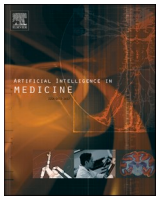
An innovative artificial intelligence-based method to compress complex models into explainable, model-agnostic and reduced decision support systems with application to

Artificial Intelligence In Medicine 151 (2024) 102841

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed



Research paper

An innovative artificial intelligence-based method to compress complex models into explainable, model-agnostic and reduced decision support systems with application to healthcare (NEAR)

Karim Kassem [a,1], Michela Sperti [a,1], Andrea Cavallo [b], Andrea Mario Vergani [c,d,e], Davide Fassino [f], Monica Moz [g], Alessandro Liscio [g], Riccardo Banali [g], Michael Dahlweid [g], Luciano Benetti [g], Francesco Bruno [h,i], Guglielmo Gallone [h,i], Ovidio De Filippo [h,i], Mario Iannaccone [j], Fabrizio D'Ascenzo [h,i], Gaetano Maria De Ferrari [h,i], Umberto Morbiducci [a], Emanuele Della Valle [c], Marco Agostino Deriu [a,*]

[a] *Polito^{BIO}Med Lab, Department of Mechanical and Aerospace Engineering, Politecnico di Torino, Turin, Italy*
[b] *SmartData@PoliTO Center for Big Data Technologies, Politecnico di Torino, Turin, Italy*
[c] *Department of Electronics, Information and Bioengineering (DEIB), Politecnico di Milano, Via Ponzio 34/5, 20133 Milan, Italy*
[d] *Department of Mathematics, Politecnico di Milano, Via Bonardi 9, 20133 Milan, Italy*
[e] *Health Data Science Centre, Human Technopole, Viale Rita Levi-Montalcini 1, 20157 Milan, Italy*
[f] *Department of Mathematical Sciences, Politecnico di Torino, Turin, Italy*
[g] *Dedalus Research Lab, Milan, Italy*
[h] *Division of Cardiology, Cardiovascular and Thoracic Department, Città della Salute e della Scienza, Turin, Italy*
[i] *Cardiology, Department of Medical Sciences, University of Turin, Turin, Italy*
[j] *Department of Cardiology, S. G. Bosco Hospital, Turin, Italy*

A R T I C L E I N F O

Keywords:
Clinical decision support systems
Explainability
Artificial intelligence
Precision medicine
Risk prediction

A B S T R A C T

Background and objective: In everyday clinical practice, medical decision is currently based on clinical guidelines which are often static and rigid, and do not account for population variability, while individualized, patient-oriented decision and/or treatment are the paradigm change necessary to enter into the era of precision medicine. Most of the limitations of a guideline-based system could be overcome through the adoption of Clinical Decision Support Systems (CDSSs) based on Artificial Intelligence (AI) algorithms. However, the black-box nature of AI algorithms has hampered a large adoption of AI-based CDSSs in clinical practice. In this study, an innovative AI-based method to compress AI-based prediction models into explainable, model-agnostic, and reduced decision support systems (NEAR) with application to healthcare is presented and validated.

Methods: NEAR is based on the Shapley Additive Explanations framework and can be applied to complex input models to obtain the contributions of each input feature to the output. Technically, the simplified NEAR models approximate contributions from input features using a custom library and merge them to determine the final output. Finally, NEAR estimates the confidence error associated with the single input feature contributing to the final score, making the result more interpretable. Here, NEAR is evaluated on a clinical real-world use case, the mortality prediction in patients who experienced Acute Coronary Syndrome (ACS), applying three different Machine Learning/Deep Learning models as implementation examples.

Results: NEAR, when applied to the ACS use case, exhibits performances like the ones of the AI-based model from which it is derived, as in the case of the Adaptive Boosting classifier, whose Area Under the Curve is not statistically different from the NEAR one, even the model's simplification. Moreover, NEAR comes with intrinsic explainability and modularity, as it can be tested on the developed web application platform (https://neardashboard.pythonanywhere.com/).

Conclusions: An explainable and reliable CDSS tailored to single-patient analysis has been developed. The proposed AI-based system has the potential to be used alongside the clinical guidelines currently employed in the

* Corresponding author.
E-mail address: marco.deriu@polito.it (M.A. Deriu).
[1] Karim Kassem and Michela Sperti contributed equally to this study.

https://doi.org/10.1016/j.artmed.2024.102841
Received 17 October 2023; Received in revised form 29 February 2024; Accepted 11 March 2024
Available online 12 March 2024

medical setting making them more personalized and dynamic and assisting doctors in taking their everyday clinical decisions.

1. Introduction

Artificial Intelligence (AI) and its subfields Machine Learning (ML) and Deep Learning (DL) are rapidly transforming many aspects of healthcare [1]. Critical clinical challenges, like automated diagnosis, prognosis, drug discovery, and treatment effects, have been greatly improved using ML and DL algorithms [2–7]. Among the most significant aims of AI in healthcare, the personalization of clinical evaluations and treatments for specific patients plays a key role [8]. Medical decisions and best clinical practices are based on clinical guidelines, which suffer from being often static, rigid, and derived from average populations. In general, guidelines result from randomized controlled trials. However, guidelines might not always fit real-world patients. In this context, data-driven methodologies may greatly improve clinical guidelines, making them more personalized, more flexible, and more adaptable to specific situations, thus, addressing the modern vision of precision medicine.

The above-mentioned trend is embodied by the rapid increase in the medical domain of Clinical Decision Support Systems (CDSSs) [9] augmented by AI algorithms [10]. A traditional CDSS is a software tool designed to support the physician during the clinical decision-making process, in which the characteristics of an individual patient are matched to a computerized clinical knowledge base, to produce patient-specific evaluations and/or recommendations to be presented to the clinician for a decision [11].

In recent years several successful examples of AI-based systems employed in clinical decision-making have been proposed, driven by the recent progress in ML/DL which has moved towards deeper and more complex architectures [12] able to tackle typical problems in biomedicine. However, complexity comes at cost. Deep models are often described as black boxes generating predictions without providing the reasons behind their outcomes [8]. This lack of understanding leads to severe consequences, especially in medical applications, where clinical decisions ultimately affect human health [13]. Other similar concerns around the use of AI-based CDSSs in healthcare include lack of transparency [1], transferability, informativeness, fairness, privacy [14], and confidence [15]. The mentioned challenges are amplified in the precision medicine setting, where experts require much more information from a CDSS than a simple binary prediction to perform their diagnosis [15]. For these reasons, eXplainability in AI (XAI) and the related concepts of interpretability and transparency have become central priorities in AI's fair and ethical application in medicine [16].

XAI in CDSSs is a relatively new area of study, but there are already several examples of explainable tools applied to clinical frameworks [17–19]. Nevertheless, most of the systems provide explanations for individual patient decisions without making the underlying ML/DL model more understandable, more transparent, and ultimately more usable, and do not effectively address the well-known trade-off between the accuracy and transparency of complex models.

In this broad scenario, there are still unresolved issues concerning clinical guidelines and unmet requirements regarding ML/DL models, such as being ethical, trustworthy, transparent, and fair. More in detail, a human-centric vision is crucial when developing CDSSs based on ML/DL models. This means that the end-user (e.g., medical doctors) should be able to interact with those tools in an easy, effective, reliable, and friendly way, without the need for technical knowledge. Moreover, AI-driven tools in healthcare should also empower medical doctors with knowledge base extracted from real-world data and population analysis to support prognosis and diagnosis of real-world patients yet without disclosing data or personal information employed for generating the knowledge base.

