Explainable AI for clinical and remote health applications: a survey on tabular and time series data

Flavio Di Martino[1] · **Franca Delmastro**[1]

Abstract

Nowadays Artificial Intelligence (AI) has become a fundamental component of healthcare applications, both clinical and remote, but the best performing AI systems are often too complex to be self-explaining. Explainable AI (XAI) techniques are defined to unveil the reasoning behind the system's predictions and decisions, and they become even more critical when dealing with sensitive and personal health data. It is worth noting that XAI has not gathered the same attention across different research areas and data types, especially in healthcare. In particular, many clinical and remote health applications are based on tabular and time series data, respectively, and XAI is not commonly analysed on these data types, while computer vision and Natural Language Processing (NLP) are the reference applications. To provide an overview of XAI methods that are most suitable for tabular and time series data in the healthcare domain, this paper provides a review of the literature in the last 5 years, illustrating the type of generated explanations and the efforts provided to evaluate their relevance and quality. Specifically, we identify clinical validation, consistency assessment, objective and standardised quality evaluation, and human-centered quality assessment as key features to ensure effective explanations for the end users. Finally, we highlight the main research challenges in the field as well as the limitations of existing XAI methods.

Keywords Explainable AI · Health · EHR · Time series · Remote patient monitoring · Clinical DSS

1 Introduction

Artificial Intelligence (AI) has become in the last few years a building block of modern health services, improving efficiency and providing a concrete support to the decision making process. However, the lack of transparency and interpretability of AI still remains one

---

✉ Flavio Di Martino
flavio.dimartino@iit.cnr.it

Franca Delmastro
franca.delmastro@iit.cnr.it

[1] Institute for Informatics and Telematics (IIT), National Research Council of Italy (CNR), Via Moruzzi 1, Pisa 56100, Italy

of the main barriers to its real adoption in the clinical practice (Topol 2019), and even more in those systems that require a direct interaction with a non-expert user (e.g., remote patient monitoring and personalised support). In fact, in the healthcare domain we have to distinguish among different types of users dealing with AI systems: (i) the clinical/medical personnel (i.e., the domain expert), who needs explanations to increase their trust in the system and, at the same time, can provide a clinical validation; (ii) the technical developer, who is in charge of the reliability of the model; (iii) the patient or monitored subject, who needs interpretable and personalised explanations. Explainable AI (XAI) techniques have the potentiality to support all these types of users by making AI models more expressive and improving human understanding and confidence in AI-empowered Decision Support Systems (DSSs) (Das and Rad 2020), but they generally do not offer a one-fits-all solution (Cinà et al. 2022).

In addition, the healthcare domain includes a variety of AI applications related to different research areas, each of them requiring appropriate explanations. To date, biomedical imaging (e.g., X-rays, CT-scans, ultrasounds) is one of the most active XAI application fields, trying to explain model classification by generating saliency maps to highlight the relevance of different image regions (aka, super-pixels) to a given prediction (Tjoa and Guan 2020). It is generally applied in Computer-Aided Diagnosis (CAD) with different targets, such as cancer (Gulum et al. 2021) and COVID-19 (Mondal et al. 2021; Faruk 2021). However, many health applications are also based on several other data types, such as tabular and time series data that can derive from clinical information, such as Electronic Health Records (EHRs), as well as from real-world data collected by IoT and personal mobile devices. Currently the explainability of models applied to those data has not gathered the same attention by the research community yet. This trend is actually unexpected and in contrast with their widespread in the real life. EHRs are the major source of tabular data for clinical settings (Payrovnaziri et al. 2020), which contain rich, longitudinal, heterogeneous, and patient-specific information including demographics, clinical information, questionnaire outcomes, lab tests, and vital sign measurements. On the other hand, the recent diffusion of e-health and m-health systems offers increasing opportunities for remote health monitoring and decision making that heavily rely on the analysis of Multivariate Time Series (MTS) (Kok et al. 2022). The integration between these data sources and data-driven AI models may provide a fundamental contribution to the delivery of early, personalised, and high-quality care both in clinical and remote settings, and explainability becomes fundamental to provide effective explanations both to expert and non-expert users.

However, several existing XAI methods are currently not suitable for tabular data due to the unique characteristics that distinguish them from images (and also text records), such as the potential interactions between features and the coexistence of continuous and categorical predictors (Sahakyan et al. 2021). In addition, the majority of methods applied to time series data are generally adapted from computer vision and Natural Language Processing (NLP) fields in order to highlight which specific signal components get the most attention from the model while the classification is performed. As a result, these methods might not account for specific features of time series data, such as recurrent spatiotemporal patterns and correlations between multiple channels and/or sensing modalities. The unintuitive nature of some time series also poses additional challenges as even domain experts may struggle in understanding the information hidden in the most relevant signal components (Rojat et al. 2021).

Selecting existing XAI methods suitable for this data is not sufficient to effectively bring explainability in healthcare applications without an extensive assessment of the generated

explanations (Markus et al. 2021). Clinical validation is currently one of the most widely discussed requirements to build trust in AI-empowered decision making in healthcare (Amann et al. 2020), as it is a critical step for a model to gain clinical credibility by matching data-driven knowledge with evidence-based assessment. In addition, evaluating the level of consistency of explanations generated by multiple methods may provide some preliminary insights into AI systems' reliability, despite sensitivity analysis may be needed to better expose model vulnerabilities and draw more accurate conclusions about stability and robustness properties (Linardatos et al. 2020). However, evaluating clinical soundness and consistency of explanations does not enable a formal quality assessment as well as a systematic comparison of XAI methods, but standard metrics and practices are still missing in the research community (Guidotti et al. 2018). According to Doshi-Velez and Kim (2017), quality evaluation approaches can be divided into *functionality-grounded*, *application-grounded*, and *human-grounded*. The former represents an initial and objective assessment of explanations based on the definition of quantitative metrics, which enable to select the best method regardless end users' needs and preferences. The other two approaches are complementary: human-in-the-loop evaluations are necessary to tune the explanations with respect to the target audience, by considering both domain experts and non-expert users (generally the patient). However, evaluating visual and textual explanations supplied by an algorithm is necessary but not sufficient to enable informed and confident decision-making if the interactivity with the end users is neglected (Arrieta et al. 2020). In fact, multidisciplinary collaboration is the premise to detect relevant interactions between end users and AI systems leading to a better interpretation of model predictions, which should be integrated by design or iteratively added to meet end users' needs.

1.1 Contribution

This survey focuses on the application of XAI techniques to models learned from clinical data stored in EHRs, and real-world data collected by IoT and personal mobile devices. However, it is not sufficient to focus on the application of existing XAI methodologies in this field but it is essential to understand how explanations can be validated and evaluated to close the loop. For this reason, we investigated those works who include also clinical validation, consistency assessment and quality evaluation of the proposed XAI techniques for healthcare. During our research we found several works experimenting XAI methodologies without providing any kind of assessment of the generated explanations. We summarise them in a table to highlight the impact of this research field on the scientific literature, but then we focus on more structured studies including the explanations evaluation, from different perspectives. Specifically, we consider clinical validation to satisfy the strict constraints imposed by the medical domain, consistency assessment of explanations across multiple models and/or XAI methods, and finally formal quality evaluations including novel, objective, and quantitative metrics as well as user-centered studies. Therefore, the contribution of this survey can be summarised as follows:

- an overview of the most prominent XAI methods applicable to tabular and time series data (Section 2);
- a literature survey (methodology is reported in Section 3) related to the usage of XAI in healthcare applications targeting these data types and explanations' assessment (Section 4);

- a discussion section to highlight the main limitations of the presented methods, and open research challenges to improve explainability from both a methodological and user-based perspective (Section 5).

2 Background

Several technical features come into play when analysing the emerging landscape of XAI, which makes the taxonomy of existing methods not unique. Prior surveys addressing XAI from a more general and application-independent perspective classifying methods based on different aspects, which can be summarised as follows:

- **Scope**: *local* or *global*. Local methods aims to explain predictions only for single data instances, whereas global methods enable understanding the reasoning of a learning algorithm as a whole.
- **Stage of applicability**: explainability may be applied throughout the main stages of AI development pipeline, namely pre-modelling, model-development, and post-modelling. However, this classification generally distinguishes between *ante-hoc* and *post-hoc* methods. In the first case, explainability is embedded in the structure of the model and is available directly at the end of the learning phase, whereas in the second case explanatory techniques are used to unveil the *"black-box"* of complex models after their training.
- **Target model**: *model-agnostic* methods can be theoretically applied to any kind of AI model, whereas *model-specific* ones are tailored to certain model classes, such as Convolutional Neural Networks (CNN).
- **Explanation form**: attribution methods generate importance scores for each input, also providing input ranking. Similarly, *heatmaps* such as saliency and attention maps, compute and visualise adaptive weights related to the relevance of each data point. *Decision rules* (i.e., IF-THEN rules), as well as decision trees, represent other common explanation formats. Finally, *dependency plots* show the expected target response as a function of the input features of interest, thus potentially revealing both relationships between target and inputs (e.g., linear, non-linear, monotonic) and interactions among input variables.
- **Algorithmic nuances**: the underlying algorithm used to extract explanations. *Perturbation-based* methods manipulate parts of the input by replacing, removing, or masking them in order to generate attributions for individual features, data points, or signal regions. *Gradient-based* methods are tailored to Deep Neural Networks (DNN), as they obtain attributions by using gradient (i.e., partial derivatives) to compute the impact of each input on model outcomes via one or more forward/backward pass through the network. On the other hand, *instance-based* methods extract a subset of relevant features that is needed to retain/change a given prediction without applying any perturbation to original data.

In addition, taxonomy of XAI methods also depends on the data type that is fed as input to the model to be explained, which can be images, text, graph, tabular, or time series data. As already outlined in Section 1, most of existing techniques have been originally conceived for images or text data, therefore they could not be suitable or readily applicable to

tabular and time series data. For this reason, in the next subsections we first provide the reader with a summary of current XAI methods that are best suited for these data types.

2.1 XAI for tabular data

From the literature analysis, it may be noticed that the majority of the existing XAI techniques applicable to tabular data are model-agnostic. Feature ablation and permutation methods are straightforward options to estimate feature importance for any black-box estimator, by measuring how the prediction error changes when removing a given feature or randomly shuffling its values, respectively. *Mean Decrease in Accuracy* (MDA) is a popular choice in permutation studies, but other scoring metrics can be used as well. Tree-based models also provide an alternative measure of feature importance based on the *Mean Decrease in Impurity* (MDI), in which impurity is quantified by the splitting criterion of the decision tree (normally, Gini's index). Therefore, MDI computes feature importance as the total decrease in node impurity (i.e., homogeneity of labels within the node) for every splits across all trees that include a given feature, weighted by the proportion of samples reached at each node.

Shapley Additive Explanations (SHAP) (Lundberg and Lee 2017) is probably the state-of-the-art method for XAI, and it is built on the concept of Shapley values coming from the coalitional game theory. This concept has been transferred to the Machine Learning (ML) domain by considering a prediction task as a game, features as players, and coalitions as all possible feature subsets, thus making it very suitable for tabular data. SHAP computes feature importance scores as the *average marginal contribution* that each feature brings to an individual prediction, where *"marginal"* stands for the difference between the actual predicted value and a *base value* used as a reference. According to Lundberg and Lee (2017), this value is defined as *"the value that would be predicted if we did not know any feature for the current output"*; in other words, it represents the average prediction over training/test set. On the other hand, the *"average"* terms implies computing the mean value across all permutations, i.e., all the possible subsets that include a specific feature. To apply SHAP provides several advantages. First, local explanations can also be aggregated to get global explanations. In addition, due to the axiomatic assumptions included in SHAP theoretical foundations, global explanations are more reliable than those obtained by most feature attribution methods. Finally, SHAP offers different algorithmic implementations to explain any kind of model.

Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al. 2016) technique is another popular model-agnostic methods to obtain local interpretability. Although a model may be very complex globally, LIME produces an explanation by approximating it by an interpretable surrogate model (generally, a sparse linear model) only in the neighborhood of the instance to be explained. This is achieved by first creating a new dataset consisting of data points randomly drawn in the proximity of the instance of interest, along with the corresponding predictions of the original model. Then, a linear classifier is trained using the perturbed data set, in which each sample is also weighted by its proximity to the target instance through an appropriate weighting kernel. Finally, a very similar method to Least Absolute Shrinkage and Selection Operator (LASSO) regularisation is applied to keep only the most important features. As a result, regression coefficients are used as feature importance scores.

LIME works better for local interpretability, but global explanations may also be derived. A first option is to simply average importance scores across data instances, but

this approach may suffer from a high variance due to multiple local approximations. On the other hand, *LIME SubModular Pick* (LIME-SP) optimisation algorithm allows the selection of a representative, non-redundant set of explanations as exemplary cases of how the model behaves for each class. However, this method just provides some global understanding, and not a comprehensive picture of the overall model reasoning.

Despite the key intuition of using local surrogate models cuts down LIME computational complexity (and time), it reduces outcome stability as well. The choice of simple sparse linear model implies that if the underlying model is highly non-linear even in the locality of the prediction, the explanations may not be faithful. In addition, explanations are originated from random perturbations of the original input space, which may not be representative of the instance to explain. Therefore, several techniques have been proposed trying to improve LIME stability. ALIME (Shankaranarayana and Runje 2019) exploits an auto-encoder as weighting function, whereas hierarchical clustering is adopted in DLIME (Zafar and Khan 2019) instead of random perturbations to group training data, and then it selects the cluster closer to the target instance. In addition, an alternative weighting approach has been proposed in ILIME (ElShawi et al. 2019), in which each perturbed instance is weighted based on its influence on the target instance to be explained, and the distance from it.

Anchors algorithm (Ribeiro et al. 2018) represents an evolution of LIME that exploits reinforcement learning and graph search to detect a region in the neighbourhood, defined by a range of values for some features, representing a sufficient and high-precision condition (i.e., an *"anchor"*) to guarantee local prediction, such that any changes to other features do not essentially alter model outcomes. These range values are then converted into IF-THEN rules, which can be used to explain not only the target instance but also every other instance meeting the anchor.

