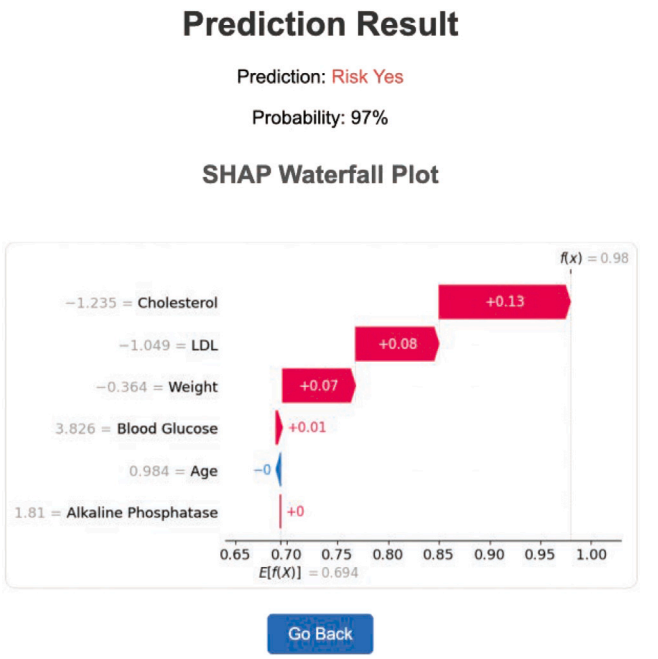


Fig. 12. Representation of the MORIX web user interface, representing the Risk Mortality (Yes), with a probability of 97%, for a subject with MAFLD.

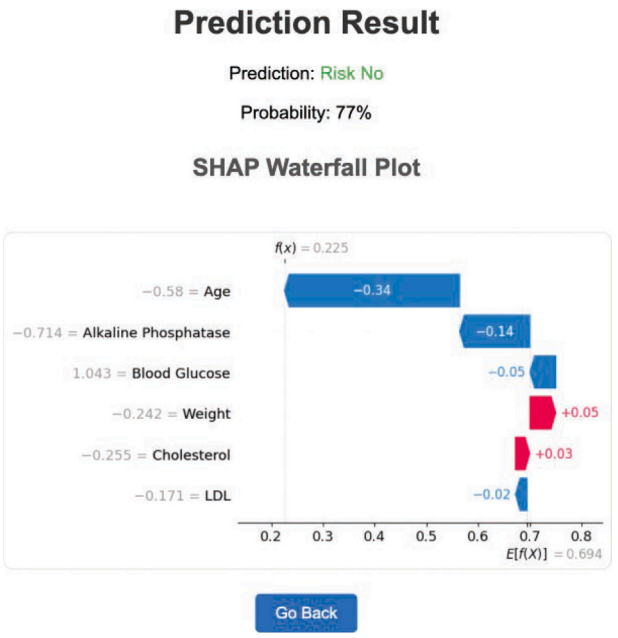


Fig. 13. Representation of the MORIX web user interface, representing the Risk Mortality (No), with a probability of 77%, for a subject with MAFLD.

the cities of Castellana and Putignano (Italy), constituting a unified demographic population [18] and [6]. Currently, no comparable studies are developing a MAFLD mortality classifier, which limits our ability to benchmark against other models.

Future work should focus on expanding the dataset size and diversity to enhance model robustness and generalizability also in other kinds of population. Efforts will include incorporating data from local hospitals and research institutions, allowing the dataset to better reflect a broader population and capture a wider range of demographic and clinical characteristics. Additionally, the exploration of Deep Learning (DL) algorithms could offer new insights and improve predictive accuracy, particularly in predicting mortality in MAFLD subjects. Using methods such as bootstrapping or cross-validation has mitigated some limitations by assessing model stability; however, larger and more varied datasets will further enhance confidence in these findings. Expanding the dataset and applying advanced DL techniques will not only strengthen the reliability of our results but also support the development of predictive models with broader clinical applicability. Furthermore, we will explore the relationships between the disease and its risk factors, which, while beyond the scope of the present study, represent a promising direction for future research.