To tackle the above-mentioned open issues and to meet the listed ML/DL-based product requirements, we propose NEAR, an AI-based method to compress complex models into an explainable, model-agnostic, and reduced decision support system with potential applications in healthcare. NEAR is a modular environment that transfers the prediction capabilities of a standard ML/DL classifier to a simpler explainable model through the Shapley Additive Explanations (SHAP) framework [20]. This approach creates an effective tool for clinical risk prediction, with several advantages compared to the initial ML/DL model. First, each prediction comes with an error score that describes its reliability. Then, each prediction is supported by its feature importance, i.e., the impact of each input value on the score computed on the single patient. In conclusion, the model provides results also if the input feature set is not complete and indicates the relevance of the missing values.

2. Methods

NEAR design, development, and evaluation are described in this section. A schematic workflow of NEAR design is presented in Fig. 1.

2.1. Computational framework design

NEAR was developed using Python version 3.8.10, with an object-oriented approach aimed at obtaining a flexible and modular tool. NEAR was designed to be compatible with both scikit-learn [21] binary classifiers and Feed-Forward Neural Networks (FFNNs) based on Keras [22] or PyTorch [23] libraries. Compatibility was attained by developing a custom "wrapper" class (see Section 3.1). In detail, here the SHAP library [20] was used to explain models, NumPy [24] and SciPy [25] to fit curves, pandas [26] to handle data, and plotly [27] to display graphics.

2.2. SHAP application

The SHAP tool was employed to obtain the so-called Partial

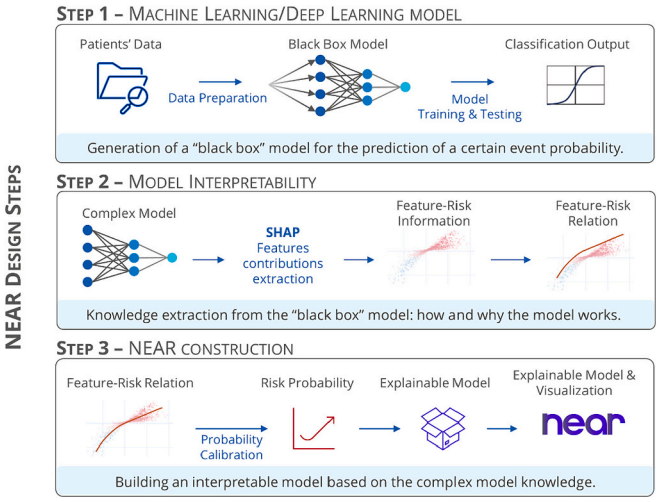


Fig. 1. NEAR design can be summarized into three main steps: 1) Development of an Artificial Intelligence (AI)-based model, to generate a risk score prediction. 2) Analysis of the relation among the features and the output risk score to interpret the above-mentioned black box model. 3) Generation of NEAR, a reduced model built by combining the different features contributions (expressed in terms of simple equations) to produce the output risk score.

Dependence Plots (PDPs), i.e., representations of the impact of the single input features on the output score of the model (Fig. 1, step 2). First, the SHAP explainer object [28] and the expected output of the ML/DL model (the so-called baseline) were derived using 100 randomly selected samples from the training set (Section 2.5 explains how the training set was defined). Then, the SHAP explainer was used to obtain the marginal contribution, commonly known as SHAP value, of each input feature to the predicted score. This computation was performed for every sample of a subset of the validation set made of 200 entries (Section 2.5 explains how the validation set was defined). The PDPs were then derived by coupling marginal contributions and actual values of each considered feature, as obtained for the various samples. Additionally, each feature was assigned a global importance score, defined as the average of the absolute values of the individual contributions over the 200 validation samples. The global feature importance was expressed in percentage terms, in this study.

2.3. Partial dependence plots fitting

PDPs were fitted by curves to estimate the mathematical relationship between the patient feature and its impact on the final score. Each patient feature was first scaled to the $[-1, 1]$ range. Then, the best fitting curve was identified among the ones included in the NEAR library (Table S1) which contains equations of several possible fitting curves (e.g., linear, polynomial, sigmoidal, exponential). For each patient feature, together with the fitting curve, uncertainty margins were also quantified. More details on the best curve identification and quantification of uncertainty are given in Supplementary material, Sections 1 and 2.

2.4. Risk score generation and calibration

At the end of the curve selection step the original ML/DL model is reduced to a simpler and more easily interpretable model, NEAR, which uses the optimized curves for each feature and their respective margins of error. Specifically, given a value x for a feature, the output $f(x;\overline{p})$ of the optimized curve is chosen to define the contribution (positive or negative) of that feature to the total classification score for the sample the feature belongs to. Furthermore, using the margins of error of the curve, it is possible to quantify the budget of uncertainty related to the input value x of the feature. In the presence of a missing value for the feature, the related uncertainty contribution to the final score can be estimated using its positive and negative average contributions to the PDP. The global score for a sample is obtained by summing up single contributions from each feature and the baseline extracted by SHAP. The margins of uncertainty for the final score are quantified summing up the budget of uncertainty evaluated for each feature (including the missing ones).

The operation of the ML/DL scores conversion into probability values resembling the real occurrences of classes, is defined as score calibration. Applying score calibration, NEAR provides the clinician with a risk probability, because providing only the classification score generated by the prediction model would not be informative enough. Technically, before the calibration step probability scores are first transformed into their log-odds, as proposed in [29]. Then score calibration is performed adopting a regression model called calibrator, which is trained to fit the probability of a class given the score as an input feature. This step is performed on the validation set to prevent overfitting. Here the Platt's logistic model [30] is adopted as the calibrator to transform the score using the formulation:

$$p(y_i = 1|f_i) = \frac{1}{1 + exp(Af_i + B)},$$

where y_i is the true label of sample i, f_i is the uncalibrated score provided by the classifier for the same sample, and A and B are real numbers that are identified using maximum likelihood estimation. The Platt's logistic model creates a mapping between the scores and their probabilistic values, and it is applied to the outputs of the NEAR predictor to obtain the final risk probability. Moreover, to calibrate the contribution of the single features, the difference between the calibrated final score and the model baseline can be calculated and distributed to each feature proportionally to its contribution.

2.5. Use case definition and evaluation

NEAR is defined as a general framework that can be applied to binary classification tasks on tabular datasets using any ML/DL model.

In this work, NEAR is evaluated on a cardiological use case to be intended as a proof of concept: the mortality prediction in patients who experienced Acute Coronary Syndrome (ACS) [7]. The aim of this evaluation is to highlight NEAR modularity, flexibility, and performance in a real clinical use case, presenting its intrinsic explainability, transparency, and utility for clinicians.

To prove the flexibility of NEAR to adapt to different models, the application of three popular ML/DL classifiers is presented: Adaptive Boosting (AdaBoost) [31] implemented with scikit-learn library; FFNN [32], implemented with PyTorch; customized Naïve Bayes (NB) [33] classifier. The dataset used to build up the PRAISE score [7] has been used to train and evaluate the ML/DL algorithms. Since the goal of the present work is not the development of a novel risk prediction score, but rather of a supporting decision tool able to augment any risk prediction model, the dataset derivation and external validation cohorts used in [7] were merged and all missing values imputed by their median value. The definition of the variables used in the present study is detailed in the Supplementary material (Section 3).