All the above mentioned methods are aimed at explaining how model outcomes are generated. However, there also exists other techniques, falling within the umbrella of *counterfactual explanations*, aiming at detecting the minimal feature changes that are necessary to drive a prediction towards a desired different output. Counterfactual explanations are generally formulated as an optimisation problem, so the main difference between existing techniques lies in the optimisation method and/or in the loss function to be minimised. A first method, called *unconditional counterfactual explanations*, has been proposed in Wachter et al. (2017) for differentiable models, such as neural networks, in which the gradients of the loss function can be computed. The loss function to be minimised in this case is the distance between the counterfactual and the original data point, subject to the constraint that the model classifies the counterfactual with the desired (and different) label. Guidotti et al. (2018) proposed *Local Rule-Based Explanations of black-box decision systems* (LORE), a model-agnostic method to extend counterfactual explanations beyond differentiable models. This approach exploits a genetic algorithm to create a synthetic neighborhood for a target instance, then it retrieves both a decision rule (similar to an anchor) and a set of counterfactual rules to identify changes leading to different predictions. More recently, Looveren and Klaise (2021) proposed a technique to obtain counterfactual explanations for differentiable classifiers based on prototypes, in which each class-specific prototype is computed as the average encoding over the K nearest instances with the same class label in the latent space generated by a CNN encoder. Once found, prototypes are embedded into the model objective function to guide the perturbations towards an interpretable counterfactual.

Sensitivity analysis is another category of XAI methods aimed at computing feature relevance that works by measuring how much model predictions are sensitive with respect

to changes in one or more input parameters. In addition, it may also be used for model inspection, to detect how altering some internal components/properties affects the model outcomes. Traditional sensitivity analysis methods estimate the importance of each input variable as its contribution to the output model variance. Morris' method (Morris 1991) is one of the most popular approaches for sensitivity analysis. It works by dividing the range of each variable and iteratively making one change at time within the range of each input variable, in order to cluster inputs in three categories: (1) features with no effect, (2) features with linear effect and no interactions, 3) features with non-linear effects and/or interaction effects. Despite this method is very complete, it is also very computational costly, in particular as the number of predictors increases. Therefore, other lightweight solutions have been proposed, such as those based on Analysis of Variance (ANOVA) decomposition (Saltelli et al. 2010). Moreover, adversarial examples represent a more recent and innovative approach to achieve sensitivity analysis, by exploiting the vulnerability of AI models against adversarial attacks as proxy of input relevance. Specifically, they apply intentional changes to input variables in order to generate new samples that can mislead model predictions, then quantify variable relevance depending on how the changed inputs are able to fool the model. However, adversarial example-based sensitivity analysis methods are currently used for computer vision and NLP tasks, while their effectiveness for other data types, such as tabular data, still need to be deeply investigated.

Visual explanation techniques are also available to highlight the relationship among target and input variables. *Partial Dependence Plots* (PDP) (Friedman 2001) show the average marginal effect of one or two features on model outcomes, assuming that the features are uncorrelated (which may not always be true). Their equivalent for local predictions, called *Individual Conditional Expectation* (ICE) plots, has been proposed by Goldstein et al. (2015) to visualise the dependence of the prediction on a feature for each instance separately. Finally, *Accumulated Local Effects* (ALE) (Apley and Zhu 2020) plots represent an unbiased alternative to PDP, as they account for feature correlation when showing feature influence on model outcomes.

To conclude this overview of XAI techniques suitable for tabular data, it is worth to briefly mention some methods available to globally explain complex models by approximating them with simpler ones, such as decision rules/trees. *InTrees* (Deng 2019) has been proposed as a framework to extract a compact set of decision rules from tree ensembles, by selecting and pruning rules according to a trade-off among their frequency within tree nodes, their error rate, as well as their length. Additional methods have been also developed for approximating DNN (Wu 2018) and Support Vector Machine (SVM) models (Barakat and Bradley 2007).

2.2 XAI for time series data

To date, Recurrent Neural Networks (RNN) generally represent the best strategy to deal with time series data, thanks to their memory state and their ability to learn relations through time. CNN with temporal convolutional layers are able to build temporal relationships as well, also extracting high level features from raw data. The introduction of these models to solve MTS classification and forecasting tasks significantly boosted predictive performance without requiring heavy data pre-processing. As a result, the majority of existing XAI methods applicable to this data are specific for these models.

As far as CNN is concerned, almost all methods are inherited from computer vision field to obtain post-hoc explainability. According to the underlying algorithm concepts,

they can be divided into gradient-based and perturbation-based methods. Gradient-based methods measure how much a change around a local neighborhood of the input corresponds to a change in the model output by running a single forward or backward pass of Gradient Descent (GD) algorithm (or similar) in the network. They have been originally conceived as pixel attribution methods, also referred to as saliency maps, in order to highlight the pixels/super-pixels that are relevant for a certain image classification. However, they may also be adapted for time series data in order to highlight the most relevant data points within a 1-D sequence.

A first gradient-based explanation method has been proposed in Simonyan et al. (2013) to create saliency maps corresponding to the gradient of an output neuron with respect to changes in a small neighborhood around the input, thus highlighting image regions that are relevant for a target class. Afterwards, Sundararajan et al. (2017) proposed *Integrated Gradients*, which essentially represent a variation of the gradient computing technique implemented in the previous method, by directly attributing the network predictions to its input features.

Deep Learning Important FeaTures (DeepLIFT) (Shrikumar et al. 2017) is one of the most popular algorithms for post-hoc explanation of deep networks. It also works by attributing importance scores to input features, in which each score represents the impact on network outcomes of changing the original feature value to a reference baseline, which can be empirically chosen by end users. This approach is also known as *"Gradient*Input"* methods, since it also multiplies the gradient by the input signal. This operation essentially represents a $1^{s}t$ order Taylor approximation of the output changes when inputs are set to zero, and it has proven to enhance the visualisation of saliency maps with respect to previous gradient-based methods. Moreover, DeepLIFT has been later combined with Shapley values to build *DeepSHAP*, a specific framework to approximate SHAP feature attributions for any Deep Learning (DL) model. This method differs from the original DeepLIFT by using a distribution of background samples instead of a single reference value to change each selected feature, and using Shapley equations to linearise network components such as softmax operators.

Deconvolution (Zeiler and Fergus 2014), is a technique to visualise CNN-based saliency maps by using *DeconvNets* networks (Zeiler et al. 2011), which leverage the same CNN layers and operators in an exactly reversed order for mapping encoded features to input (i.e.,pixels), as opposite to the standard CNN data processing pipeline. Moreover, the *guided backpropagation* technique (Springenberg et al. 2014), also known as guided saliency, has been proposed as a variant of the deconvolution approach to extend its applicability to all possible CNN architectures to visualise saliency of the learned features.

Class Activation Mapping (CAM) (Zhou et al. 2016) is another CNN-based XAI methods, originally developed to detect class-specific image regions used by the network to make predictions. CAM computes a vector by concatenating the average activations of convolutional feature maps that are placed immediately before the last prediction layer, then it feeds a weighted sum of this vector to the final layer. In this way, the relevance of class-specific image regions, and of features learned in the latent space in general, can be retrieved by projecting back the weights of the output layer onto the convolutional feature map. However, CAM implementation requires CNN to have a specific architecture in their final layers, thus limiting its applicability. In addition, it is suitable only to highlight high-level representations learned at the last stages, whereas it cannot provide any explanation of low-level representations that are learned at earlier stages. To overcome these limitations, a more general CAM implementation, *Gradient-weighted CAM* (Grad-CAM) (Selvaraju et al. 2017), has been developed to extend its applicability to any CNN, which relies

on gradient information flowing to the last convolutional layer to locate the most important image regions within a saliency map in an architecture-independent fashion.

Finally, *Layer-wise Relevant Propagation* (LRP) (Bach et al. 2015) is an interpretability method to decompose DNN by propagating their predictions backward without altering the output magnitude. By starting from neurons in the final prediction layer, and moving back to neurons of the input layer, the prediction value is backpropagated in such a way that each neuron redistributes to the preceding layer the same amount of information received from the higher layer.

Unlike gradient-based methods, perturbation-based methods compute the contribution of individual parts of the input by removing or randomly replacing them, then they exploit a distance metric to measure the difference in the model decision function. As a result, a higher difference in the prediction outcome indicate a higher contribution of the input component that has been altered. In this context, the *Occlusion* method by Zeiler and Fergus (2014) is one of the most used techniques coming from computer vision applications. It acts as sensitivity analysis method by systematically replacing different contiguous parts of an input with a given baseline, then monitoring the decrease in the prediction function. The implementation of this method is not computationally expensive and can be applied to any network architecture as it does not require specific internal components.

For what concerns RNN, *Attention* mechanism (Bahdanau et al. 2014), sometimes also referred as *Self-Attention*, is currently the state-of-the-art method for explainability. Attention originates from NLP domain, where it has been initially proposed as a solution to overcome the bottleneck of the original Encoder-Decoder RNN model employed for machine translation. It encodes the input sequence to one fixed length vector from which decoding the output at each time step. The main criticism of this approach is related to the difficulties for DL models to cope with very long sentences, especially those that are longer than the ones contained in the training corpus. Differently, the attention model does not encode input sequences as a whole, but rather it develops context vectors that are filtered specifically for each output time step. Then, it searches for a set of positions within a source sentence where the most relevant information is concentrated according to the context vectors associated with these positions and all the previous generated target words, in order to predict the next word. Besides providing a gain in predictive performance, attention also represents a powerful explanatory technique, as it highlights which words in a text corpus are the most relevant for a certain prediction.

Attention mechanism is at the foundation of *Attentive transformers* (Vaswani et al. 2017), which consist of specific RNN modules that can be embedded into the learning process of neural networks to obtain ante-hoc explainability. The inner functioning of a transformer relies on a sparse-max function to obtain a mask, which is subsequently scaled and multiplicatively applied to the input, in order to learn adaptive weights that reflect the impact of each input data on the final prediction. Transformers can be used to enable interpretability at different levels, such as input features and time points. For instance, they can detect globally important variables, persistent temporal patterns, as well as significant events within a data trajectory leading to a target outcome. Transformers have found their major applications in NLP, as demonstrated by very popular language models such as *BERT* (Devlin et al. 2018) and *GPT-3* (Radford et al. 2018). However, with the advent of vision transformers for image-based applications (Mondal et al. 2021), they have been also embedded into CNN architectures. As a result, attention is the primary XAI method to explain RNN, while attention-enhanced CNN may also be found in some cases.

Eventually, there are also additional data mining methods applicable to time series data, although less used in practice. Fuzzy inference systems are a viable solution to simulate

logical thinking, whereas *Symbolic Aggregate ApproXimation* (SAX) (Lin et al. 2003) works by converting time series into strings. Specifically, it first divides each time series into equally-length segments. Then, by assuming a Gaussian distribution of input data, it assigns a symbol to each segment by mapping the average segment value with the corresponding probability, in order to discover recurrent patterns. Prototype Learning (PL) is also a compatible approach for time series data, which generates samples to be used as reference to explain the typical pattern of all data instance belonging to the same class. In this context, *Shapelets* (Ye and Keogh 2009) is a time series-specific, prototype-based method to explain AI models by extracting input sub-sequences that are representative to each class. This method also provides more interpretable, accurate, and faster results with respect to the standard PL approach that selects class prototypes from the nearest samples in latent embedding space to a given data point.

3 Survey methodology

We searched *IEEExplore*, *Springer*, *ACM*, and *Elsevier* digital libraries, using different search strings obtained as a combination of the following subsets of keywords:

1. *"Explainable"*, *"Interpretable"*;
2. *"AI"*, *"Artificial Intelligence"*, *"Machine Learning"*;
3. *"Healthcare"*, *"Health"*, *"Wellbeing"*.

In addition, we also analysed *Scopus* to double check the screening process conducted for the previous databases. This search also highlighted an additional cluster of relevant works coming from other sources, such as Nature Research journals, which have been further investigated.

Motivated by the very recent development and application of XAI methods, and especially in the healthcare domain, we filtered out the search by selecting papers published in the last 5 years (i.e., between 2017 and 2021). Table 2 reports the percentage of selected papers for each year within the date range, and it further confirms the latest exponential increase of XAI applications in the health domain, with the majority of research studies published in the last 2 years.

In order to focus only on XAI methods suitable for tabular and time series/sequence data, we first excluded applications targeting biomedical images, such as computerised tomography scans, magnetic resonance imaging, and ultrasound images, which are typically 3-D data with additional time dimension and/or multiple channels (4-5D) represented by tensors to be fed to DNN models. Then, we also excluded NLP tasks to extract meaning from unstructured medical text records (e.g., patient prescription notes), which generally make use of high-dimensional word embeddings as input data. Finally, we did not consider genomics and molecular biology applications based on graph data structures.

In turn, tabular data include independent (i.e., "static") EHR, as well as feature datasets derived from physiological signals. In the first case, each patient's record is treated as an independent observation, or else multiple observations are aggregated over a specific time window, still resulting in static data. In the second case, feature datasets are a high-level and multidimensional input representation resulting from the application of signal processing pipelines, such as signal framing/windowing and handcrafted feature extraction. On the other hand, time series/sequence data may be

Table 1 Inclusion and exclusion criteria

Inclusion criteria	Exclusion criteria
Application to tabular data	Application to biomedical images
Application to time series/ sequence data	Application to medical text data
Post-hoc XAI	Application to graph data
Ante-hoc XAI	IML models only

Table 2 Percentage of research works by date

Year	Papers (%)
2021	60.6
2020	31.0
2019	5.6
2018	1.4
2017	1.4

Table 3 Number of research works by data source

Digital library	Papers (#)	
	Journal	Conference
IEEExplore	15	15
Springer	10	3
ACM	7	7
Elsevier	6	0
Scopus	8	0

divided into longitudinal EHR, which consist of a sequence of visits/admissions to model patient trajectories using clinical variables or medical codes (e.g., ICD-10-CM codes), and raw signal time series.

For what concerns the stage of XAI applicability while building a model, we considered as eligible both post-hoc methods to add interpretability to already developed black-box models (e.g., tree ensembles, CNN), and ante-hoc methods that embed interpretability in the structure of the model, thus making it available directly at the end of the learning phase. However, we did not include *"glass-box"* approaches in which interpretability is simply addressed in terms of development of intrinsically Interpretable ML (IML) models. This typically involves three model classes, namely sparse linear classifiers (e.g., linear/logistic regression, generalised additive models), discretisation methods (e.g., rule-based learners, decision trees), and instance-based models (e.g., k-Nearest Neighbors, (k-NN)) (Du et al. 2019). A summary of the inclusion and exclusion criteria used for the literature survey is listed in Table 1.

As a result, we reached a total of 71 publications at the end of the search, including 46 journal articles (64.8%) and 25 conference papers (35.2%) reporting original studies. The number of selected papers for each digital library is shown in Table 3, whereas Table 4 illustrates the distribution of surveyed articles across the journals.