Regarding the developed web application, improvements will include: (a) optimizing the backend to handle requests even more efficiently, possibly by exploring the use of additional lightweight web frameworks alongside Flask to enhance server-side operations; (b) improving data management capabilities by integrating advanced data validation techniques and incorporating real-time data processing to ensure accuracy and responsiveness before making predictions, and (c) strengthening security measures by incorporating the latest advances in cybersecurity to protect user data and model integrity, while continuously mitigating risks such as data breaches and unauthorized access.

Ethics statement

All procedures adhered to the ethical standards of the institutional research committee (MICOL study: DDG-CE-589/2004, DDG-CE-782/2013; NUTRIHEP study: DDGCE- 502/2005, DDG-CE-792/2014) and the 1964 Helsinki Declaration.

CRediT authorship contribution statement

Domenico Lofù: Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Conceptualization. **Paolo Sorino:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Tommaso Colafiglio:** Visualization, Supervision, Conceptualization. **Caterina Bonfiglio:** Writing – review & editing, Resources, Funding acquisition, Data curation. **Rossella Donghia:** Writing – original draft, Resources, Project administration, Funding acquisition, Data curation. **Gianluigi Giannelli:** Writing – original draft, Visualization, Validation, Resources, Project administration, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Angela Lombardi:** Visualization, Validation, Supervision. **Tommaso Di Noia:** Writing – original draft, Visualization, Validation, Supervision, Investigation, Conceptualization. **Eugenio Di Sciascio:** Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation. **Fedelucio Narducci:** Writing – original draft, Visualization, Validation, Supervision, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Micol and the Nutriep group. This work was partially supported by the following projects: “LIFE: the itaLian system wIde Frailty nEtwork”, PNRR-MAD-2022-12376656, CTEMT - ‘Casa delle Tecnologie Emergenti di Matera’. “IDENTITA - rete Integrata meDiterranea per losservazione ed Elaborazione di percorsi di Nutrizione”. Also this work has been carried out while Tommaso Colafiglio was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome in collaboration with Politecnico di Bari.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cmpbup.2024.100176.

References

[1] Y. Fazel, A.B. Koenig, M. Sayiner, Z.D. Goodman, Z.M. Younossi, Epidemiology and natural history of non-alcoholic fatty liver disease, Metabolism 65 (8) (2016) 1017–1025.

[2] A.P. Levene, R.D. Goldin, The epidemiology, pathogenesis and histopathology of fatty liver disease, Histopathology 61 (2) (2012) 141–152.

[3] Z.M. Younossi, A.B. Koenig, D. Abdelatif, Y. Fazel, L. Henry, M. Wymer, Global epidemiology of nonalcoholic fatty liver disease—meta-analytic assessment of prevalence, incidence, and outcomes, Hepatology 64 (1) (2016) 73–84.

[4] R. Cozzolongo, A.R. Osella, S. Elba, J. Petruzzi, G. Buongiorno, V. Giannuzzi, G. Leone, C. Bonfiglio, E. Lanzilotta, O.G. Manghisi, et al., Epidemiology of HCV infection in the general population: a survey in a southern Italian town, Off. J. Am. Coll. Gastroenterol. ACG 104 (11) (2009) 2740–2746.

[5] A.R. Osella, M. del Pilar Díaz, R. Cozzolongo, C. Bonfiglio, I. Franco, D.I. Abbrescia, A. Bianco, G. Buongiorno, S. Elba, J. Petruzzi, et al., Overweight and obesity in southern Italy: their association with social and life-style characteristics and their effect on levels of biologic markers., Rev. Fac. Cien. Méd. Córdoba 71 (3) (2014).

[6] G. Misciagna, M. Del Pilar Diaz, D. Caramia, C. Bonfiglio, I. Franco, M. Noviello, M. Chiloiro, D. Abbrescia, A. Mirizzi, M. Tanzi, et al., Effect of a low glycemic index mediterranean diet on non-alcoholic fatty liver disease. a randomized controlled clinici trial, J. Nutr. Health Aging 21 (4) (2017) 404–412.