The aforementioned models are trained, evaluated, and compared to the corresponding simplified models explained using NEAR. Models' performances are evaluated in terms of accuracy score, F2 score, and Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC). The best ML/DL model is also considered for a more detailed comparison with its NEAR-based simplification, performed in terms of ROC curve and Confusion Matrices (CMs). Statistically reliable performances are obtained splitting the dataset into training and test sets (proportion: 90/10 %), followed by a 5-fold cross-validation [32] applied during the training/validation phase of the model building. Finally, bootstrap [34] is applied over the test phase of the models.

3. Results

In this section, the presentation of the main technical characteristics of NEAR is followed by a presentation of the results of its application to the selected clinical real-world use case (mortality prediction in patients who experienced an ACS [7]).

3.1. NEAR architecture

The schematic architecture of NEAR is presented in Fig. 2 with implementation details.

NEAR is composed of four main implementation blocks:

a) SHAP-LOCK: this module checks (i) NEAR's compatibility with different ML/DL models (scikit-learn, Keras/PyTorch, as well as customized models are accepted as inputs), (ii) the integrity of the provided input dataset, (iii) the presence of a scaler object, and applies SHAP to the ML/DL model to generate the explainer object (as explained in Section 2.1).
b) SHAP-EX: in this module, ML/DL model baseline and PDPs for each model feature are calculated by applying SHAP. The global ranking of features is provided as output (as explained in Section 2.2).
c) Automatic VIrtual Sensor Calibration (AVISC): in this module, each PDP is automatically best fitted using curves in the NEAR library

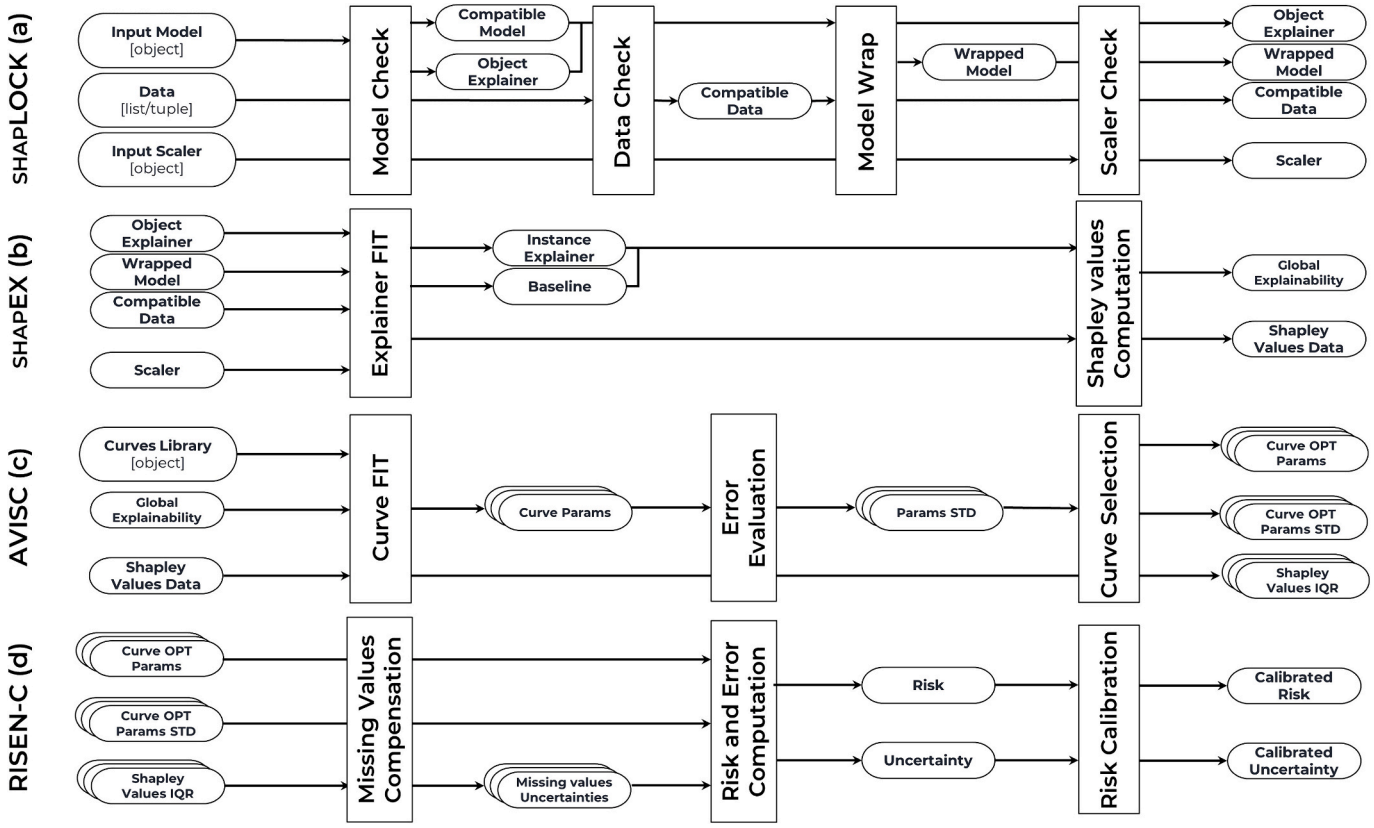


Fig. 2. NEAR architecture is depicted as a sequence of operations, grouped into four main blocks: a) SHAP-LOCK, b) SHAP-EX, c) Automatic VIrtual Sensor Calibration (AVISC), and d) RIsk, miSsing valuEs maNagement, and Calibration (RISEN-C). In the workflow, square boxes represent actions, while circle boxes represent objects. If an arrow is represented behind a box, it means that it is skipping that action. STD: standard deviation, OPT: optimal, IQR: interquartile range. Note that the workflow shows NEAR operations performed on one single dataset split. The workflow is extendable to a cross-validation setting, in which the process is repeated multiple times, as in the NEAR development phase.

(Table S1) and the error score over the single fit is provided (as explained in Section 2.3).

d) RIsk, miSsing valuEs maNagement, and Calibration (RISEN-C): in this module, the final risk score probability is calculated together with its error score (considering the presence or absence of missing values). Then, the risk score and the associated error are calibrated (as explained in Section 2.4).

The NEAR object, that encapsulates the above-mentioned four blocks, can process any dataset, deal with any ML/DL model, and implement different splitting techniques. Once fitted, NEAR provides a detailed analysis of each single sample in terms of contribution of the single features, errors, calibrated and not calibrated output risk score.

3.2. NEAR evaluation

A real-world clinical use case was considered for NEAR evaluation: the mortality risk prediction in patients who experienced an ACS. As described in the Methods Section 2.5 NEAR prediction performances were compared against the performances of ML/DL models upon which NEAR was built: AdaBoost, FFNN, and customized NB classifier. Results reported in Table 1 show accuracy, F2, and AUC scores with associated

Table 1
NEAR performances in terms of accuracy, F2, and Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) scores are reported and compared with the ones of the scikit-learn-based classifier (namely, Adaptive Boosting, AdaBoost), the ones of the PyTorch-based classifier (namely, Feed-Forward Neural Networks, FFNN), and the ones of the customized model (namely, Naïve Bayes, NB) used to compute the NEAR reduced model. The results are shown for multiple dataset splits into validation (val) and test sets. The statistical variance given by the cross-validation and bootstrap applied during the training phase and NEAR generation is also shown as Interquartile Range (IQR) over the performance results.