Table 4 Number of surveyed articles by journal

Journal	Publications (#)
IEEE Journal of Biomedical Health Informatics	7
BMC Medical Informatics and Decision Making	7
Nature Scientific Reports	6
IEEE Access	4
ACM Transactions on Computing for Healthcare	3
Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)	3
IEEE Transactions on Visualization and Computer Graphics	2
Computers and Biology in Medicine	2
Artificial Intelligence in Medicine	2
Nature Communications	2
IEEE Internet of Things Journal	1
BMC Bioinformatics	1
IEEE Journal of Translational Engineering in Health and Medicine	1
Journal of Translational Medicine	1
Journal of Management Information Systems	1
Computer Methods and Programs in Biomedicine	1
ACM Transactions on Management Information Systems	1
The Lancet Digital Health	1

4 Results

In the previous section, we outlined the search strategy adopted in this survey, with particular reference to the input data type and the stage of application of XAI methods within the AI development pipeline. In this section, we present the revised works based on their main contribution in evaluating the explainability applied to the target health application. For a complete list of acronyms used in tables and throughout the text please refer to Table 9 in Section 1 of the Appendix.

A first research branch consists of exploratory studies aimed at experimenting XAI methodologies in order to demonstrate their possible integration with complex models to enable global/local interpretability in specific health applications, while maintaining good predictive performances. The most relevant preliminary works using tabular and time series data are summarised in Table 10 and Table 11 in Section B of the Appendix, respectively. However, these studies do not perform any evaluation of the proposed explanations, leaving room to further investigation as fundamental step to enhance confidence and trust of the medical community in decision making based on predictive AI. After this analysis, we move a step forward by focusing on studies that also included an explanation assessment from one or more of the following perspectives:

1. *clinical validation*: alignment with existing medical knowledge/practice;
2. *consistency assessment*: level of agreement of explanations provided by multiple models and/or XAI methods;
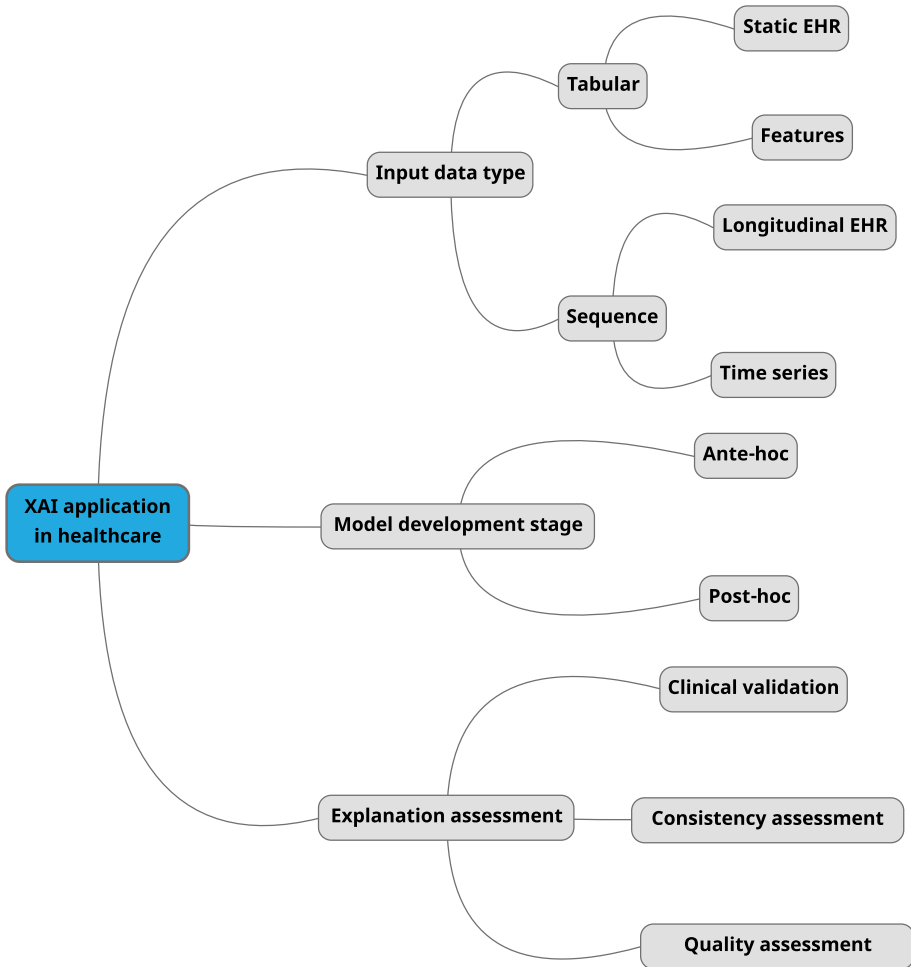
Fig. 1 Taxonomy mind map of XAI methods used for this survey

3. *quality assessment*: it includes both quantitative evaluations based on novel metrics, and qualitative evaluations through clinician ratings and feedback.

Such evaluations are complementary, so they may be conducted concurrently to strengthen the trust in a model. In the case more than one of the above explanation assessment procedures is performed, we present the most relevant findings. A taxonomy mind map is shown in Figure 1 to support the reader in understanding the main aspects covered in the overall process of selection, grouping, and analysis of the collected research works.

For each category, the revised works are also summarised in tables specifying the target application, input data type and datasets, AI models, and XAI methods. These tables also include a citation analysis derived from Google Scholar and updated to early September 2022, to highlight the impact of the research works. However, it is worth noting that this literature survey is limited to the past 5 years due to the very recent application of XAI in healthcare, and that > 90% of the selected papers has been published in the last 2 years.

Therefore, the number of citations provides only a preliminary analysis of the impact of the research in this field.

4.1 Clinical validation

Evaluating explanations from a clinical standpoint is crucial to guarantee that the inner reasoning of a model follows the domain knowledge, at least with respect to the most important and well-established clinical guidelines. In other words, it demonstrates that model behaviour appears to be "human-like"), which also adds clinical credibility to the model itself. The research works presented in this section propose a preliminary clinical validation of the generated explanations, which is conducted either through domain expert surveys or through the comparison with the related medical literature. For tabular data this generally includes to investigate the global relationships between a target health condition and the input predictors evidenced by feature attribution or rule-based methods. Related research studies are summarised in Table 5. On the other hand, clinical validation of explanations generated by AI systems based on time series data is performed for two main reasons:

- for longitudinal EHR, to assess if the evolution of patient trajectories resulting in a target health condition is aligned with the general clinical knowledge.
- for physiological signals, to investigate whether the most important components highlighted by the model are clinically relevant, in order to assert decisions are made upon meaningful features.

Research studies in this area are shown in Table 6. Eventually, clinical comparison may also support knowledge discovery to learn novel relationships and patterns with emergent risk factors, which might be further investigated for a future integration in the current clinical practice.

XAI methods have been applied to a wide variety of CAD applications based on tabular data, and especially on EHR. Several solutions have been identified in this survey explaining the detection of neurological disorders. Beebe-Wang et al. (2021) proposed an explainable risk assessment for imminent (i.e., within 3 years) dementia diagnosis, by training XGBoost, Multi-Layer Perceptron (MLP), and Long Short-Term Memory (LSTM) models with multi-year, extensive cognitive testing data coming from the Religious Orders Study and Rush Memory and Aging Project (ROSMAP) (Bennett et al. 2012).

Then, they applied SHAP to the best XGBoost classifier for both feature selection and model explanation. Results indicate that the most relevant features come from cognitive tests collected in the most recent year, which is further confirmed by the absence of improvements in model performance when considering older visits or cumulative data to make predictions. This finding suggests that longitudinal testing may not be necessary for future dementia diagnosis, which is consistent with other studies reporting the modest value of gradual cognitive changes in predicting future dementia onset. In addition, SHAP feature ranking shows that the final model focuses on few cognitive tests that can be collected in a single visit in less than 20 minutes. Specifically, combining episodic memory tests with executive functioning or language tests led to a predictive accuracy comparable with that of a full cognitive test battery (98 minutes), in line with the current neuropsychology.

Sha et al. (2021) proposed a novel computational framework, *Systems Metabolomics using Interpretable Learning and Evolution* (SMILE), for supervised metabolomics data

Table 5 List of XAI studies with tabular data performing clinical validation of explanations

References	# Cit.	Application	Input Data	AI model(s)	XAI method(s)	Dataset(s)
Beebe-Wang et al. (2021)	7	Imminent dementia diagnosis	EHR	XGBoost	SHAP	ROSMAP dataset Bennett et al. (2012)
Sha et al. (2021)	4	AD diagnosis	Plasma metabolite concentrations	Evolutionary algorithms	LGP	AD metabolomic dataset Wang et al. (2014)
Kim et al. (2021)	3	END detection in stroke patients	EHR	LightGBM	SHAP	K-Attention dataset Jung et al. (2019)
Rashed-Al-Mahfuz et al. (2021)	6	CKD diagnosis	EHR	RF, GBDT, XGBoost	SHAP	UCI CKD dataset Rubini and Eswaran (2015)
Pang et al. (2019)	10	Early childhood obesity prediction	EHR	XGBoost	SHAP	Retrospective study
Zeng et al. (2021)	5	Post-operative complication risk prediction	EHR	XGBoost	SHAP	Retrospective study
Zhang et al. (2021)	10	post-LT AKI prediction	EHR	RF	SHAP	Retrospective study
Lu et al. (2020)	5	COVID-19 diagnosis and prognosis	EHR	GBDT, RF	SHAP	Retrospective study
Gupta et al. (2021)	1	COVID-19 recovered subject detection	HR/HRV features	XGBoost	SHAP	Retrospective study
Pal and Sankarasubbu (2021)	50	COVID-19 diagnosis	Cough sounds, symptoms metadata	TabNet, DNN	Attention	Pilot study
Dissanayake et al. (2020)	18	Heart anomaly detection	PCG features	stacked LSTM-CNN-MLP network	SHAP, occlusion maps	PhysioNet DB Goldberger et al. (2000)

Table 6 List of XAI studies with time series data performing clinical validation of explanations

References	# Cit.	Application	Input Data	AI model(s)	XAI method(s)	Dataset(s)
Shashikumar et al. (2021)	15	Sepsis prediction	Longitudinal EHR	WCPH-RNN	Saliency	Retrospective study
Sun et al. (2021)	1	Survival prediction	Longitudinal EHR	RNN, GNN	Attention	WHAS Lemeshow et al. (2011), SUPPORT Knaus et al. (1995), META-BRIC Curtis et al. (2012), MIMIC-III, PLAGH
Zheng et al. (2020)	9	Mortality prediction, AKI prediction	Longitudinal EHR	TITV network	FiLM, attention	NUH-AKI, MIMIC-III
Kwon et al. (2018)	196	HF diagnosis	Longitudinal EHR	RETAIN variant	Attention	HIRA-NPS Kim et al. (2014)
Ivaturi et al. (2021)	4	AF detection	ECG	CNN	Feature ablation/permutation, LIME, guided saliency	PhysioNet 2017 Clifford et al. (2017)
Mousavi et al. (2020)	42	AF detection	ECG	Bi-LSTM ensemble	Attention	PhysioNet 2017, MIT-BIH AFIB[1]
Filtjens et al. (2021)	3	pre-FoG movement detection in PD patients	3-D kinematic joint trajectories	CNN	LRP	Gait dataset Spildooren et al. (2010)

analysis aimed at Alzheimer disease (AD) diagnosis. SMILE exploits Linear Genetic Programming (LGP) as evolutionary algorithm to generate a compact predictive model, and it uses metabolite concentrations as input features stored in separated registers in a LGP program. The algorithm has been implemented using a metabolomic dataset reported in Wang et al. (2014), which includes plasma concentration of 242 metabolites from 57 AD subjects, 58 Mild Cognitive Impairment (MCI) subjects, and 57 healthy controls. For explainability purpose, features have been ranked according to their occurrence in the best evolved models, thus providing a way to assess their importance in predicting the disease. In addition, the co-occurrence frequency between each pair of features has been considered as a correlation measure. SMILE analysis highlighted many key metabolites that have been previously linked to AD, but also others less clinically investigated that can be potentially correlated with AD onset. In addition, SMILE performance degraded when detecting AD from MCI subjects, suggesting for possible similarities in biomarkers between the two conditions.

Kim et al. (2021) proposed an interpretable model to predict Early Neurological Deterioration (END) in stroke patients from heterogeneous EHR data. They trained several models by using data from 2363 subjects included in the Korean Atrial Fibrillation Evaluation Registry in Ischemic Stroke Patients (K-ATTENTION) (Jung et al. 2019), a real-world dataset composed of multi-center prospective registries that mainly focus on characteristics, oral anticoagulant use, and clinical outcomes of stroke patients. Then, SHAP has been applied to the best performing Light Gradient Boosting Machine (LightGBM) model to identify the most relevant risk factors for END. Specifically, results obtained from the analysis of SHAP dependence plots reveal clear cut-off values associated with positive and negative probabilities of END occurrence for the 4 main representative features, whose clinical implication may be applicable to real-world clinical practice. For instance, they suggest that patients with severe stroke tend to develop END, thus imposing a special attention for them, while patients with mild to moderate stroke have a lower probability to develop END. In addition, the cut-off value of fasting glucose concentration for predicting END is very similar to the current diagnostic criteria used for diabetes diagnosis (126 mg/dL).

XAI methods can also be used to assess the impact of several risk factors to develop chronic health disorders. Rashed-Al-Mahfuz et al. (2021) proposed a ML framework to detect Chronic Kidney Disease (CKD) from lab test data. To this aim, they evaluated different tree ensemble algorithms on the UCI CKD dataset (Rubini and Eswaran 2015), which consists of clinical tests collected from 400 patients with a total of 24 variables. SHAP values have been computed for each model, then 13 features with both the highest ranking and overlap across the models have been chosen as optimised subset to train a smaller RF model. Moreover, these predictors have been further categorised according to their source of acquisition (.e.g,, blood test, urine test, others) to develop additional RF models based on all possible combinations. The optimised and pathologically-selected subsets obtained by analysing SHAP explanations reach similar performance to the full input set, demonstrating that an accurate and early CKD diagnosis may be achieved by using few, low-cost, and clinically-relevant screening tools.

Pang et al. (2019) proposed an interpretable ML approach to analyse the main risk factors associated to early childhood obesity, including demographic characteristics, lab parameters, as well as anthropometric markers and vital sign measurements. The authors trained several ML models on the Pediatric Big Data repository, a clinical database derived from the EHR system at the Children's Hospital of Philadelphia with more than 860 children, then they applied SHAP to explain the best performing XGBoost model. Results

show that Well-known obesity risk factors such as weight, height, weight-for-height, geographic location, race and ethnicity appear among the most important features to the model. On the other hand, SHAP analysis also highlights novel factors that are related to human metabolism, such as Body Temperature (BT) and respiratory rate, which may deserve further investigation to unveil possible physiological mechanisms causing these associations.