[7] F. Procino, G. Misciagna, N. Veronese, M.G. Caruso, M. Chiloiro, A.M. Cisternino, M. Notarnicola, C. Bonfiglio, I. Bruno, C. Buongiorno, et al., Reducing NAFLD-screening time: A comparative study of eight diagnostic methods offering an alternative to ultrasound scans, Liver Int. 39 (1) (2019) 187–196.

[8] M. Eslam, A.J. Sanyal, J. George, A. Sanyal, B. Neuschwander-Tetri, C. Tiribelli, D.E. Kleiner, E. Brunt, E. Bugianesi, H. Yki-Järvinen, et al., MAFLD: a consensus-driven proposed nomenclature for metabolic associated fatty liver disease, Gastroenterology 158 (7) (2020) 1999–2014.

[9] M. Eslam, P.N. Newsome, S.K. Sarin, Q.M. Anstee, G. Targher, M. Romero-Gomez, S. Zelber-Sagi, V.W.-S. Wong, J.-F. Dufour, J.M. Schattenberg, et al., A new definition for metabolic dysfunction-associated fatty liver disease: An international expert consensus statement, J. Hepatol. 73 (1) (2020) 202–209.

[10] G. Semmler, S. Wernly, S. Bachmayer, I. Leitner, B. Wernly, M. Egger, L. Schwenoha, L. Datz, L. Balcar, M. Semmler, et al., Metabolic dysfunction-associated fatty liver disease (MAFLD)—rather a bystander than a driver of mortality, J. Clin. Endocrinol. Metab. 106 (9) (2021) 2670–2677.

[11] G. Casalino, G. Castellano, A. Consiglio, N. Nuzziello, G. Vessio, Microrna expression classification for pediatric multiple sclerosis identification, J. Ambient Intell. Humaniz. Comput. (2021) 1–10.

[12] F. Castellana, S. Aresta, P. Sorino, I. Bortone, D. Lofù, F. Narducci, T. Di Noia, E. Di Sciascio, R. Sardone, An artificial neural network model to assess nutritional factors associated with frailty in the aging population from southern Italy, in: 2022 IEEE International Conference on Systems, Man, and Cybernetics, SMC, 2022, pp. 3228–3233.

[13] E. Lella, A. Pazienza, D. Lofù, R. Anglani, F. Vitulano, An ensemble learning approach based on diffusion tensor imaging measures for Alzheimer's disease classification, Electronics 10 (3) (2021) 249.

[14] G. Casalino, G. Castellano, G. Zaza, Evaluating the robustness of a contact-less mhealth solution for personal and remote monitoring of blood oxygen saturation, J. Ambient Intell. Humaniz. Comput. (2022) 1–10.

[15] F. Calivà, N.K. Namiri, M. Dubreuil, V. Pedoia, E. Ozhinsky, S. Majumdar, Studying osteoarthritis with artificial intelligence applied to magnetic resonance imaging, Nat. Rev. Rheumatol. 18 (2) (2022) 112–121.

[16] P. Sorino, V. Paparella, D. Lofu, T. Colafiglio, E. Di Sciascio, F. Narducci, R. Sardone, T. Di Noia, A Pareto-optimality-based approach for selecting the best machine learning models in mild cognitive impairment prediction, in: 2023 IEEE International Conference on Systems, Man, and Cybernetics, SMC, IEEE, 2023, pp. 3822–3827.

[17] Y.-T. Shen, L. Chen, W.-W. Yue, H.-X. Xu, Artificial intelligence in ultrasound, Eur. J. Radiol. 139 (2021) 109717.

[18] P. Sorino, M.G. Caruso, G. Misciagna, C. Bonfiglio, A. Campanella, A. Mirizzi, I. Franco, A. Bianco, C. Buongiorno, R. Liuzzi, et al., Selecting the best machine learning algorithm to support the diagnosis of non-alcoholic fatty liver disease: A meta learner study, PLoS One 15 (10) (2020) e0240867.

[19] P. Sorino, A. Campanella, C. Bonfiglio, A. Mirizzi, I. Franco, A. Bianco, M.G. Caruso, G. Misciagna, L.R. Aballay, C. Buongiorno, et al., Development and validation of a neural network for NAFLD diagnosis, Sci. Rep. 11 (1) (2021) 1–13.