Data split	Accuracy	F2 score	AUC	Accuracy	F2 score	AUC	Accuracy	F2 score	AUC
		AdaBoost			*FFNN*			*NB*	
Val	0.80 (0.00)	0.36 (0.02)	0.80 (0.03)	0.84 (0.01)	0.38 (0.05)	0.81 (0.00)	0.82 (0.01)	0.35 (0.03)	0.79 (0.01)
Test	0.81 (0.00)	0.40 (0.02)	0.84 (0.00)	0.85 (0.02)	0.41 (0.00)	0.84 (0.01)	0.82 (0.01)	0.38 (0.00)	0.82 (0.00)

Data split	Accuracy	F2 score	AUC	Accuracy	F2 score	AUC	Accuracy	F2 score	AUC
		NEAR-AdaBoost			*NEAR-FFNN*			*NEAR-NB*	
Val	0.79 (0.01)	0.35 (0.02)	0.81 (0.01)	0.88 (0.01)	0.37 (0.02)	0.82 (0.01)	0.86 (0.01)	0.37 (0.04)	0.80 (0.02)
Test	0.81 (0.01)	0.37 (0.07)	0.83 (0.04)	0.88 (0.00)	0.43 (0.09)	0.84 (0.04)	0.86 (0.01)	0.39 (0.06)	0.82 (0.04)

Interquartile Ranges (IQRs) over multiple dataset splits. Table 1 elucidates the comparative performance metrics of the models employed in this study. Reading the upper part of the Table one can observe, e.g., how the AUC of the three different ML/DL tested models is comparable, both on the validation and on the test set (AdaBoost reached a value of 0.80 and 0.84 on the validation and test set, respectively, FFNN of 0.81 and 0.84, and NB of 0.79 and 0.82). Reading the lower part of the Table, and comparing it with the upper part, one can observe, e.g., that NEAR performances are comparable with the ones of the complex models upon which it was built. As an example, when compared to AdaBoost, which has an AUC of 0.80 and 0.84, respectively on the validation and test set, NEAR shows an AUC of 0.81 and 0.83.

Since the performances (in terms of F2 score, that we chose as the most informative parameter) of AdaBoost and FFNN are comparable and slightly better than those of NB (as proved in Tables 1 and 2), we selected one of them, namely AdaBoost, as illustrative example to better investigate NEAR performances in terms of ROC curves and CMs (Fig. 3).

Table 2 illustrates that the NEAR-modified models maintain robust performance metrics, evidencing no significant deterioration when benchmarked against the ML/DL models from which they are derived. For instance, an analysis of the AdaBoost and its NEAR-AdaBoost counterpart (specifically in the lower left quadrant of the table) reveals that the majority of *p*-values, derived from statistical tests on accuracy, F2 score, and AUC measures, exceed the threshold of statistical significance.

The ROC curves and the CMs for the other two ML/DL classifiers are reported in the Supplementary material (Figs. S2 and S3). Fig. 3 presents a detailed analysis of NEAR proficiency when built upon the AdaBoost model. In detail, in Fig. 3 (left panel) is reported that NEAR built upon AdaBoost (both on the validation and test set) exhibits the same level of performance (in terms of ROC curve) of the AdaBoost model. In the figure right panel, we can observe the CMs corresponding to each ROC curve on the left panel. If we compare the CM built using AdaBoost on the test set and the one built using NEAR on the test set, we can observe how NEAR maintains the same level of performance in terms of true negatives (81.36 % vs 81.13 %) with a slight degradation over the true positives (74.33 % vs 71.89 %).

3.3. NEAR functionalities

In this section, we present the main functionalities provided by NEAR when compared to the state-of-the-art ML/DL models:

- The possibility of calculating a local features importance (for the single patient) with a relative uncertainty over the importance itself, which is then propagated to the final risk score probability provided by NEAR (Fig. 4 and methodological Section 2.4). All the scores are calibrated.
- The possibility of treating missing values in the prediction phase, providing a relative error over the presence/absence of specific features and their effect on the risk probability associated error (Fig. 5 and methodological Section 2.4).
- The possibility of inspecting the explainable model (NEAR) in terms of its set of building equations and of comparing it with the clinical domain knowledge.

An example of NEAR functionalities applied to a single patient (female, 81 years old, deceased) is presented in Fig. 4, where: the personalized final risk score (32.10 %), obtained summing up the contributions of all the single features, is reported with upper and lower uncertainty margins (46.28 % and 17.92 %, respectively; Fig. 4, first row). The overall risk uncertainty for this patient is thus 46.28 %–17.92 % = 28.36 %. Then, the contribution to the final score of single features is reported as a positive or negative score (identifying increasing or decreasing risk of death contribution, respectively; Fig. 4, second row) and the final risk can be derived by adding each feature contribution to the model's baseline (1.96 %). Finally, the relative error of each feature is reported (in Fig. 4, third row, for an age value equal to 81 the relative error is 34 %, which represents about one-third of the overall risk uncertainty).

NEAR provides indications about the impact that the presence of missing values might have on the final risk score increasing the error over the final prediction, also providing in which relative percentage, with respect to other features. In the specific example in Fig. 5 (male, 54 years old, survived), even if the contributions of missing variables Prasu, Vascular_access, and CKD_EPI (the complete features description can be found in the Supplementary material, Section 1) are equal to 0, they still increase the prediction error of 14 %, 9 %, and 14 %, thus making the final score less reliable.

The NEAR pilot version (an interactive visualization interface that enables NEAR tests on real patients' data) can be tested at the following link: https://neardashboard.pythonanywhere.com/.

4. Discussion

A growing number of CDSSs [9] enhanced by AI algorithms [10] have been reported in the literature in recent years, paving the way to more and more personalized medicine strategy. A robust and affordable clinical system supports healthcare providers to make decisions and improve patients care. In this sense, a CDSS that uses knowledge management and enables integrated workflows to obtain clinical advice on a specific patient based on many parameters provides support at the time of care and personalized care plan suggestions [11]. Increasingly

Table 2
To better investigate the statistical differences among the three Machine Learning/Deep Learning (ML/DL) models tested in the present study and among each NEAR model and the ML/DL model from which it was derived, a Mann Whitney *U* test was performed [35]. More in detail, AdaBoost and FFNN, which are the best performing ML/DL models, does not show any statistical difference. Moreover, NEAR maintains the same level of performances with respect to each ML/DL models. Due to the small number of folds, the *p*-value resolution is reduced. As a result, few *p*-values are equal to one, nonetheless verifying the hypothesis of NEAR and ML/DL model performance comparability.

Data split	Accuracy	F2 score	AUC	Accuracy	F2 score	AUC	Accuracy	F2 score	AUC
	AdaBoost-FFNN			*FFNN-NB*			*AdaBoost-NB*		
Val	p = 0.15	p = 0.22	p = 0.55	p = 0.15	p = 0.15	p = 0.10	**p = 0.01**	p = 0.42	p = 0.55
Test	p = 0.15	p = 0.69	p = 0.42	p = 0.15	**p = 0.01**	p = 0.06	**p = 0.02**	**p = 0.01**	p = 0.15

Data split	Accuracy	F2 score	AUC	Accuracy	F2 score	AUC	Accuracy	F2 score	AUC
	NEAR-AdaBoost			*NEAR-FFNN*			*NEAR-NB*		
Val	**p = 0.03**	p = 0.55	p = 1	**p = 0.01**	p = 0.84	p = 1	**p = 0.01**	p = 0.31	p = 0.42
Test	p = 0.84	p = 0.42	p = 1	**p = 0.01**	p = 0.84	p = 1	**p = 0.01**	p = 0.69	p = 0.69

A p-value presented in bold denotes statistical significance, indicating a value below the threshold of 0.05.