Clinical validation is also fundamental when evaluating explanations related to risk prediction models developed for clinical settings, such as those related to life expectancy, post-operative outcomes, and hospital attendance. Zeng et al. (2021) proposed an explainable ML model for post-operative complication risk prediction of congenital heart surgery patients from patient- and surgery-specific features and intra-operative physiological time series. To this aim, they trained a XGBoost model on a private dataset containing data from 1964 pediatric patient, reaching 83.1% accuracy and 0.85 Area Under the ROC Curve (AUC) values for multi-label classification with five complication types (i.e., lung, cardiac, rhythm, infectious, others). Then, SHAP has been used to detect the most relevant factors and to perform an extensive clinical comparison of the main risk profiles learned by the model, all of which resulted to be clinically relevant. In particular, high blood pressure and prolonged cardiopulmonary bypass time patterns confirmed a high correlation with worse post-operative outcomes.

Zhang et al. (2021) developed an explainable model for the early prediction of Acute Kidney Injury (AKI) after Liver Transplantation (LT) using more than 100 variables mainly covering patient/donor demographic and clinical characteristics, such as comorbidities, laboratory values, and medications. They performed a retrospective data collection of adult LT cases to build two separate datasets, which have been used for internal and external validation of several ML models, respectively. SHAP explanation analysis for the best performing RF classifier indicated that higher pre-operative indirect bilirubin concentration, lower urine output, lower platelet count, longer anesthesia time, and high graft steatosis percentage are associated with a higher probability of AKI onset. The distribution and relation of these risk factors with AKI diagnosis match with the current physio-pathology, thus adding clinical credibility to the final model.

Explaining health predictions based on static EHR may be useful to detect the most important features, and to evaluate their relevance and correlation with respect to the model outcome from a clinical standpoint. On the other hand, modeling longitudinal EHR data collected during multiple visits/examinations or hospital admissions may enable learning how the evolution of patients' clinical trajectories impact on a target health condition. Shashikumar et al. (2021) presented *Deep Artificial Intelligence Sepsis Expert* (DeepAISE), a novel interpretable recurrent survival model for periodical sepsis prediction after ICU admission from longitudinal lab tests and physiological measurements, such as HR, mean arterial pressure, pulse oximetry (SpO_2), respiration, and BT. DeepAISE combines a Gated Recurrent Unit (GRU)-based RNN with a Weibull Cox Proportional Hazards (WCPH) semi-parametric survival model (Cox 1992) to learn predictive features related to higher-order interactions and temporal patterns among clinical risk factors that maximise the data likelihood of observed time to septic events. Specifically, it starts by producing risk scores 4 hours after ICU admission, then it predicts the probability of onset of sepsis with 2-hour resolution for the next 12 hours. The model has been subjected to a internal validation using the Emory cohort dataset (ICU patients admitted to the hospitals within the Emory Healthcare system in Atlanta, Georgia from 2014 to 2018), and to an external validation using a patient cohort taken by the Medical Information Mart for Intensive Care (MIMIC)-III database (Johnson et al. 2016), in order to ensure robustness against potential changes due to different internal procedures and patients' characteristics.

To get model explanations, feature relevance scores have been computed as the gradient of the sepsis risk score with respect to all input features, in a similar way to CNN-based saliency maps. Analysis of the top-10 features confirmed that the system exploits predictors that have been already identified as risk factors for sepsis, such as recent surgery, length of ICU stay, HR, Glasgow coma score, white blood cell count, and temperature, as well as some less appreciated but known factors such as low blood phosphorous levels. A feature permutation study has also been performed by replacing the top-10 features at both global and local levels with random and/or missing values in order to assess the impact on the model performance. Obtained results indicate that locally important features may provide a better overview of the actual top contributing factors to individual risk scores, since local perturbations yield a significant drop in the model performance with respect to the global replacement strategy.

Sun et al. (2021) proposed *AttenSurv*, an attention-based RNN for Heart Failure (HF) survival prediction of seriously ill patients from longitudinal and heterogeneous EHR. The network consists of three modules: 1) a Bidirectional LSTM (Bi-LSTM) network to learn the latent representation of a patient trajectory; 2) a MLP network for survival prediction; 3) an attentive transformer to detect global critical risk factors. In addition, the authors also proposed an enhanced variant, named *GNNAttenSurv*, which also incorporates a Graph Neural Network (GNN) module to extract the latent correlation between risk factors. Both networks have been tested on three public follow-up datasets, namely WHAS (Lemeshow et al. 2011), SUPPORT (Knaus et al. 1995), and METABRIC (Curtis et al. 2012), and on two EHR datasets, the MIMIC-III DB and the Chinese PLAGH dataset, using different sets of dataset-specific features including lab tests, vital signs, demographics, and treatment information. The top-10 features identified by the model for different survival time horizons (ranging from 3 days up to 2 years) have been reviewed by different medical experts, and the resulting assessment demonstrated that they represent truly informative risk factors, with some of them currently adopted for HF survival prediction in the clinical practice.

Zheng et al. (2020) proposed a general DL framework, *TRACER*, to facilitate accurate and interpretable decision making in healthcare applications, using in-hospital acquired AKI and mortality prediction as cases studies. The framework relies on a RNN model as core component, named Time-Invariant and Time-Variant (TITV) network, which is designed to learn both time-variant and time-invariant feature importance scores for each patient into two separate sub-modules, by using a self-attention network and a Feature-wise Linear Modulation (FiLM)-based network (Perez et al. 2018), respectively. The proposed network has been evaluated on the NUH-AKI dataset (> 100k patients from the National University Hospital in Singapore) for AKI diagnosis, and on the MIMIC-III dataset for mortality prediction. Then, extensive feature-level clinical interpretation has been performed in both domains. This analysis highlighted similar temporal patterns for features that share a similar physiological functionality, whereas diverging patterns have been found for features that have contrasting functionalities or reflect different patient clusters. Overall, feature rankings generally agree with the clinical relevance of the corresponding risk factors.

Kwon et al. (2018) developed a novel visual analytics tool, named *RetainVis*, to enhance interpretability and interactivity of RNN outcomes for disease diagnosis from longitudinal EHR, by integrating model explanations with additional functionalities, such as visualisation of historical patient trends, patient grouping according to desired criteria, and comparison with reference values in selected patient cohorts. For what concerns the DL model, RetainVis relies on a variant of the original REverse Time AttentIoN (RETAIN)

network (Choi et al. 2016), named RETAINEx, which exploits Bi-LSTM modules with non-uniform time interval embedding to model irregular time spacing across consecutive visits, and a double attention mechanism at both time and feature levels. The network has been tested over the HIRA-NPS dataset (Kim et al. 2014), containing medical information of approximately 1.4 million of Korean patients, using HF diagnosis as main case study. In addition, the authors conducted an extensive analysis to review if the medical diagnoses and treatments that received the highest attention within the trajectory of patients that develop heart failure are supported by clinical evidence. Obtained results confirmed the premise that hypertensive, metabolic, and ischaemic heart diseases, and obesity are the main leading factors for heart failure, as well as one of the main related comorbidities.

The recent COVID-19 pandemic has also fostered the development of several solutions aimed at explaining disease diagnosis based on different data sources and analysis approaches, involving both tabular data and signal processing methods. Lu et al. (2020) proposed an explainable system to diagnose COVID-19 in suspected patients and then to predict mortality of confirmed cases using lab test data (e.g., nucleic acid test, blood test), also including medical text reports of basic diseases and symptoms. Specifically, they used a Gradient Boosting Decision Tree (GBDT) model for disease diagnosis, whereas RF was the best choice to predict mortality. In the first case, the GBDT model has been evaluated on a private COVID-19 dataset coming from Wuhan hospital (EHR data from 350 patients), whereas a public dataset available in Yan et al. (2020) has been used for the mortality prediction task (485 patients) by considering only lactic dehydrogenase, lymphocyte, and C-Reactive Protein (CRP) as features. SHAP analysis shows that procalcitonin and white blood cell are the most relevant features for COVID-19 diagnosis, in line with the current clinical findings. Unfortunately, the analysis of the explanatory power of textual features extracted through Term Frequency Inverse Document Frequency (TF-IDF) (Weiss et al. 2010) method is not reported, which might add further value to explanation assessment. As far as mortality prediction is concerned, results demonstrate that when the level of LDH and CRP rises and the level of lymphocyte decreases, the death probability is higher, which agrees with clinical features of death cases. In addition, SHAP dependence plots also highlight clear boundaries associated with rising and decreasing patterns of death probability for each feature, which may act as starting point to further investigate the impact of these risk factors.

Gupta et al. (2021) detected COVID-19 recovered subjects from healthy controls using ECG-based Heart Rate (HR) and Heart Rate Variability (HRV) features. They trained seven ML models by using 1-minute ECG recordings from a total of 447 subjects collected at two hospitals in Delhi, India, then they applied SHAP to the best performing XGBoost model. From this study, it may be inferred that high-frequency power, normalised high-frequency power, HRV standard deviation, low-frequency power, and low-to-high frequency power ratio are the most influenced features after COVID-19 infection, and that changes exhibited by these features are related to an increased vagal activity. These findings match with earlier studies suggesting that heart vagal stimulation increases in the post-COVID recovery phase (Bonaz et al. 2020).

Pal and Sankarasubbu (2021) proposed a mixed approach for early COVID-19 diagnosis by integrating symptoms metadata and cough sounds. The model architecture consists of two sub-components that are concatenated to obtain a final prediction: a TabNet (Arık and Pfister 2021) for generating embeddings from patient characteristics, diagnosis, and symptoms, and a DNN to generate cough embeddings from temporal and spectral acoustic features extracted through signal processing. Both networks integrate an attentive transformer to learn feature relevance from each data modality. The evaluation conducted on

a medical dataset containing 30k cough audio segments and associated symptoms from 150 patients with four cough classes (COVID-19, asthma, bronchitis, and healthy) pointed out that more accurate predictions can be achieved using symptoms metadata than cough features, while the overall performance increase by combining both data sources. As it may be expected, attention distribution highlight fever, cough, dizziness, and chest pain as the most recurrent symptoms for infected subjects. In addition, the authors perform an in-depth clinical analysis of the main significant differences in the energy distribution of the cough spectrum between COVID-19 and other cough types. Overall, results confirm that the model is able to learn the main relationships between the frequency distribution of the most discriminating features and the underlying cough sound characteristics for each class. These findings are also supported by t-distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten and Hinton 2008) visualisation, showing a clear separation between the four clusters of cough features learned by the model.

Several solutions based on signal processing methods have also been proposed to explain the detection of heart disorders, which are mainly based on the analysis of ECG recordings. Ivaturi et al. (2021) presented a XAI framework for AF prediction from single-lead ECG signals. To this aim, they first trained a *MobileNet* (Howard et al. 2017) CNN architecture on the PhysioNet/Computing in Cardiology Challenge 2017 dataset (Clifford et al. 2017). Then, global explanation analysis has been performed by dividing each ECG cycle into 8 equally size segments, each one corresponding approximately to a region of interest (e.g., P wave, T wave, isoelectric baseline), and applying feature ablation, feature permutation, and LIME methods to highlight the most relevant segments. Moreover, saliency maps with guided back-propagation technique have also been used to compare the direct contribution of raw input data to local predictions with global, segment-based analysis. Clinical analysis of explanations shows that the network effectively focuses on physiological features that match with those used by cardiologists for the clinical AF diagnosis, such as the absence of P-wave, variability of R-R intervals, and electrical activity in the isoelectric region of the ECG.

Mousavi et al. (2020) proposed *HAN-ECG*, an alternative solution for explaining AF predictions from single-lead ECG. This system differs from Ivaturi et al. (2021) as it relies on a stacked Bi-LSTM ensemble with a hierarchical attention mechanism to learn relevant components of the input signal at different levels, namely beat, wave, and time windows, respectively. The network has been evaluated over two datasets, including the PhysioNet 2017 and the MIT-BIH AFIB[1], then the visualisation of attention layers has been exploited to demonstrate that the model focuses on clinically relevant heart beats and waves for detecting AF arrhythmia. As in Ivaturi et al. (2021), the absence of P-waves, which may be occasionally replaced with a series of small waves called fibrillation waves, and the irregular R-R intervals in which the heartbeat intervals are not rhythmic played an essential role in AF detection.

Dissanayake et al. (2020) developed an interpretable DL framework for heart anomaly detection from Mel-Frequency Cepstral Coefficient (MFCC) spectral features (Clifford et al. 2017) extracted from phonocardiogram (PCG) signals (i.e., heart sounds). The framework combines a pre-trained LSTM network for automatic segmentation of the input MFCC maps, a CNN encoder to perform spatial feature learning on the supplied feature map, and a MLP network to get the final prediction. Different network architectures have

---

[1] https://physionet.org/content/afdb/1.0.0/

been tested through the combination of the above modules in order to explicitly examine the importance of signal segmentation as a prior step to classification. Then, both SHAP and occlusion maps have been used to explain the hidden representations learned by the model. Experimental results obtained on the benchmark PhysioNet database (Goldberger et al. 2000) indicate that the network architecture without segmentation module reaches the highest accuracy, outperforming the state-of-the-art methods. In addition, both XAI methods show that a correct classification of PCG signals occur when the model mainly focuses on learned features that are placed within (or between) two fundamental heart sound locations, namely S1 and S2 segments. This model behaviour accords with the clinical assessment followed in digital phonocardiography. For what concerns the role of signal segmentation, these findings suggest that if the model is robust enough to learn the segmentation function while extracting associated features from S1 and S2 locations, then signal segmentation may not be necessary as preliminary data processing.