[20] Q. Saihood, E. Sonuç, A practical framework for early detection of diabetes using ensemble machine learning models, Turk. J. Electr. Eng. Comput. Sci. 31 (4) (2023) 722–738.

[21] H.A. Al-Jamimi, Synergistic feature engineering and ensemble learning for early chronic disease prediction, IEEE Access (2024).

[22] R. Curci, A. Bianco, I. Franco, A. Campanella, A. Mirizzi, C. Bonfiglio, P. Sorino, F. Fucilli, G. Di Giovanni, N. Giampaolo, et al., The effect of low glycemic index mediterranean diet and combined exercise program on metabolic-associated fatty liver disease: A joint modeling approach, J. Clin. Med. 11 (15) (2022) 4339.

[23] A. De, N. Bhagat, M. Mehta, S. Taneja, A. Duseja, Metabolic dysfunction-associated steatotic liver disease (MASLD) definition is better than MAFLD criteria for lean patients with NAFLD, J. Hepatol. 80 (2) (2024) e61–e62.

[24] D.-Q. Sun, Y. Jin, T.-Y. Wang, K.I. Zheng, R.S. Rios, H.-Y. Zhang, G. Targher, C.D. Byrne, W.-J. Yuan, M.-H. Zheng, MAFLD and risk of CKD, Metabolism 115 (2021) 154433.

[25] S. Lin, J. Huang, M. Wang, R. Kumar, Y. Liu, S. Liu, Y. Wu, X. Wang, Y. Zhu, Comparison of MAFLD and NAFLD diagnostic criteria in real world, Liver Int. 40 (9) (2020) 2082–2089.

[26] M. Decraecker, D. Dutartre, J.-B. Hiriart, M. Irles-Depé, F. Chermak, J. Foucher, V. de Lédinghen, Long-term prognosis of patients with metabolic (dysfunction)-associated fatty liver disease by non-invasive methods, Aliment. Pharmacol. Ther. 55 (5) (2022) 580–592.

[27] V.H. Nguyen, M.H. Le, R.C. Cheung, M.H. Nguyen, Differential clinical characteristics and mortality outcomes in persons with NAFLD and/or MAFLD, Clin. Gastroenterol. Hepatol. 19 (10) (2021) 2172–2181.

[28] D. Kim, P. Konyn, K.K. Sandhu, B.B. Dennis, A.C. Cheung, A. Ahmed, Metabolic dysfunction-associated fatty liver disease is associated with increased all-cause mortality in the United States, J. Hepatol. 75 (6) (2021) 1284–1291.

[29] K. Drożdż, K. Nabrdalik, H. Kwiendacz, M. Hendel, A. Olejarz, A. Tomasik, W. Bartman, J. Nalepa, J. Gumprecht, G.Y. Lip, Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: a machine learning approach, Cardiovasc. Diabetol. 21 (1) (2022) 240.

[30] Q. Huang, X. Zou, X. Wen, X. Zhou, L. Ji, NAFLD or MAFLD: which has closer association with all-cause and cause-specific mortality?—results from NHANES III, Front. Med. 8 (2021).

[31] A. Mirizzi, L.R. Aballay, G. Misciagna, M.G. Caruso, C. Bonfiglio, P. Sorino, A. Bianco, A. Campanella, I. Franco, R. Curci, et al., Modified WCRF/AICR score and all-cause, digestive system, cardiovascular, cancer and other-cause-related mortality: A competing risk analysis of two cohort studies conducted in southern Italy, Nutrients 13 (11) (2021) 4002.

[32] P. Sever, New hypertension guidelines from the national institute for health and clinical excellence and the british hypertension society, J. Renin-Angiotensin-Aldosterone Syst. 7 (2) (2006) 61–63.

[33] World Health Organization, et al., WHO Guidelines on Drawing Blood: Best Practices in Phlebotomy, World Health Organization, 2010.