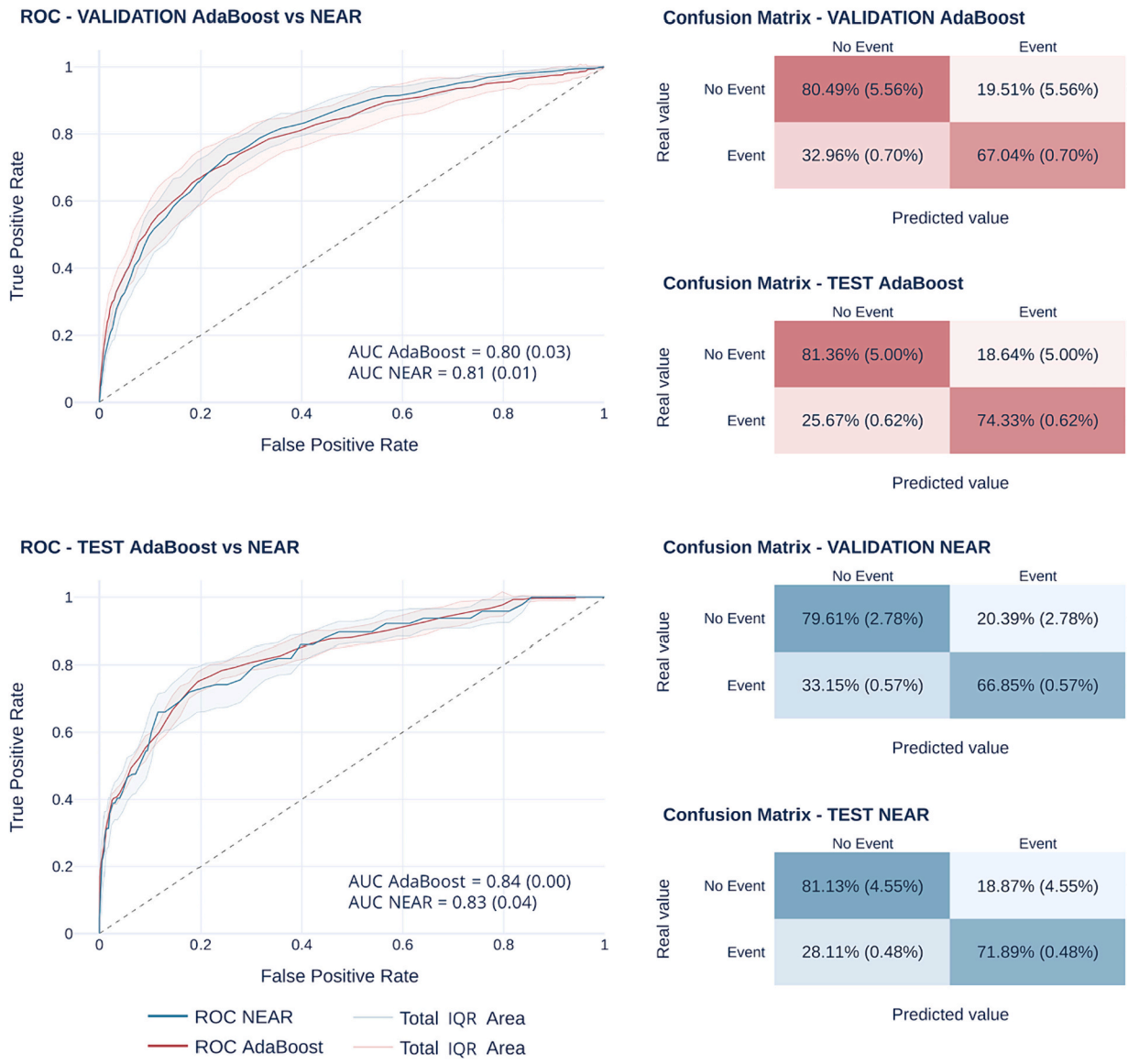


Fig. 3. NEAR performances in terms of the Receiver Operating Characteristic (ROC) curves, the Area Under the ROC Curve (AUC), and the Confusion Matrices (CM) are reported and compared with the ones of the selected ML-based classifier (namely, AdaBoost) used to compute the NEAR reduced model. Both the ROC curves and the CM report the statistical variance given by the cross-validation and the bootstrap applied during the AdaBoost's training phase and NEAR generation. The performances are shown in the case of validation (figure's first row) and test (figure's second row) sets of the clinical use case.

complex models (e.g., Deep Neural Networks) [12], have been implemented as main core in CDSSs since they repeatedly proved to outperform alternative approaches [2–7]. However, such systems present unique challenges in healthcare applications, since they are often in the form of black box, with little or no understanding of how and why the tool suggested a particular action [8].

In this scenario, an AI-based modular framework (NEAR) able to circumvent the opaqueness of ML/DL algorithms by turning complex models into explainable and transparent ones (as shown in Figs. 4 and 5, providing an explanatory example of how clinicians can be informed on the contribution of single variables to the final risk score) without relevant loss of information (as shown in the comparative Table 2, in which the computed *p*-values show no significant difference in performance between each tested ML/DL model and NEAR), is here proposed leveraging the functionalities of the SHAP tool [21]. The methodology implemented in NEAR is analogous to the concept of "knowledge distillation", which was first introduced as model compression in year 2006 to condense complex models (e.g., ensembles) into a single model that is simpler and less computationally expensive [37] and then generalized in 2015 [38]. The idea behind "knowledge distillation" is to train a small model to match the predictions of the large one. NEAR relies on SHAP to (i) decouple the marginal contributions of the features on the large model outputs, and (ii) capture the relationship between SHAP and feature values. In this way, a reduced, computationally more efficient, transparent, and reliable model can be built from the complex one, enabling relevant properties and making more clear cause-effect relationships. In this regard, one NEAR key asset is that it provides not only a binary classification but also a risk score that represents the likelihood of a clinical condition to occur. The score is obtained through calibration, and it is easily interpretable as the sum of the contributions of the single features. As a result, NEAR fulfils the fundamental requirements of explainability (in terms of, e.g., medical interpretability, model limitations, and description of the medical decision-making process) [16]. Moreover, NEAR allows for the identification of the most important features that constitute the final risk, indicating to the clinicians which patient parameters might be most critical. The presence of missing variables with their relevance in the composition of the risk might also be a crucial suggestion for the clinicians in understanding