Eventually, XAI is also gaining increasing attention in the field of Human Activity Recognition (HAR), based both on wearable sensing and smart home environments (Arrotta et al. 2022). Most HAR applications are related to human well-being and fitness through physical activity monitoring, as well as to active and healthy aging by supporting older and impaired subjects in the correct execution of daily activities in their home environment, and/or by detecting abnormal behavioural and locomotion patterns (Khodabandehloo et al. 2021). However, using XAI methods in these applications is currently limited to explain why (and how) simple/complex activities are detected, so the impact of explainability methods on decision making may be limited. Differently, the contribution of XAI to HAR applications in the healthcare domain is much higher, as it provides evidences for the final diagnosis and justifications for successive interventions, with particular reference to gait analysis to detect orthopedic or neurodegenerative disorders, such as Parkinson Disease (PD), and/or fall detection. Such applications are often based on a vast stream of inertial and/or kinematic data, which are in current need of interpretability in order to detect which signal characteristics are effectively used by AI models to take meaningful decisions. To this aim, Filtjens et al. (2021) proposed an interpretable DL framework to detect movements preceding the occurrence of Freeze of Gait (FoG) episodes in PD patients. The framework is based on a CNN to model the reduction of movement prior to a FOG episode from 3-D kinematic joint trajectories of hip, knee, and ankle, respectively, and on LRP explanatory technique to identify the most influential features. The model has been built by using an existing dataset (Spildooren et al. 2010) containing 3-D motion data from 28 PD patients with and without FoG, and 14 healthy subjects. LRP interpretability analysis indicated that the movements perceived as the most relevant by the model are fixed knee extension during the stance phase, reduced peak knee flexion during wing phase, and the fixed ankle dorsiflexion during the wing phase. On the other hand, very little relevance has been observed for these movements in PD patients without FOG and in healthy controls. Therefore, this behaviour suggests that model decisions are made upon meaningful kinematic features, which are actually related to movement reductions during gait.

4.2 Explanation consistency assessment

The evaluation of the level of agreement between explanations generated by different methods is often used by researchers to get some preliminary insights into stability and robustness of AI models. These properties are often used interchangeably as they both refer to the model ability to withstand perturbations introduced in input data, even if a slight difference

among the two concepts exists. Indeed, stability is evaluated with respect to unintentional perturbations that may occur in the real world, such as data noise, while robustness refers to subtle yet intentional changes in input data, namely adversarial attacks.

Given a target high-performing model, if similar explanations are generated by different methods, then the model should provide correct outcomes for the same reasons for equal or similar data instances over the time. On the other hand, similar explanations obtained by multiple models with the same explanatory technique(s) might indicate that common patterns are discovered within data and used to make decisions. As a result, the model should be able to deal with both random changes and adversarial examples without leading to systematic misclassification. However, more specific XAI approaches, such as sensitivity analysis, should be required to draw more accurate conclusions about these model desiderata.

From a practical standpoint, consistency assessment applies to methods providing the same explanation format, generally by matching rankings coming from feature attribution methods or by evaluating the degree of overlap between different decision rule sets. These methods are mainly applied to tabular data, while saliency and attention remain the benchmark methods to explain models learned from time series. Moreover, there are also other motivations limiting explanation consistency assessment to time series. First, saliency methods mainly work for local predictions, so evaluating explanation consistency may not provide any global understanding of models, unless a huge number of data instances is analysed individually. Moreover, the usefulness of comparing maps obtained by different methods can be questionable if end users have difficulties in understanding the high-level content hidden in the input sub-sequences showing the highest relevance. As a result, this section focuses on the most relevant XAI studies targeting tabular data and performing explanation consistency assessment. Research works are also summarised in Table 7.

Thimoteo et al. (2022) compared post-hoc explanations for COVID-19 diagnosis with other glass-box AI approaches. Specifically, they applied SHAP to SVM and RF classifiers trained on lab test data provided by the COVID-19 Data Sharing/BR (over 50k suspected COVID-19 patients), then they compared feature relevance with both LR coefficients and feature importance scores provided by Explainable Boosting Machine (EBM) algorithm (Nori et al. 2019). All global explanations converged in indicating eosinophils and leukocytes, and in general white blood cell-related parameters among the essential features to help diagnose the infection from blood test and pathogen variables.

Alves et al. (2021) proposed a *Decision Tree-based eXplainer* (DTX) and applied it for COVID-19 diagnosis from lab test (i.e., blood, urine, and others) data. This approach produces readable tree structure that provides classification rules to reflect the local behavior of complex models, and it can be considered similar to LIME method using decision trees as surrogate models instead of sparse linear models. In addition, DTX outcomes have been aggregated over many patients for the identification of global patterns, named criteria graphs. DTX rules and criteria graphs have been extracted from a RF classifier trained on the same public COVID-19 dataset used in Leung et al. (2021), then they have been compared with SHAP and LIME explanations at global and local stages, respectively. Results showed a high level of overlap of the proposed method with respect to these well-established XAI techniques, in particular highlighting a correspondence between the 5 largest nodes in the graph and the top-5 features in SHAP ranking.

Okay et al. (2021) developed an interpretable ML approach for early stage diabetes detection from sign and symptom data, which can be easily collected through patient

Table 7 List of XAI studies performing explanation consistency assessment

References	# Cit.	Application	Input Data	AI model(s)	XAI method(s)	Dataset(s)
Thimoteo et al. (2022)	2	COVID-19 diagnosis	EHR	SVM, RF	SHAP	COVID-19 Data Sharing/BR
Alves et al. (2021)	36	COVID-19 diagnosis	EHR	RF	DTX, Criteria graph, LIME, SHAP	COVID-19 dataset[b]
Okay et al. (2021)	1	Diabetes diagnosis	EHR	RF, GBDT	SHAP, LIME	Sylhet Diabetes dataset[2]
Oba et al. (2021)	1	Diabetes diagnosis	EHR	TabNet, XGBoost, LightGBM, CatBoost	SHAP (all), attention (TabNet)	Retrospective study
Elshawi et al. (2019)	132	Hypertension prediction	EHR	RF	feature permutation, PDP, ICE, global surrogate models, LIME, SHAP	Pilot study
Seedat et al. (2020)	1	Voice pathology assessment	Audio features	ExtraTrees	SHAP, Morris sensitivity analysis	Pilot study
Kapcia et al. (2021)	0	Lung cancer life expectancy prediction	EHR	RF	LIME, SHAP	Simulacrum dataset[3]
Duell et al. (2021)	9	Lung cancer mortality prediction	EHR	XGBoost	LIME, SHAP, Anchors	Simulacrum dataset
Moncada-Torres et al. (2021)	43	Breast cancer survival prediction	EHR	XGBoost	SHAP	Retrospective study
Ang et al. (2021)	0	ICU mortality risk prediction	EHR	RF, MLP	SHAP	MIMIC-III
Song et al. (2020)	33	AKI prediction	EHR	GBDT	SHAP	Retrospective study
Duckworth et al. (2021)	5	Hospital readmission prediction	EHR	XGBoost	SHAP	Retrospective study
Tahmassebi et al. (2020)	3	Eye state detection	EEG	XGBoost, DNN	SHAP	Pilot study
Antoniadi et al. (2021)	7	QoL assessment in ALS caregiving	EHR	XGBoost	SHAP	Retrospective study
Ward et al. (2021)	5	Pharmaco-vigilance monitoring	EHR	RF, XGBoost, ExtraTrees	MDA, MDI, LIME, SHAP	Retrospective study

questionnaires. They first trained RF and GBDT models on the Sylhet Diabetes dataset[2] (520 instances, 17 categorical features), then they applied SHAP to compare global explanations. Results indicated that the top-3 features (polyuria, polydipsia, and gender) are shared across the two models using SHAP, with also a high degree of overlap for the top-10 attributes. LIME also provided similar feature rankings among RF and GBDT for the selected local predictions, but this is not enough to assert the convergence of its global explanations. Oba et al. (2021) performed a similar study to analyse explanations related to diabetes aggravation detection from medical records that integrate patient interviews with lab tests and physiological measurements. They used a medical check-up dataset collected by a Japanese health care center to train different tree ensemble models (i.e., XGBoost, LightGBM, and CatBoost) and a TabNet, then they compared SHAP values obtained from the former with attention weights generated by the network. In this case, the top-3 features ranked by SHAP, namely current severity status of diabetes, blood sugar level, and glycated hemoglobin, were the same among all the models, and also equal to those learned by attentive transformers. In addition, results obtained by TabNet pointed out many highly-ranked indicators that can be obtained by non-invasive tests and interviews, which are less burdensome and expensive to patients.

Elshawi et al. (2019) performed an extensive analysis for investigating the outcomes of ML models for hypertension prediction from cardio-respiratory fitness data obtained after a treadmill test. The authors exploited data from $> 23k$ patients and models coming from the FIT project (Sakr et al. 2018), then they selected the best performing RF classifier to compare a variety of XAI methods, including feature permutation, PDP, ICE plots, feature interaction with Friedman's H statistic (Friedman and Popescu 2008), and surrogate models for global explanations, as well as LIME and SHAP for local explanations. Results obtained by this experiment suggested that integrating different global interpretations may allow to generalise the overall conditional distribution modeled by the trained response function, but local interpretations should be preferred for a better understanding of smaller variations in the conditional distribution for specific instances.

Another automated and interpretable diagnostic application has been proposed by Seedat et al. (2020) for voice pathology assessment from smartphone-based microphone recordings. To this aim, they conducted a pilot study to collect and analyse voice recordings obtained from 33 healthy and diseased subjects, then they trained several ML models using a set of handcrafted features extracted through audio signal processing. By choosing ExtraTrees as the best performing model, they compared global explanations obtained through SHAP, Morris sensitivity analysis, and feature permutation. All methods converged in identifying the most relevant features, and they also highlighted 6 clinically used MFCC features as the top relevant ones.

Assessing stability should be imperative for predictive models that are supposed to be deployed for survival analysis. Kapcia et al. (2021) proposed *ExMed*, a tool to enable XAI data analytics and visualisation for clinicians by supporting multiple ML models and feature attribution algorithms, and they tested it using lung cancer life expectancy prediction from EHR as main application. By exploiting the public Simulacrum dataset[3], a cancer dataset provided by the National Cancer Registration and Analysis Service of Public Health England, they trained different ML models and then selected the best RF classifier to compare SHAP values with the global average of LIME local scores obtained over all

---

[2] https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.

[3] https://simulacrum.healthdatainsight.org.uk/

the test patients. This analysis show a very similar impact of almost all 20 patient features used. In particular, cancer grade and M-best (i.e., presence/absence of distant metastatic spread) are the two most relevant features with very close importance scores (also in line with the current clinical knowledge), while there is a disagreement on age attribution. By using the same dataset, Duell et al. (2021) presented an extended comparison of explanations obtained for lung cancer survival prediction. Specifically, they compared SHAP, LIME, and anchors methods applied to a XGBoost model, as well as the feature importance ranking derived from an EBM model. Overall, all methods converge in identifying M-best as the most relevant feature globally, while the ranking for the remaining features differs between SHAP and LIME. For instance SHAP consider N-best (i.e., extent of involvement of regional lymph nodes) as the second most important feature, whereas LIME considers the behaviour of the tumour. Given such discrepancy between LIME and SHAP, the authors studied the scale of their differences by analysing the first 1000 instances of the test set individually to identify priority features, regardless of whether they are shared or not for every instance between the two methods. Results highlight M-best, N-best, and T-best (i.e, size and extent of the primary tumor) as the three most important features, further supporting both consistency and clinical relevance of knowledge representation.

Similarly, Moncada-Torres et al. (2021) compared the performance of a conventional WCPH regression model against three different ML methods, namely Random Survival Forest (Ishwaran et al. 2008), Survival SVM (Pölsterl et al. 2016), and XGBoost, for breast cancer survival prediction from patient, tumor, and treatment-related characteristics. Models have been trained on a dataset built through a retrospective data collection from the Netherlands Cancer Registry between 2005 and 2008, then SHAP has been applied to investigate the differences between a reference WCPH model and the best performing XGBoost model. This comparison resulted in a high degree of overlap between global explanations across the models, while XGBoost reached a considerably better accuracy. Therefore, this increase in model performance may be attributed to XGBoost's ability to model non-linearities and complex interactions among input variables with respect to a simpler semi-parametric approach.

Ang et al. (2021) performed a study to compare the most salient features for ICU mortality risk prediction with different ML models, all of which have been trained on the benchmark MIMIC-III database. Specifically, the authors compared SHAP values obtained from MLP and RF classifiers, as well as Logistic Regression (LR) coefficients and Gini's feature importance scores derived from a decision tree. Obtained results highlighted a high degree of commonality for age, blood urea nitrogen level, and patient dischargement from cardiac surgery as main mortality risk factors in ICU settings.

Song et al. (2020) proposed an alternative perspective for analysing explanation consistency, aimed at cross-site validation of an AKI prediction model from multi-center EHR data. To this aim, they developed a GBDT classifier using a huge amount of input features collected from patients enrolled in the Greater Plain Collaborative (Waitman et al. 2014), a research network including twelve healthcare systems in the US, including demographics, diagnoses, procedures, lab tests, medications, as well as vital signs. To evaluate this approach, a reference site has been chosen for model training and internal validation, while data coming from other 6 clinical sites have been used for external validation. System transportability has been assessed through the definition of a novel metric, the adjusted Maximum Mean Discrepancy (adjMMD), to infer performance deterioration (i.e., drop in AUC) between transported and refitted (i.e., retrained) models at each site. adjMMD is a modified version of the traditional MMD metric, which is widely used in transfer learning for measuring the data distribution shift between source

and target data (Pan et al. 2010). SHAP analysis has also been performed to investigate changes in the marginal effects of top-10 features in predicting moderate-to-severe AKI. This work demonstrated that cross-site performance deterioration is likely and it is generally associated with significant disparities in feature importance, which may be caused by heterogeneity of risk factors across different populations. As a result, a joint analysis of model explanations and performance could be used to estimate the transportability of AI models, which in turn can foster the adaptation process across different clinical settings (e.g., health institutions) and/or patient cohorts.

In addition to cross-site transportability, another fundamental issue related to the deployment of predictive models in clinical settings is the possible occurrence of both data and concept drift over the time. The former is a systematic shift in the underlying distributions of input parameters, while the latter represents a substantial change in the relationship between inputs and target clinical outcomes. Both of them represent threats to model validity; therefore, a periodical monitoring of explanation consistency (together with predictive performance) may be used as model temporal validation, with particular reference to data drift detection. To the best of our knowledge, Duckworth et al. (2021) is the first study suggesting the analysis of explanations over the time to measure data drift, by exploiting COVID-19 as an exemplary case for monitoring hospital readmission risk prediction models. Specifically, this work analyses attendance records of all adults to Southampton General Hospital's ED between 2019 and 2020, including patient descriptors medical history, attendance characteristics, as well as any vital sign measured at the point of triage to perform temporal validation. Specifically, a XGBoost model has been trained only on pre-pandemic attendances, then it has been evaluated in weekly bins over a test period ranging from October 2020 to May 2021 in terms of predictive performance (i.e, AUC) and explanations (i.e., average SHAP values). Results outline a clear drop in model performance starting from March 2020 (i.e., onset of COVID-19), which are also associated with significant changes in the importance of several features. In particular, symptoms data that we now know to be common in infected subjects, such as shortness of breath and chest pain, show a higher relevance starting from COVID-19 spread. Similarly, respiration rate, BT, and SpO_2 present peaks in importance scores starting from March 2020, and also in December 2019 (potentially reflecting the flu season). These outcomes demonstrate that short-time and unpredictable changes in the data distribution, for instance related to a new emergent disease, can negatively affect model outcomes. Therefore, periodical analysis of explanations may act as proxy measure of data drift, with the final aim to timely plan model retraining/updating interventions.