[34] C. Bonfiglio, A. Campanella, R. Donghia, A. Bianco, I. Franco, R. Curci, C.B. Bagnato, R. Tatoli, G. Giannelli, F. Cuccaro, Development and internal validation of a model for predicting overall survival in subjects with MAFLD: A cohort study, J. Clin. Med. 13 (4) (2024) 1181.

[35] C.-C. Wu, W.-C. Yeh, W.-D. Hsu, M.M. Islam, P.A.A. Nguyen, T.N. Poly, Y.-C. Wang, H.-C. Yang, Y.-C.J. Li, Prediction of fatty liver disease using machine learning algorithms, Comput. Methods Programs Biomed. 170 (2019) 23–29.

[36] G. Feng, K.I. Zheng, Y.-Y. Li, R.S. Rios, P.-W. Zhu, X.-Y. Pan, G. Li, H.-L. Ma, L.-J. Tang, C.D. Byrne, et al., Machine learning algorithm outperforms fibrosis markers in predicting significant fibrosis in biopsy-confirmed NAFLD, J. Hepato-Biliary-Pancreatic Sci. 28 (7) (2021) 593–603.

[37] S.K. Dey, A. Hossain, M.M. Rahman, Implementation of a web application to predict diabetes disease: an approach using machine learning algorithm, in: 2018 21st International Conference of Computer and Information Technology, ICCIT, IEEE, 2018, pp. 1–5.

[38] W.G. Baxt, Application of artificial neural networks to clinical medicine, The Lancet 346 (8983) (1995) 1135–1138.

[39] M. Pal, Random forest classifier for remote sensing classification, Int. J. Remote Sens. 26 (1) (2005) 217–222.

[40] W.S. Noble, What is a support vector machine? Nature Biotechnol. 24 (12) (2006) 1565–1567.

[41] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, et al., XGBoost: extreme gradient boosting, 1, (4) 2015, pp. 1–4, R package version 0.4-2.

[42] J. Fan, X. Ma, L. Wu, F. Zhang, X. Yu, W. Zeng, Light gradient boosting machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data, Agricult. Water. Manag. 225 (2019) 105758.

[43] A. Alanazi, Using machine learning for healthcare challenges and opportunities, Inform. Med. Unlocked 30 (2022) 100924.

[44] S.M.D.A.C. Jayatilake, G.U. Ganegoda, Involvement of machine learning tools in healthcare decision making, J. Healthc. Eng. 2021 (1) (2021) 6679512.

[45] M.A. Sarwar, N. Kamal, W. Hamid, M.A. Shah, Prediction of diabetes using machine learning algorithms in healthcare, in: 2018 24th International Conference on Automation and Computing, ICAC, IEEE, 2018, pp. 1–6.

[46] M. Ferdous, J. Debnath, N.R. Chakraborty, Machine learning algorithms in healthcare: A literature survey, in: 2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT, IEEE, 2020, pp. 1–6.

[47] P.K. Kushwaha, M. Kumaresan, Machine learning algorithm in healthcare system: A review, in: 2021 International Conference on Technological Advancements and Innovations, ICTAI, IEEE, 2021, pp. 478–481.
[48] F.K. Došilović, M. Brčić, N. Hlupić, Explainable artificial intelligence: A survey, in: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO, IEEE, 2018, pp. 0210–0215.
[49] M.T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
[50] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).
[51] A. Lombardi, D. Diacono, N. Amoroso, P. Biecek, A. Monaco, L. Bellantuono, E. Pantaleo, G. Logroscino, R. De Blasi, S. Tangaro, et al., A robust framework to investigate the reliability and stability of explainable artificial intelligence markers of mild cognitive impairment and Alzheimer's disease, Brain Inform. 9 (1) (2022) 1–17.
[52] T. Calders, S. Jaroszewicz, Efficient AUC optimization for classification, in: European Conference on Principles of Data Mining and Knowledge Discovery, Springer, 2007, pp. 42–53.
[53] B. Efron, R. Tibshirani, The bootstrap method for assessing statistical accuracy, Behaviormetrika 12 (17) (1985) 1–35.
[54] J. Duckett, HTML & CSS: Design and Build Websites, vol. 15, Wiley Indianapolis, IN, USA, 2011.