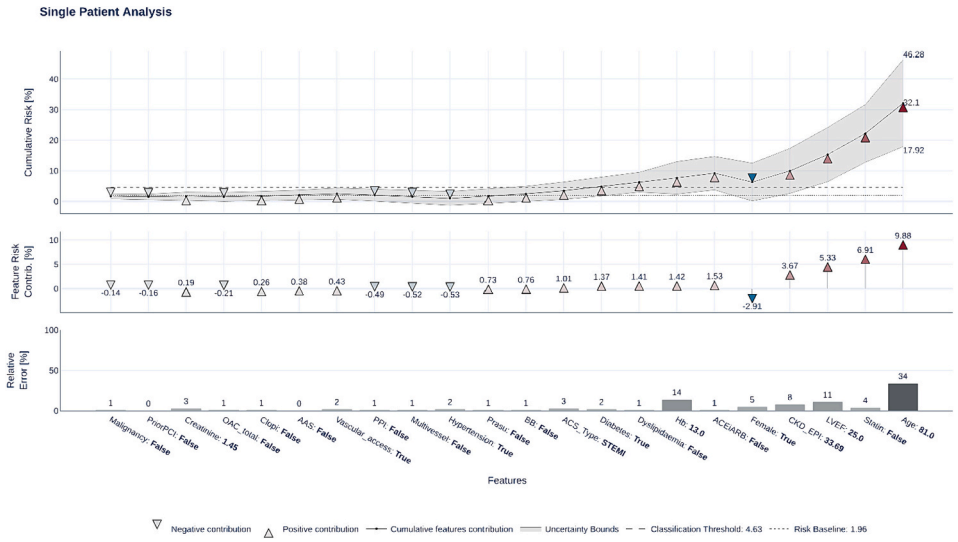


Fig. 4. NEAR explanations on the single patient's features contributions: in the first row, an improved version of the so-called SHAP waterfall plot [36] is shown. The plot is presented in its cumulative version, to highlight how each single feature contribution and its relative error are summed up to the other ones to contribute to the final risk score. A positive contribution (increasing the risk of death) is depicted as an arrow pointing upward, while a negative contribution (decreasing the risk of death) is depicted as an arrow pointing downward. The results are reported for a single patient (using the Adaptive Boosting model to build the NEAR reduced model), considered here as an example. In the second row, the local feature importance, in its absolute value, for the single patient is shown. In the third row, the relative error for every single feature is shown. For visual clarity, only the 22 most important features are reported in the figure.

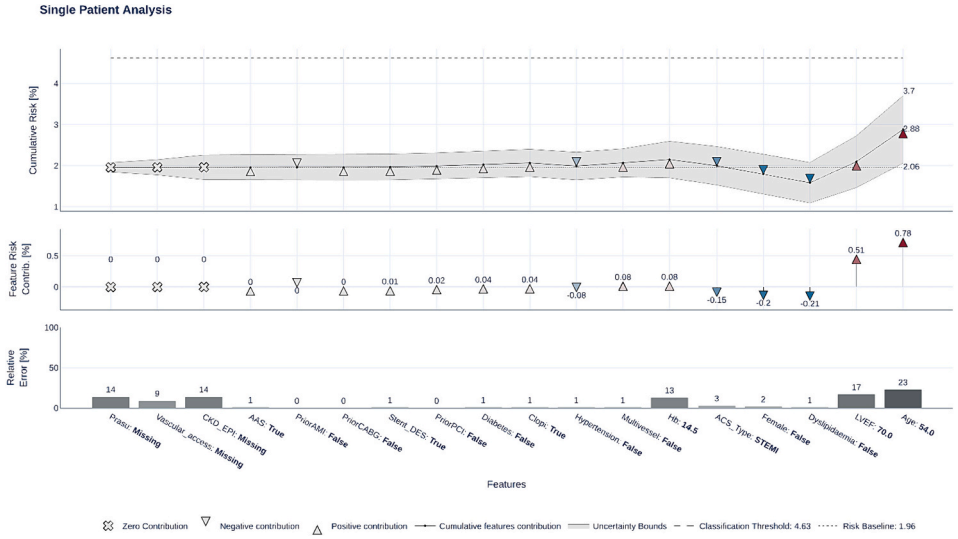


Fig. 5. NEAR capability of handling potential missing features. The above plot is similar to Fig. 5, with the addition of missing values (represented by a cross symbol). Thus, the missing value error is propagated to the final risk score. For visual clarity, only the 18 most important features are reported in the figure.

which data is needed to be collected first, and NEAR provides a measure of the impact of missing values on the final risk score. Finally, NEAR's versatility, here proved by its compatibility with different implementation frameworks (scikit-learn, Keras/PyTorch, and customized models) makes the framework suitable for a large class of applications.

In the context of the existing literature, NEAR has the potential to expand the body of knowledge for practitioners and researchers, as indicated by comparisons with established risk prediction scores [7,39–42]. In Table 3, NEAR is described and compared with five literature examples of popular risk predictors in the context of cardiology. In detail, in the mentioned Table, the model type and the predicted event are reported for each prediction score. The scores are compared based on their main characteristics, such as confidence, interpretability, usability, and explainability. Explainability is discussed at population (global) as well as at individual patient (local) level.

Referring to Table 3, it is observed that NEAR effectively addresses certain limitations hampering the adoption of more complex models such as ensemble and DL models which, despite offering high performance, compromise on interpretability and often lack confidence at the individual (patient-level) and exhibit low explainability [43]. In contrast NEAR, while still achieving high performance, aligns more closely with rule-based and graphical models by offering an inherently transparent structure that facilitates confidence on a local level. The comparison with the five selected literature examples of risk predictors in the context of cardiology, detailed in Table 3, showcases NEAR as a tool marked by its usability (enabled through a dedicated web platform), flexibility (due to its capability to be trained on any ML/DL model), interpretability (with a minimal loss in AUC, under 2 %), and the ability to provide local confidence and explainability for individual predictions. These features collectively support practitioners by enhancing their

Table 3
Comparison between NEAR and five Clinical Decision Support Systems (CDSSs) adopted in cardiology, each one applying a different Machine Learning/Deep Learning (ML/DL) approach, except for the Framingham Risk Score, associated with a rule-based system and broadly acknowledged as a benchmark in the domain of cardiovascular risk prediction. The five CDSSs have been trained and validated on specific datasets as described in related references (indicated in column Model Type). The approaches used include Framingham Risk Score (FRS), Support Vector Machine (SVM), Neural Networks (NNs), Adaptive Boosting (AdaBoost), and Bayesian Networks (BNs). For each approach, the predicted event and the main characteristics, such as the confidence (at the level of the overall model, global, or at the single prediction level, local), the explainability (at the level of the overall model, global, or at the single prediction level, local), the interpretability, and the usability are reported.

Model type	Predicted event	Main characteristics
Framingham Risk Score (Rule-based Model) [39]	Development of Coronary Heart Disease (CHD)	The FRS is a widely known cardiovascular risk predictor for the general population, with several versions published over the years. Among them, a particularly user-friendly version employs a rule-based system founded on a scoring methodology. The inherent interpretability of this score, stemming from its straightforward structure, facilitates both global and local explanations. Additionally, the availability of multiple web-based platforms further enhances its accessibility for calculation.
Support Vector Machine (Machine Learning Model) [40]	Prediction of medication adherence in heart failure patients	Son et al. created a predictive model for medication adherence in heart failure patients using an SVM. The robustness of the proposed methodology lies in the model's predictive efficacy, coupled with its potential utility as a CDSS for patient stratification. This approach not only offers global explainability but also global confidence in the predictions.
Neural Networks (Deep Learning Model) [41]	First cardiovascular event (over 10 years)	The contribution by Weng et al. is in the construction of a cardiovascular risk predictor based on NNs, surpassing the efficacy of alternative ML models. This score not only enhanced the performance of existing cardiovascular risk predictors but also underscored the substantial performance capabilities inherent in DL methodologies. It affords a comprehensive confidence level and explainability on a global scale.
AdaBoost (Machine Learning Ensemble Model) (PRAISE) [7]	Mortality prediction in patients with ACS	The PRAISE score proposed by D'Ascenzo et al., operating as a CDSS, adopts the AdaBoost methodology, demonstrating excellent performance and global confidence. It supplies a comprehensive elucidation of the most important clinical variables pivotal to predictions on a global scale. Furthermore, its accessibility to practitioners is facilitated through a user-friendly web-based platform.
Bayesian Networks (Machine Learning Graphical Model) [42]	Cardiovascular risk prediction based on variables relationships	Ordovas et al. introduced a Bayesian network model designed for cardiovascular event prediction in the general population. The system can be used as a CDSS. The efficacy of this method is underscored by the inherent explainability of the model, both globally and locally. The Bayesian model facilitates the exploration of interrelations among cardiovascular risk factors, enabling in-depth inferences about these factors. Moreover, the work is complemented by freely available software enhancing its utility for practitioners.
NEAR	Mortality prediction in patients with ACS	NEAR acts as a CDSS approximating ML/DL models with a concise set of equations. This system not only sustains a consistent level of performance but also embodies an inherently interpretable form. Its architectural design facilitates model interpretability on both global and local scales, concurrently offering global and local prediction confidence. Additionally, NEAR is complemented by a web-based platform, augmenting its user-friendly interface.

decision-making processes.