To conclude this section, we report additional innovative health applications characterised by a lack of benchmark AI solutions, which have been also less investigated from a clinical standpoint. Here, explanation consistency assessment is used for preliminary model validation beyond predictive performance analysis. Tahmassebi et al. (2020) compared SHAP explanations for XGBoost and DNN models for eye state detection from multi-channel Electroencephalogram (EEG) signals. The authors conducted a pilot study to build a dataset of approximately 15K EEG recordings, then they trained both classifiers by considering each of the 14 employed channels as a separate input feature in order to investigate the activation of the corresponding brain regions. Results indicated that the top-3 important features are shared across models with slight variations in their ranking; however, the greatest non-linearity of DNN also provided higher-order interactions among features with respect to a XGBoost model with shallow trees as base learners, which in turn led to a higher contribution of low-impact regions for the final prediction.

Antoniadi et al. (2021) proposed the first XAI framework for Quality of Life (QoL) assessment in Amyotrophic Lateral Sclerosis (ALS) patient caregiving from a heterogeneous set of EHR data, which include demographics, usage of health service, questionnaire data from both patients and their caregivers, as well as caregiving duties and patients' clinical attributes. Data have been collected by 90 ALS subjects (and their caregivers) who attended the National ALS/MND multidisciplinary Clinic at Beaumont Hospital, Dublin, then they have been used to compare performance and explanations of a XGBoost model with those of a LASSO-LR reference model, due to the absence of state-of-the-art methods for the target application. In addition, consistency check of the top ranked features by SHAP has been performed across several XGBoost model configurations, each one built with a smaller feature subset chosen through ensemble feature selection. This analysis is geared towards the creation of a human-centered DSS to alert clinicians about worsening QoL conditions in ALS patient caregiving. In addition, the system should be able to reveal and explain the main risk factors within a limited set of data that can be easily and routinely collected by both patients and caregivers, in order to achieve a better usability and assessment.

Ward et al. (2021) proposed a XAI approach for pharmacovigilance monitoring by analysing the impact of patient features related to drug history, comorbidities, and current drug dispensing in developing adverse health conditions. To this aim, they collected data from patients aged > 65yo from the Western Australian Department of Health, by considering acute coronary syndrome as adverse outcome to be predicted with different tree ensemble models (i.e., RF, XGBoost, and ExtraTrees). Then, they compared feature importance scores obtained by different XAI techniques, namely LIME, SHAP, as well as MDI and MDA feature permutation. To reduce cross-model variance, the authors first compared the average feature importance scores across all models, then they performed a per-model analysis of top ranked features across explanatory methods. Overall results indicated that sex and age are ranked highly by all methods, an expected outcome as these are the most important auxiliary inputs in clinical/epidemiological studies. Results also outlined repeating peaks of importance for all XAI methods falling within musculo-skeletal and cardio-vascular drug classes. For what concerns per-model rankings, results indicated that highly ranked comorbidity and drug dispensing features are shared among models under SHAP and MDA analysis, while LIME outcomes are inconsistent, which may be due to a higher variance introduced by averaging LIME local explanations to obtain global patterns. Finally, random integer features have been also included into the models as confounding factors, which should be ranked last as they cannot have any predictive power. Nevertheless, MDI ranked some of them as relevant, unveiling a potential bias in this method as variables with high cardinality may be selected for many tree splits, thus appearing important to the model.

4.3 Explanation quality assessment

Clinical validation and consistency assessment are valuable but indirect ways to estimate the quality and relevance of the proposed explanations from different perspectives. However, there are not well-established and globally recognised metrics or practices that enable a formal assessment and a systematic comparison of methods (Markus et al. 2021). To fill this gap, quantitative metrics have been recently proposed for an initial and objective evaluation. However, their applicability is often limited to the target task or to the format

Table 8 List of XAI studies performing explanation quality assessment

References	# Cit.	Application	Input Data	AI model(s)	XAI method(s)	Dataset(s)
Cheng et al. (2021)	6	Surgical complication risk prediction	EHR	Not found	SHAP	PIC Zeng et al. (2020)
Kumarakulasinghe et al. (2020)	13	Sepsis detection	EHR	RF	LIME	PhysioNet 2019 Reyna et al. (2019)
Barda et al. (2020)	24	Mortality risk prediction	EHR	RF	SHAP	Retrospective study
Penafiel et al. (2020)	4	Stroke prediction	EHR	DS-GD	DS rules, LIME	Retrospective study
Hatwell et al. (2020)	18	CAD	EHR	Ada-WHIPS	Ada-WHIPS rules, Anchors, LORE	9 CAD datasets from UCI ML repository
Zhang et al. (2021)	8	Signal quality assessment	PPG	CNN	SHAP, integrated gradients, guided saliency	Training: ICU dataset Drew et al. (2014); Test: stroke dataset Pereira et al. (2019)
Wickstrøm et al. (2020)	9	AMI detection, SSI detection	Echo, CRP measures	FCN	CAM	ECG200 dataset Dau et al. (2019), SSI dataset Mikalsen et al. (2016)
Slijepcevic et al. (2021)	5	Gait disorder detection	3-D GRF	CNN, SVM, MLP	LRP	GaitRec dataset Horsak et al. (2020)
Hsieh et al. (2016)	24	Seizure detection, fist movement detection	EEG	CNN	Attention	Seizure dataset Schalk et al. (2004), Physionet DB

of explanations generated (e.g., saliency maps, decision rules), while a standardised framework is still missing.

Quantitative evaluations should be followed by human-centered assessment made by healthcare professionals through surveys, feedback, and ratings. Moreover, the collaboration with domain experts is also necessary for the design and development of novel analytics and visualisation tools that effectively incorporate visual/textual explanations of AI models into clinical decision making workflow, also highlighting which interactions with end users are needed to improve the interpretation of machine-based predictions. Given the limited number of research works found in the literature, studies targeting tabular and time series data are grouped together in Table 8.

Cheng et al. (2021) developed *VBridge*, a novel visual analytics tool to address three key challenges related to XAI adoption in healthcare, namely clinicians' unfamiliarity with ML features, lack of contextual information, and need for cohort-level evidence. To this aim, the system first provides a hierarchical display of attribution-based feature explanations, by grouping the most relevant features semantically for a better analysis. In addition, it includes enriched interactions to link relevant features with raw data (both static and time series), reference values from desired patient clusters, with the ultimate goal of providing a complete overview of each patient. As running example for their study, they trained several ML models on the Pediatric Intensive Care (PIC) database (Zeng et al. 2020) for surgical risk prediction in pediatric patients, then they selected SHAP for feature explanation. VBridge has been evaluated through interviews with four expert clinicians, by applying both forward and backward data analysis as separate case studies. In the first case, clinicians inspect data following the data processing flow (i.e., from original EHR, to features, predictions and explanations), which is also similar to the current clinical practice; in the second case, the order is reversed by starting with the output of AI models. Feedback obtained from both case studies demonstrate that visually associating model explanations with patients' situational records can help clinicians to better interpret and use predictions for decision-making.

Kumarakulasinghe et al. (2020) proposed a framework to evaluate clinical relevance and quality of LIME local explanations, by using a RF classifier trained on the PhysioNet/Computing in Cardiology Challenge 2019 dataset (Reyna et al. 2019) for sepsis prediction in ICU patients as case study. The framework enables a multidimensional and semi-quantitative evaluation. It first evaluates the percentage of model explanations accepted by clinicians and the percentage of overlap between LIME and clinicians for the top-K ranked features (both general match and exact ranking match), then it includes a survey adapted from Jian et al. (1998) for a qualitative assessment of trust and reliance on AI outcomes. Additionally, physicians were also asked to report their level of satisfaction with respect to the textual and visual representation of LIME explanations. As a first pilot study, it has been used by 10 clinicians, each of them should have inspected 10 separate local predictions to reach a total of 100 cases. Obtained results indicated that LIME explanations have a very high level of agreement with those provided by physicians, with also trust and reliance scores fairly high.

Barda et al. (2020) conducted a multidisciplinary research study aimed at designing effective user-centered displays of ML explanations for healthcare applications. To this aim, the first trained several ML models for in-hospital mortality risk prediction by using EHR data obtained from ICU pediatric patients admitted at Children's Hospital of Pittsburgh between 2015 and 2016, then they applied SHAP to the best performing RF classifier. SHAP explanations have been analysed through focus group session attended by a total of 21 clinicians, in which different design options and rationales have been proposed

for explanation presentation, such as display formats, risk representation formats, dimensionality, and many others. Therefore, focus group sessions provided critical reviews and highlighted users' preferences, which in turn have been used to define a final version for explanations' content and display. Specifically, the preferred visualisation tool integrates additional information, such as raw feature tables, time series plots, and contextual factors, to help a better understanding of model explanations. In addition, feedback from focus group participants positively support the adoption of model-agnostic, instance-level explanations based on feature relevance (such as those obtained by SHAP) to understand AI predictions in healthcare.

Explanation quality assessment may also be applied to novel XAI algorithms in order to compare them with benchmark techniques. Penafiel et al. (2020) proposed a novel interpretable classifier to make predictions from incomplete/missing EHR data. The proposed approach is able to integrate expert knowledge into the learning process through the a priori definition of rules by medical users, then it exploits the Dempster-Shafer (DS) mathematical framework (Shafer 2016) to reason with data uncertainty, along with a GD optimisation to derive a sub-optimal set of the best performing decision rules. The DS-GD classifier has been trained on EHR data provided by the regional hospital in Okayama, Japan (28k records) to perform stroke prediction, providing superior performance with respect to the state-of-the-art ML models and clinical stroke assessment methods. Moreover, the authors conducted an expert survey by presenting to different neurologists the global rules learned by the model that are associated with an increased stroke risk, including also less contributory and random rules to prevent biased answers. For each rule, clinicians were asked to report whether they consider it true, false, or if there is no correlation. Results obtained by analysing the percentage of clinical adherence confirmed that almost all the rules discovered by the model are also the most accepted statements among experts.

Hatwell et al. (2020) proposed a novel approach to explain the AdaBoost (AB) classification algorithm, a black-box model widely used in the CAD literature. The new algorithm, named Adaptive-Weighted High Importance Path Snippets (Ada-WHIPS), has been designed to provide accurate and highly-interpretable disease detection by extracting simple and logical rules from the AB model. Specifically, Ada-WHIPS assigns unique weights among individual branch nodes of each tree learner within the ensemble, then it performs a simple heuristic search over the weighted nodes to find out a single rule that dominates the final decision. This XAI algorithm has been compared with two state-of-the-art, rule-based explanatory techniques, namely Anchors and LORE, by conducting separate experiments over 9 different CAD datasets (Cleveland, Breast Cancer, Thyroid, etc...) available at the UCI ML repository. Explanation assessment has been conducted through non-parametric hypothesis tests in order to find significant differences in 4 quality metrics, namely efficiency, precision, coverage, and stability. Efficiency just refers to the computational time required to derive an explanation, whereas coverage measures the amount of test instances that could be predicted after looking at a given rule (i.e., rule generalisability). On the other hand, precision represents the fraction of correct predictions that can be made by applying the rule (i.e., rule specificity). Finally, stability is a novel metric introduced by the authors for optimising rule-based explanation algorithms, as it represents a regularised version of precision including a trade-off with coverage to explicitly avoid over-fitting, thus preventing tautological rules that correctly applies only to a very limited set of instances. Ada-WHIPS explanations result in better generalisation capabilities in a two-way comparison with Anchors for all datasets, while significant differences are less consistent in case of a three-way comparison. Moreover, Ada-WHIPS remained competitive in terms of specificity with respect to state-of-the-art, also providing a comparable computational time.

Explanation quality assessment is even more challenging in the time series field, as there are no gold standard metrics applicable to signal processing methods that can offer a global and objective evaluation of saliency/attention mechanisms generally used for explaining DNN models. As a result, most of existing approaches are tied to the availability of extensive expert annotations for the comparison of local saliency with human ground truth over a significant number of data instances. Zhang et al. (2021) proposed novel evaluation metrics for the validation of saliency maps with respect to expert annotations, which can be used for rigorous comparisons between DNN models. Specifically, they include *Congruence*, which measures the proportion of model saliency within the expert annotations, and *Annotation Classification*, which determines the proportion of expert annotations covered by the saliency maps (i.e., the spread of model saliency). To apply these metrics, the authors developed a *ResNet* (He et al. 2016) CNN architecture for photoplethysmography (PPG) signal/image quality assessment (i.e., normal VS. abnormal). The network has been trained using PPG data from an ICU dataset used in Drew et al. (2014) and tested over PPG data from a stroke dataset collected in Pereira et al. (2019). Model saliency has been visualised for the whole test set using different methods, namely DeepSHAP, Integrated Gradients, and guided back-propagation, then it has been compared with manual expert annotations. Obtained results demonstrate that signal-based models act in a more explainable fashion with respect to image-based ones, with guided saliency outperforming the other techniques in both evaluation metrics. The comparison with model classification outcomes also highlights a weak-to-moderate positive correlation between explainability and performance metrics, such as accuracy and specificity, suggesting that the higher the saliency correctly allocated with respect to human ground truth, the higher the model performance.

Wickstrøm et al. (2020) proposed additional metrics to specifically address the lack of uncertainty measures of XAI techniques, with particular reference to saliency methods applied to DNN. To this aim, they first developed a Fully Convolutional Network (FCN) ensemble, in which each CNN is trained independently. Then, uncertainty in the relevance scores has been estimated by taking the standard deviation across importance scores produced by each CNN through CAM, which in turn has been used to threshold the time steps that most models agree on to make explanations more reliable. The proposed model has been tested for AMI detection from echocardiogram signals using the ECG200 dataset (Dau et al. 2019), and for Surgical Site Infection (SSI) detection from longitudinal blood measurements of CRP using a EHR dataset collected in Mikalsen et al. (2016). To evaluate model saliency, *Relevance Accuracy* metric has been computed to determine the amount of actually relevant data points within a sample time series that are recognised by the model, similarly to the congruence metric defined in Zhang et al. (2021). In addition, *Relevance Consistency* metric has been also introduced by the authors to estimate how the ensemble correctly detects the most relevant time points across different network initialisation with respect to a single FCN model. Results demonstrate that the proposed FCN model is accurate in locating relevant time points in both case studies, and it is also more consistent as it indicates mostly the same time steps as relevant for its predictions when trained with different random initialisation.