Furthermore, NEAR distinguishes itself from many explainable artificial intelligence (XAI) methodologies by creating parametric models that leverage insights from the input complex model. This approach is particularly beneficial in scenarios where computational resources are limited, as it integrates seamlessly without the need for specific libraries, thereby offering potential compatibility across various IT development environments.

The usability, flexibility, interpretability, and explainability of NEAR significantly facilitate practitioners' interactions with the tool, enhancing their decision-making capabilities. These attributes support practitioners to correlate patient characteristics with pathology risks and foresee individual responses to personalized treatments based on their impact on clinical variables. Consequently, future versions of NEAR could serve as instrumental design aids for clinical trials or in the analysis of real-world patient data. Simultaneously, researchers could utilize NEAR in various domains to better understand the mechanisms of disease progression or identify contributing factors that trigger pathological events. It allows for the examination of the connections between features and outcomes across multiple scenarios and on a sample-by-sample basis.

The innovative contribution of the NEAR approach to support the process of clinical decision-making lies in its unique methodology. Instead of focusing on explaining an existing model, NEAR seeks to construct a new parametric model that harnesses the knowledge embedded in the complex model it takes as input. This approach not only facilitates a deeper understanding of the model's insights but also significantly broadens the accessibility of clinical decision support systems. By ensuring that NEAR can be implemented on platforms with

limited computational resources, it democratizes the use of advanced analytical tools in clinical settings, making sophisticated decision support accessible to a wider range of healthcare providers.

In the evolving field of clinical practice, the integration of tools like NEAR offers a forward-thinking approach to augmenting patient care. NEAR aims to address the inherent challenges of clinical guidelines, such as their resource-intensive development, the complexities involved in their updates, and the integration of expert knowledge. By facilitating personalized patient assessments, NEAR enhances the adaptability and application of clinical guidelines to individual patient needs, making these guidelines more dynamic and data-informed.

NEAR's role is to complement existing clinical guidelines by providing an additional layer of decision support, enabling risk evaluations and treatment plans at individual level, thus ensuring that the care delivered is tailored to the unique circumstances of each patient. This method maintains the relevance of clinical guidelines and introduces a new dimension to the evaluation of therapeutic options. The capability of NEAR to adapt to cases where patient data may be incomplete or not fully aligned with guideline requirements is particularly valuable. It uses predictive analytics to fill gaps in patient information, enabling clinicians to apply guidelines more effectively to the patient's condition.

Currently, NEAR operates independent of direct guideline recommendations, offering clinicians a supplementary tool for decision-making. This approach is designed to support, not supplant, the clinician's judgment and the use of clinical evidence in patient care.

Future developments of NEAR are anticipated to foster closer integration with clinical guidelines. The intention is to use guideline frameworks (such as specific recommendations and threshold values) as a contextual filter for NEAR's outputs. This evolution will align NEAR's analytical capabilities more directly with guideline-based recommendations, enhancing the personalization and relevance of patient care.

NEAR seeks to deepen the analysis available to clinicians, helping to tailor interventions more closely to patient needs and ensuring practices align with the highest standards of care. This approach represents a commitment to advancing healthcare delivery through innovation, ensuring it remains patient-centered, precise, and informed by the latest in clinical intelligence.

This study faces possible limitations. First, NEAR is currently only applicable to tabular data and binary classification tasks. Second, NEAR assumes that the dataset's features are uncorrelated. Third, applying NEAR on large amounts of data may become unfeasible due to SHAP's long execution time. Furthermore, it is crucial to stress that NEAR is intended to be a support tool in the healthcare setting, with the last decision always left to the human expert. However, methods also processing patient-specific correlated input features can be easily implemented in the future, making NEAR even more and more usable in the clinical setting.

5. Conclusions

In this work, we presented NEAR, a breakthrough clinical decision support system which challenges main drawbacks related to the day-by-day medical practice supported by clinical guidelines. NEAR leverages AI techniques to provide a personalized, flexible, modular, transparent, explainable, and trustworthy solution with potential applications in healthcare. In this sense, the NEAR framework has the capability of simplifying and explaining black box models, with high versatility in processing a variety of datasets and ML/DL models. The core idea behind NEAR is to fit the PDPs generated by SHAP for each model's variable with a curve. The fitted curve is then used to calculate the contributions of the input variables, and the sum of all contributions yields the final score. This approach has several advantages: first, the final reduced (NEAR) model becomes transparent, explainable, and trustworthy; second, missing variables can be ignored while providing a prediction reliability score; third, the overall error score over the final risk probability can be provided. NEAR fits the case of explainable CDSSs where the risk score provided by the system needs to be unfolded by the patient features and may improve the current clinical guidelines making them more flexible and patient tailored.

CRediT authorship contribution statement

Karim Kassem: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Michela Sperti:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Supervision, Validation, Writing – original draft, Writing – review & editing. **Andrea Cavallo:** Data curation, Formal analysis, Methodology, Software, Validation, Writing – original draft. **Andrea Mario Vergani:** Data curation, Formal analysis, Methodology, Software, Validation, Writing – original draft. **Davide Fassino:** Data curation, Formal analysis, Methodology, Software, Validation, Writing – original draft. **Monica Moz:** Conceptualization, Supervision, Writing – original draft. **Alessandro Liscio:** Methodology, Writing – original draft. **Riccardo Banali:** Methodology, Writing – original draft. **Michael Dahlweid:** Methodology, Writing – original draft. **Luciano Benetti:** Methodology, Writing – original draft. **Francesco Bruno:** Data curation, Validation. **Guglielmo Gallone:** Data curation, Validation. **Ovidio De Filippo:** Data curation, Validation. **Mario Iannaccone:** Data curation, Validation. **Fabrizio D'Ascenzo:** Data curation, Supervision, Validation. **Gaetano Maria De Ferrari:** Data curation, Supervision, Validation. **Umberto Morbiducci:** Supervision, Writing – original draft, Writing – review & editing. **Emanuele Della Valle:** Conceptualization, Supervision. **Marco Agostino Deriu:** Conceptualization, Data curation, Formal analysis, Project administration, Supervision, Writing – original draft, Writing – review & editing.

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

D. F. is a member of the Italian INdAM-GNCS research group. A. M. V. acknowledges the support by MUR, grant Dipartimento di Eccellenza 2023-2027.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.artmed.2024.102841.

References

[1] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med Jan. 2019;25(1):44–56. https://doi.org/10.1038/s41591-018-0300-7.