Slijepcevic et al. (2021) proposed a complementary strategy to evaluate class-specific explanations for gait classification from 3-D ground reaction force (GRF) sensors data, including both quantitative and qualitative analysis. To this aim, they first trained CNN, SVM, and MLP classifiers on the GaitRec dataset (Horsak et al. 2020), a clinical database including bilateral GRF measurements from 132 patients with 3 classes of orthopedic gait disorders and from 62 healthy controls, then LRP technique has been used to explain the most relevant signal characteristics learned by the models. Explanation assessment

has been performed through two interrelated approaches: a quantitative statistical analysis using Statistical Parametric Mapping (SPM) (Pataky 2010), and a qualitative evaluation conducted by clinical experts. In case of binary classification (i.e., normal VS. disordered gait), SPM analysis shows significant differences in signal regions that are also highly relevant according to LRP scores, supporting the premise that models based their predictions primarily on features that are also significantly different between the two classes. In addition, regions with high relevance according to LRP can be largely associated with gait analysis literature and are also plausible from a clinical standpoint according to domain experts' review.

Preliminary approaches for quality assessment of attention mechanisms are quite similar to the evaluation methods proposed for CNN-based saliency maps. To make an example, Hsieh et al. (2016) developed *Locality-Aware eXplainable Convolutional ATtention network* (LAXCAT), a DNN architecture for MTS classification and forecasting consisting of a CNN encoder for feature extraction and a dual attention network for simultaneously learning informative variables and time intervals. The network has been evaluated for both seizure detection and fist movement detection tasks, using multi-channel EEG signals as input data. In the first case, the network has been trained and tested on a seizure dataset collected in Schalk et al. (2004), whereas the benchmark Physionet DB has been used in the second case. Afterwards, the authors defined an *Attention Allocation Measure* (AAM) metric to evaluate the average percentage of attention that is correctly detected by the model across all local predictions. The proposed model outperformed other attention-based networks available in the literature in correctly detecting relevant signal time points, even if the global AAM value obtained in the best case is modest (approximately 25%).

5 Discussion

From the literature analysis, it may be noticed that learning from static or longitudinal data clearly determines the selection of candidate models, and hence it impacts on the application of XAI methods. Static data have no time dimension by definition, and hence tree ensembles, such as RF and XGBoost, are generally the best-performing models to learn complex non-linearities between target and inputs. Shallow MLP networks, ad-hoc networks (e.g., TabNet), as well as data mining approaches (e.g., fuzzy inference, evolutionary algorithms) are used only in limited cases. Although less complex than DNN, tree ensembles are still black-boxes and demand for explainability. Post-hoc methods are generally preferred to explain these models, as they do not impose any trade-off between interpretability and predictive performance. In terms of explanation output, feature attribution methods overcome decision rule-based algorithms like *Anchors* and *InTrees*, with LIME and SHAP being, by far, the most comprehensive and dominant methods across the literature to detect and visualise feature importance. In particular, SHAP carries several advantages with respect to LIME that makes it a reference explanatory technique:

- *Solid theoretical foundations:* SHAP is backed by the game theoretically optimal Shapley values. In particular, SHAP satisfies three desirable properties of explanations: local accuracy, missingness, and consistency (Lundberg and Lee 2017), which contribute to generate trustworthy explanations. Differently, LIME is a more heuristic approach based on the assumption of a sparse linear model as local surrogate model, which may

make explanations unfaithful if the underlying model is highly non-linear even in the locality of the target instance.

- *Scope of explanations:* SHAP provides both local and global explanations, while LIME is well suited only for local interpretability. As already mentioned in Section 2.1, simply averaging LIME across multiple local predictions is subject to high variance and may negatively impact on explanation stability. On the other hand, LIME-SP just provides a set of representative explanations for each class, which is definitely not enough to get insights into the inner working of a model as a whole.
- *Model-agnosticism:* SHAP provide different implementations to deal with any kind of black-box model. However, it is important to outline that only *TreeSHAP* algorithm is used for tree-based models as it provides fast and exact results (Lundberg et al. 2020).
- *Feature interactions:* SHAP can also break down input-target relationships into main effects and interaction effects, making possible to investigate how and to what extent interactions between each variable pair contribute to model predictions, both at local and global stages.

In the time series domain, RNN and CNN currently represent the best performing solutions to learn from raw data without the need of heavy signal pre-processing methods. On the other hand, their complexity combined with the non intuitive nature and lower interpretability of time series data makes understanding these models even more complicated. By looking at the results of our literature survey, attention is the reference ante-hoc method to integrate explainability at variable and time levels when developing RNN, while gradient-based methods are usually applied to generate post-hoc explanations for CNN, with Grad-CAM and LRP being the most influential algorithms. Example-based methods involving prototype learning and perturbation-based methods (e.g., occlusion) can also be found, as well as CNN equipped with attention modules in place of post-hoc methods. Attention and saliency methods are generally used only for local interpretability, and they both provide heatmaps to be overlapped onto the input signal to highlight which specific components get the most attention from the model while classification is performed. However, they do not specify how the information contained in the most relevant signal sub-sequences is used to make a given prediction.

Despite their increasing adoption to bring explainability to AI models developed to accomplish medical tasks, an extensive and multi-dimensional assessment of explanations encompassing clinical relevance, consistency, and quality is generally missing for most health applications, and especially in terms of quality. The main goal of quality evaluation is twofold: 1) enabling a formal comparison among suitable methods in order to choose which one should be preferred; 2) determining if the offered form of explainability meets end users' needs. These objectives should not be in contrast, indeed they require integration to define the solution providing the best evidence of success.

Even if some preliminary evaluation processes conducted in the revised works provide promising results, the reference XAI methods highlighted in our survey still carry out several limitations that may undermine their safety and reliability in real scenarios. As a result, employing them to understand models deployed in high-stake health applications can be risky. In the next section, we outline some of the main limitations and current challenges related to the application of XAI methods in healthcare, also highlighting potential yet under-developed research directions to address these issues. Specifically, we begin with a focus on the challenges introduced by the evaluation process of explanations from both user-centered and user-independent perspectives, then we outline some technical/methodological flaws and potential advancements.

5.1 XAI Challenges in Healthcare

5.1.1 Clinical Validation

Evidence-based assessment is still a prevailing paradigm in healthcare with respect to data-driven knowledge, and hence it plays a central role in the decision making process. Therefore, clinical validation is a basic requirement for the acceptance and deployment of AI systems in real-life medical applications. However, clinical comparison is also valuable for other reasons. First, it can be used to mitigate *AI bias*. This can be achieved by detecting systematic deviations in the generated explanations with respect to the expected behaviour, such as counter-intuitive and clinically misaligned input-target relationships. However, this aspect is often missing or underestimated, as only the top input-target relationships that are in line with the current clinical knowledge are generally reported. Detecting and investigating such possible model failures is necessary as well, and even more before AI systems go through certification and approval processes as it may enable saving time, resources, and user efforts. In addition, clinical comparison may also be used for *knowledge discovery*, by highlighting possible emergent biomarkers/risk factors that might not be considered (or less used) by domain experts, thus potentially enabling clinical utility of AI models. In turn, medical AI systems may benefit of knowledge discovery for a better learning through iterative model design and training, involving both model selection and data collection/curation stages.

5.1.2 Stability and robustness

AI models should not be only clinically sound, but they should also be stable and robust. Models should generate similar predictions and explanations for equal or close data points over the time. Evaluate consistency among explanations provided by multiple methods at global/local stage is a straightforward and inexpensive approach to get insights into model stability and robustness, but results must be handled cautiously. Empirical and theoretical analysis demonstrated that the majority of popular feature importance and counterfactual explanation methods are non-robust (Mishra et al. 2021). In particular, most works focused on XAI methods that are specific to DNN models. Kindermans et al. (2021) demonstrated that simply adding a constant shift to input data causes several gradient-based methods to attribute wrong saliency. Similar results have been also shown by Alvarez-Melis and Jaakkola (2018) for model-agnostic methods such as LIME and SHAP, and the authors also claimed such methods are even more non-robust than gradient-based ones. As a result, comparing the level of agreement of explainability brought by most of these methods cannot exhaustively assess the stability and robustness in AI models, whereas sensitivity analysis is currently the most targeted analytical approach to better expose model vulnerabilities. For example, Siddiqui et al. (2019) applied perturbation on the most salient part of the input sequences leading to a huge drop in classification, thus highlighting model sensitivity to noises. Sensitivity analysis conducted in Hartl et al. (2020) highlighted the most salient features as the same with highest potential to cause misclassification, thus making the model easily susceptible to adversarial attacks.

However, these aspects may be only partially investigated but not definitely solved through XAI, and they still require novel evaluation metrics and practices that will guarantee the "right" behavior of a model. Extensive training is currently the main solution to

mitigate sensitivity against noise and adversarials, while extensive and external validation enables the detection of several issues. Song et al. (2020) and Duckworth et al. (2021) are two pioneer research studies reviewed in this survey that proposed a joint analysis of performance degradation and explanation consistency to reveal model transportability and temporal validity concerns, respectively. Moreover, an in-depth quality evaluation of collected data is equally important as it enables understanding the limitations of the resulting model. However, such evaluations can be costly and consuming, and hence often impractical. In addition, they might never be enough to guarantee robustness and stability, as random perturbations leading to wrong decisions may occur anyway. Therefore, a potential countermeasure to face with input perturbations could be evaluating how different is a given prediction with respect to other data points in the validation set, for instance by detecting samples that are far from the input data distribution (i.e., outliers). As a result, every prediction should be associated with a score indicating how confident is the model in its decision, leaving end users the autonomy to choose whether to trust such prediction or not.

5.1.3 Quality assessment

When using XAI to build trust and confidence, evaluating the quality of explanations is key. Quantitative proxy metrics are necessary for an initial and objective assessment as well as a formal comparison of explanation methods to identify the best-quality option regardless of end users' needs and preferences. However, this research line is currently under development, and there are no standard evaluation frameworks available in the community yet. This might be due to several factors, such as the different nature of explanations, the different models, as well as the different input data types. As a result, most of existing quantitative evaluations are based on metrics that measure performance degradation through input perturbation analysis (Alvarez Melis and Jaakkola 2018). Other quantitative evaluations often perform a comparison with expert annotations. This approach has been originally proposed for image-based applications to compare saliency maps with expert annotations (Mohseni et al. 2018), then it has been adapted for the time series domain, as it can be noticed in Section 4.3. However, one major flaw is related to the introduction of human bias in such evaluation process, which requires large-scale studies to be canceled or at least reduced. Moreover, the unintuitive nature of some time series (e.g., inertial measurements) poses additional challenges as even domain experts cannot provide an exact and reliable ground truth to compare saliency methods with.

Although quantitative metrics may enable a formal evaluation, human-centered assessment is a fundamental yet missing step for tuning explanations in order they can effectively and successfully reach target users. In addition, user studies can also integrate both qualitative assessment related for instance to usefulness, satisfaction, and trust on explanations, as well as quantitative measurements of human-machine task performance. As a result, quantitative evaluation framework should be complemented with user-centered assessment before employing AI systems in real-life. As a first note, the intended audience of XAI strongly impacts on the scope and purpose of explanations, as different users imply different responsibilities, and hence different needs and preferences. AI experts as well as non-expert users added with responsibility in case of system failure, may be probably more interested in global explanations as they provide an overview of model behaviour that might help to decide whether using the model is reliable or too risky. On the other

hand, consumers of the model (i.e., clinicians as well as patients) will rather look for local explanations related to a single prediction (or group of predictions) that affect them, while they cannot take any advantage from understanding the model as a whole. Despite the huge differences that can be shown by different pools of humans, there is a lack of user studies exploring the needs of clinicians (and also patients) in the XAI literature. Non-expert users generally have difficulties due to the gap between algorithmic outputs and human-consumable explanations (Antoniadi et al. 2021). Therefore, pushing interdisciplinary collaboration moving forward is necessary not only to explains AI systems in a more transparent and interpretable manner, but also to determine how end users would like to receive information and share meaning with each other.

5.1.4 Interactivity

Human-AI interaction in healthcare is an open challenge especially in terms of trust, usability, and acceptability, both from physicians and patients, and this can be closely related to explainability. A recent statistical study with more than 1300 physicians demonstrated that 88% positively evaluated AI systems accompanied by model-agnostic explanations (Diprose et al. 2020) in terms of trust and system understanding. Improving user understanding of the decision making process is even more important in remote healthcare, as AI algorithms embedded in e-health and m-health apps often require a direct interaction with non-expert users (i.e., patients). In a recent study conducted in Su et al. (2020), 40 popular commercial m-health apps coming from different domains (e.g., fitness, mental health, nutrition) have been reviewed by 400 consumers, and obtained results indicate that most of the involved users were not able to understand the motivation of AI outcomes due to the lack of information and explanations.

However, simply associating model outcomes with the corresponding explanations is not sufficient to build trust and confidence in a system. This should include specific interactions by design to correctly provide system outputs to end users. According to Wang et al. (2019), designing effective human-centered XAI strategies should first bridge algorithm-generated explanations with user explanation goals and expected reasoning methods. An in-depth analysis of these relationships is essential to provide effective data structures and visualisations, as well as to mitigate common cognitive biases in the human decision making leading diagnostic errors in medicine. Then, more specific approaches are necessary to fill the gap between experimental implementations and solutions acceptable by physicians in daily practice. For instance, Kovalchuk et al. (2022) highlights the need of explicit references to existing clinical norms (e.g., scales and recommendations) and domain-specific explanation of the results, in order to improve decision making while keeping the problem-specific reasoning interface similar to well-known clinical tools.