[2] Hwang EJ, et al. Development and validation of a deep learning–based automated detection algorithm for major thoracic diseases on chest radiographs. JAMA Netw Open Mar. 2019;2(3):e191095. https://doi.org/10.1001/jamanetworkopen.2019.1095.

[3] Chilamkurthy S, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet Dec. 2018;392(10162):2388–96. https://doi.org/10.1016/S0140-6736(18)31645-3.

[4] Burbidge R, Trotter M, Buxton B, Holden S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. Comput Chem Dec. 2001;26(1):5–14. https://doi.org/10.1016/S0097-8485(01)00094-8.

[5] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8–17. https://doi.org/10.1016/j.csbj.2014.11.005.

[6] Halasz G, et al. A machine learning approach for mortality prediction in COVID-19 pneumonia: development and evaluation of the Piacenza score. J Med Internet Res May 2021;23(5):e29058. https://doi.org/10.2196/29058.

[7] D'Ascenzo F, et al. Machine learning-based prediction of adverse events following an acute coronary syndrome (PRAISE): a modelling study of pooled datasets.

Lancet Jan. 2021;397(10270):199–207. https://doi.org/10.1016/S0140-6736(20)32519-8.
[8] Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. WIREs Data Min Knowl Discov Jul. 2019;9(4). https://doi.org/10.1002/widm.1312.
[9] Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med Feb. 2020;3(1):17. https://doi.org/10.1038/s41746-020-0221-y.
[10] Magrabi F, et al. Artificial intelligence in clinical decision support: challenges for evaluating AI and practical implications. Yearb Med Inform Aug. 2019;28(01):128–34. https://doi.org/10.1055/s-0039-1677903.
[11] Sim I, et al. Clinical decision support systems for the practice of evidence-based medicine. J Am Med Inform Assoc Nov. 2001;8(6):527–34. https://doi.org/10.1136/jamia.2001.0080527.
[12] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature May 2015;521(7553):436–44. https://doi.org/10.1038/nature14539.
[13] Antoniadi AM, et al. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. Appl Sci May 2021;11(11):5088. https://doi.org/10.3390/app11115088.
[14] Mireshghallah F, Taram M, Vepakomma P, Singh A, Raskar R, Esmaeilzadeh H. Privacy in deep learning: a survey. Apr. 2020.
[15] Barredo Arrieta A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion Jun. 2020;58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012.
[16] Vellido A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural Comput Applic Dec. 2020;32(24):18069–83. https://doi.org/10.1007/s00521-019-04051-w.
[17] Antoniadi AM, Galvin M, Heverin M, Wei L, Hardiman O, Mooney C. A clinical decision support system for the prediction of quality of life in ALS. J Pers Med Mar. 2022;12(3):435. https://doi.org/10.3390/jpm12030435.
[18] Du Y, Rafferty AR, McAuliffe FM, Wei L, Mooney C. An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus. Sci Rep Jan. 2022;12(1):1170. https://doi.org/10.1038/s41598-022-05112-2.
[19] Antoniadi AM, Galvin M, Heverin M, Hardiman O, Mooney C. Development of an explainable clinical decision support system for the prediction of patient quality of life in amyotrophic lateral sclerosis. In: Proceedings of the 36th annual ACM symposium on applied computing. New York, NY, USA: ACM; Mar. 2021. p. 594–602. https://doi.org/10.1145/3412841.3441940.
[20] Lundberg S, Lee S-I. A unified approach to interpreting model predictions [Online]. Available: http://arxiv.org/abs/1705.07874; May 2017.
[21] Pedregosa F, et al. Scikit-learn: machine learning in Python [Online]. Available: http://arxiv.org/abs/1201.0490; Jan. 2012.
[22] Keras: the Python deep learning API. Accessed: Jan. 27, 2023. [Online]. Available: https://keras.io/.
[23] PyTorch. Accessed: Jan. 27, 2023. [Online]. Available: https://pytorch.org/.
[24] NumPy. Accessed: Jan. 27, 2023. [Online]. Available: https://numpy.org/.
[25] SciPy. Accessed: Jan. 27, 2023. [Online]. Available: https://scipy.org/.
[26] pandas - Python Data Analysis Library. Accessed: Jan. 27, 2023. [Online]. Available: https://pandas.pydata.org/.
[27] Plotly: low-code data app development. Accessed: Jan. 27, 2023. [Online]. Available: https://plotly.com/.
[28] shap.Explainer — SHAP latest documentation. Accessed: Jan. 27, 2023. [Online]. Available: https://shap.readthedocs.io/en/latest/generated/shap.Explainer.html.
[29] Leathart T, Frank E, Holmes G, Pfahringer B. Probability calibration trees. Jul. 2018.
[30] (PDF) probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Accessed: Feb. 09, 2023. [Online]. Available: https://www.researchgate.net/publication/2594015_Probabilistic_Outputs_for_Support_Vector_Machines_and_Comparisons_to_Regularized_Likelihood_Methods.
[31] Schapire RE. Explaining AdaBoost. In: Empirical inference. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013. p. 37–52. https://doi.org/10.1007/978-3-642-41136-6_5.
[32] Michelucci U. Applied deep learning with TensorFlow 2. Berkeley, CA: Apress; 2022. https://doi.org/10.1007/978-1-4842-8020-1.
[33] Webb GI, Keogh E, Miikkulainen R, Miikkulainen R, Sebag M. Naïve Bayes. In: Encyclopedia of machine learning. Boston, MA: Springer US; 2011. p. 713–4. https://doi.org/10.1007/978-0-387-30164-8_576.
[34] Michelucci U, Venturini F. Estimating neural network's performance with bootstrap: a tutorial. Mach Learn Knowl Extr Mar. 2021;3(2):357–73. https://doi.org/10.3390/make3020018.
[35] McKnight PE, Najab J. Mann-Whitney <scp>U</scp> Test. In: The Corsini encyclopedia of psychology. Wiley; 2010. p. 1. https://doi.org/10.1002/9780470479216.corpsy0524.
[36] Waterfall plot — SHAP latest documentation. Accessed: Jan. 30, 2023. [Online]. Available: https://shap.readthedocs.io/en/latest/example_notebooks/api_examples/plots/waterfall.html.
[37] Buciluǎ C, Caruana R, Niculescu-Mizil A. Model compression. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. New York, NY, USA: ACM; Aug. 2006. p. 535–41. https://doi.org/10.1145/1150402.1150464.
[38] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. Mar. 2015.
[39] D'Agostino RB, et al. General cardiovascular risk profile for use in primary care: the Framingham heart study. Circulation Feb. 2008;117(6):743–53. https://doi.org/10.1161/CIRCULATIONAHA.107.699579.
[40] Son Y-J, Kim H-G, Kim E-H, Choi S, Lee S-K. Application of support vector machine for prediction of medication adherence in heart failure patients. Healthc Inform Res 2010;16(4):253. https://doi.org/10.4258/hir.2010.16.4.253.
[41] Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PloS One Apr. 2017;12(4):e0174944. https://doi.org/10.1371/journal.pone.0174944.
[42] Ordovas JM, et al. A Bayesian network model for predicting cardiovascular risk. Comput Methods Programs Biomed Apr. 2023;231:107405. https://doi.org/10.1016/j.cmpb.2023.107405.
[43] Yang G, Ye Q, Xia J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: a mini-review, two showcases and beyond. Inf Fusion Jan. 2022;77:29–52. https://doi.org/10.1016/j.inffus.2021.07.016.