Moreover, a major finding highlighted both in Wang et al. (2019) and Kovalchuk et al. (2022) is the need of integrating different XAI interactions and facilities to improve human interpretability, either by the joint design with healthcare professionals or through successive refinements following human feedback. Some examples may include: linking explanations with raw and/or processed data to address clinicians' difficulties with ML features; clustering patient data to provide cohort-level evidence; adding context information to generate a more comprehensive overview of a patient. Moreover, clinicians often outlined the need of WHAT-IF analysis tools to probe AI systems when their output is not as expected, such as in case of wrong or

"suspicious" decisions. This is generally achieved through the investigation of *hypothetical* questions (i.e., *how can specific input updates change model outcomes?*) and/or *counterfactual* questions (i.e., *what minimum change should I apply to a given input to drive the model towards a desired different prediction?*) Wexler et al. (2019). In this way, clinicians can edit specific data instances and see how they changes or influences model outcomes. However, WHAT-IF analysis is generally missing in most XAI visual analytics tools proposed in the literature. Indeed, among the reviewed works, only *VBridge*Cheng et al. (2021) and *RetainVis*Kwon et al. (2018) systems include this functionality to assess how and to what extent abnormal feature values contribute to a given prediction. Eventually, end users also require multiple explanation methods and/or format to be included in the same system. Outlining how different explanations can be best combined in a user interface and how these combined AI systems should be then evaluated is another open research problem.

5.1.5 Human-consumable explanations for time series data

By looking at XAI methods, and especially in the time series field, one question arises: *are explanations really explaining?*. Attention and saliency methods just highlight the most relevant input components for a given data instance, without providing any knowledge about how they are used by the inner model reasoning to derive the final decision. Whereas such explanations can be useful for image and text data as humans are generally able to understand their content at first sight, the situation is more challenging for time series because expert knowledge might be needed. Prototypes learning-based approaches are even harder to be put in practice, as comparing target and exemplary sequences might result counter-intuitive. However, saliency can still be considered a naïve approach to provide explanations and increase the trust in the system in some cases, as a user or a developer can rely more on a system if he/she knows that the model gives most attention on well-known parts of the input signal. A practical example can be the usage of saliency methods to highlight relevant ECG patterns and/or waveforms for detecting heart anomalies, such as in Ivaturi et al. (2021). This approach does not differ so much from the current clinical practice based on the visual inspection of ECG traces, hence it can effectively assist cardiologists in their decision making. Conversely, the trust brought by these methods can be questioned when using unintuitive and less inherently interpretable time series, which even domain expert may struggle to understand. A typical example in healthcare is the analysis of human gait from inertial measurements (e.g., accelerometer, gyroscope, GRF sensor, etc...) to detect neurodegenerative diseases (Filtjens et al. 2021). In such cases, some data mining methods should be applied to automatically extract meaningful information hidden within the highlighted signal components (Rojat et al. 2021). Moreover, novel algorithms may be needed for generating human-friendly explanations from time series. *Concepts* have gained an increasing attention as a new class of XAI methods, which is able to deduce high-level and human-understandable contents (i.e., concepts) from lower-level sensor data and to use them as meta-explanations. Some recent concept-based methods include *Testing with Concept Activation Vectors* (T-CAV) (Kim et al. 2018), *Automatic Concept-based Explanations* (ACE) (Ghorbani et al. 2019), *Causal Concept Effect* (CaCE) (Goyal et al. 2019), as well as *ConceptSHAP*

(Yeh et al. 2020). By designing explanations as concepts, these methods allow integrating standard saliency-based local explanations with additional meta information on the most contributing factors.

5.1.6 Understanding DNN models

Beyond the underlying limitations of saliency methods highlighted in Section 5.1.5, with particular reference to the time series field, the adoption of DNN for decision making in healthcare is further limited by the lack of reference methods to provide global explanations and to investigate the inner functioning of such models. As far as global explanations is concerned, saliency methods may provide useful information only if structured and meaningful data are passed as input (e.g. handcrafted features, clinical attributes), but this condition rarely occur while developing DNN models. On the other hand, explaining features learned in the latent space might be useless as they often do not have a human-understandable meaning. Very few research works extended saliency methods to produce global explanations, but most of them have been originally designed for image data. For instance, Oviedo et al. (2019) averaged CAM saliency maps across all training examples to detect the main discriminating features for each class. More formal algorithmic approaches can be found as well. Activation Maximization method is probably the most recognised explanation method for generating human-interpretable representations of different intermediate filters within a deep CNN architecture (Simonyan et al. 2013). However, the above approaches are less suitable for explaining signal-based methods as time series data are less intuitive than images, so their output is more difficult to understand by end users. Other existing global explanation methods, such as Tree Regularization (Wu 2018) or Network Dissection (Bau et al. 2017), are also unfit for signal processing. Indeed, approximating DNN with decision trees may become hard to interpret with high dimensional data. On the other hand, the Network Dissection algorithm works by quantifying the network response to human understandable concepts contained in image data, such as colors, objects, and scenes. As a result, it requires densely labeled time series data to ideally match concepts with sensor data samples, which is often inapplicable.

The literature contribution to global explanation methods for DNN models learned from time series data is even more limited. Siddiqui et al. (2019) clustered CNN filters according to their activation pattern, based on the assumption that filters with similar activation patterns basically detect the same content. Cho et al. (2020) proposed an alternative clustering approach by grouping input sub-sequences that activate the same nodes, and associating each cluster to a representative example. These methods represent preliminary works to achieve some global insights, but much more effort is still required to generate human-consumable explanations, and especially to explain the latent space of CNN.

In addition, no methods seem to exist to probe DNN in case of wrong or "strange" explanations. Indeed, benchmark methods such as Grad-CAM and LRP are based on heuristic to produce locally interpretable information, but the internal process and computations cannot be discerned. For example, changing one (or more weights) in the hidden layers does not easily reveal any useful clue about model functioning. This clearly limits the interactions with DSS based on DNN models as clinicians cannot investigate how and why a certain outcome is reached. Trust and confidence in such systems is negatively affected as well.

5.1.7 Causability

Human assessment of post-hoc, model-agnostic methods often criticised that explanations mainly show correlations (Bruckert et al. 2020), while they cannot fulfil expectations when it comes to enhance medical decisions on tasks where there is a need to understand causal relationships. Therefore, a consistent mapping of explainability with causability is crucial to design effective interactions for physicians and, consequently, successful user interfaces Holzinger and Muller (2021). To this aim, Holzinger et al. (2020) proposed a novel evaluation framework, namely System Causability Scale (SCS), which allows to quantify the level reached by an explanation in providing a cause-effect understanding with effectiveness, efficiency, and satisfaction by end users. However, simply measuring the degree of causality of explanation statements may not necessarily facilitate medical decisions, such as for detecting the impact of different health interventions on model outcomes. The next level of data-driven decisions require understanding how the systems react to external stimulation and why the change occurs. Answering to these questions through XAI may be tricky, as changing the statistical distribution of affected input variables may invalidate the basic assumptions on which the models were built, making results unreliable. To this aim, *Causal AI* is emerging as a novel AI paradigm combining traditional ML algorithms with the principles of causal reasoning to build causal inference systems (Schölkopf et al. 2021). As a result, we expect causal AI to bridge the gap between predictions and decision-making, with the potential to enable AI researchers and clinicians to jointly design and simulate an intervention and infer causality by relying on already available data. This might also bring a huge innovation in evidence-based medicine, as it may support and complement clinical randomised controlled trials for measuring the impact of health interventions over the long term.

6 Conclusions

Tabular and time series data are widely used in clinical and remote AI-empowered health applications, enabling the accurate detection of a variety of diseases and conditions in a more responsive, less-invasive, and low-cost fashion. However, they are not commonly analysed from XAI perspective with respect to biomedical images and medical text records. Nevertheless, XAI techniques are required to bring explainability to complex models learned from this data. An extensive evaluation of generated explanations including clinical relevance, consistency, and quality, is also a necessary (but often missing) step to ensure XAI methods effectively improve human understanding and confidence in AI decision making, thus representing a natural step towards *Trustworthy AI*. However, it should be acknowledged that the list of potential XAI challenges is much broader than only building trustworthiness. For instance, XAI may contribute to ensure the adherence to ethical principles and values. Recently, Muller et al. (2021) elaborated ten practical principles from the most recent and relevant works on the ethical application of AI, with particular reference to the medical field, addressing several issues such as clear identification of decisions, actions, and communications performed by an AI agent, accountability, lawfulness, and compliance with the state-of-the-art theories and practices. Explainability can play a central role in respecting most of these principles. As already discussed, explanations can be compared with medical knowledge for clinical validation, as well as for revealing

hidden AI bias. Moreover, XAI can contribute to ensure the fairness of AI algorithms (commandment 9 in the proposed checklist) by verifying that machine decisions are made without any discrimination based on patient characteristics and/or groups, such as gender and ethnicity (Hardt et al. 2016).

XAI may also help make the system compliant with the patient-centered care paradigm. As patients are nowadays considered active partners in the care provisioning process, they have the right to choose and control their medical treatments and recommendations (Bjerring and Busch 2021). Therefore, XAI can boost patient acceptance and comfort in undergoing AI-empowered medical practices as long as clinicians are able to comprehend and report the reasons behind a given decision. From the medical perspective, XAI enables the definition of a shared meaning of the decision process, so that clinicians can support their decisions. Moreover, XAI may also facilitate the resolution of disagreements between human and machine-based decisions. However, a clear definition of liability in case of wrong decisions is a core touch-point that is in current need of strong clarification by legislative bodies.

By combining the above ethical, social, and medical implications with the technical challenges discussed in this survey, it can be easily noticed that we are far from reaching end-to-end XAI systems ready for large-scale deployment with minimal human supervision. Methodological improvements, user-centered studies, and also clear and full regulations are still necessary for such systems to be accepted and used in the medical practice. Focusing only on the technical aspects, some may argue that explainable modelling is the unique solution for high-stake domains such as healthcare. However, the development of novel and inherently interpretable AI algorithms often imposes a trade-off in terms of predictive performance, which can raise further criticism as it may be exchanging better medical outcomes with an increased transparency. On the other hand, post-hoc explanations can prioritise model accuracy in healthcare, but they are generally approximations of its inner reasoning. As a result, they cannot be completely trustable by definition.

Before XAI will be able to reach a robust way to handle interpretability, AI predictions will inevitably carry some risks and failures, as for any new technology, treatment, and drug we aim to introduce in healthcare. Until that moment, XAI should be considered as a complementary support and not a replacement of standard medical practice, and domain expert supervision is still necessary to make the final decision. According to this perspective, XAI and evidence-based assessment can safely coexist and improve medical outcomes. However, their coexistence needs a careful orchestration to avoid a constant conflict between innovative and gold standard approaches, which can undermine the effective usage of medical AI systems with negative consequences on both patients and clinicians.

Appendix A: Acronyms

See Table 9

Table 9 List of acronyms sorted in alphabetic order

Acronym	Definition
AAM	Attention Allocation Measure
AB	AdaBoost
ACE	Automatic Concept-based Explanations
ACSRS	Asymptomatic Carotid Stenosis and Risk of Stroke
AD	Alzheimer Disease
Ada-WHIPS	Adaptive-Weighted High Importance Path Snippets
adjMMD	Adjusted Maximum Mean Discrepancy
AF	Atrial Fibrillation
AI	Artificial Intelligence
AKI	Acute Kidney Injury
ALE	Accumulated Local Effects
ALS	Amyotrophic Lateral Sclerosis
AMI	Acute Myocardial Infarction
ANN	Artificial Neural Network
AUC	Area Under the ROC Curve
BGC	Blood Glucose Concentration
Bi-LSTM	Bidirectional Long Short-Term Memory
BP	Blood Pressure
BT	Body Temperature
CaCE	Causal Concept Effect
CAD	Computer-aided Diagnosis
CAM	Class Activation Mapping
CHO	Carbohydrates
CNN	Convolutional Neural Network
COPD	Chronic Obstructive Pulmonary Disease
CRP	C-Reactive Protein
CTA-TCN	Channel Temporal Attention-Temporal Convolutional Network
CTG	Carditocogram
DeepAISE	Deep Artificial Intelligence Sepsis Expert
DL	Deep Learning
DNN	Deep Neural Network
DS	Dempster-Shafer
DSS	Decision Support System
DTX	Decision Tree eXplainer
EBM	Explainable Boosting Machine
ECG	Electrocardiogram
ED	Emergency Department
EDA	Electrodermal Activity
EEG	Electroencephalogram
EHR	Electronic Health Record
END	Early Neurological Deterioration
FCN	Fully Convolutional Network
FHS	Fetal Health Status
FiLM	Feature-wise Linear Modulation

Table 9 (continued)

Acronym	Definition
FoG	Freeze of Gait
FSL	Few-Shot Learning
GBDT	Gradient Boosting Decision Tree
GD	Gradient Descent
GE	Grammar Evolution
GNN	Graph Neural Network
Grad-CAM	Gradient-weighted Class Activation Mapping
GRF	Ground Reaction Force
GRU	Gated Recurrent Unit
HAR	Human Activity Recognition
HF	Heart Failure
HiTANet	Hierarchical Time-aware Attention Network
HR	Heart Rate
HRV	Heart Rate Variability
ICE	Individual Conditional Expectation
ICU	Intensive Care Unit
IML	Interpretable Machine Learning
k-NN	k-Nearest Neighbors
LASSO	Least Absolute Shrinkage and Selection Operator
LAXCAT	Locality-Aware eXplainable Convolutional ATtention network
LC-LSTM-DAE	Locally Connected Long Short-Term Memory Denoising Auto-Encoder
LGP	Linear Genetic Programming
LightGBM	Light Gradient Boosting Machine
LIME	Local Interpretable Model-agnostic Explanations
LIME-SP	LIME Submodular Pick
LONN	Logic-Operator Neuronal Network
LORE	Local Rule-Based Explanations of black-box decision systems
LRP	Layer-wise Relevant Propagation
LSTM	Long Short-Term Memory
LT	Liver Transplantation
MAC-LSTM	Multi-Attention Collaborative Long Short-Term Memory
MCI	Mild Cognitive Impairment
MDA	Mean Decrease in Accuracy
MDI	Mean Decrease in Impurity
MDW	Medicaid Data Warehouse
MESA	Multi-Ethnic Study of Atherosclerosis
MFCC	Mel-Frequency Cepstral Coefficient
MIMIC	Medical Information Mart for Intensive Care
ML	Machine Learning
MLP	Multi-Layer Perceptron
MMD	Maximum Mean Discrepancy
MTS	Multivariate Time Series
NB	Naïve Bayes
NCD	Non-Communicable Disease

Table 9 (continued)

Acronym	Definition
NLP	Natural Language Processing
PCG	Phonocardiogram
PD	Parkinson Disease
PDP	Partial Dependence Plot
PIC	Pediatric Intensive Care
PL	Prototype Learning
PPG	Photoplethysmogram
PRN	Partial Response Network
QoL	Quality of Life
RETAIN	REverse Time AttentIoN
RF	Random Forest
RNN	Recurrent Neural Network
ROSMAP	Religious Orders Study and Rush Memory and Aging Project
SAX	Symbolic Aggregate ApproXimation
SCS	System Causability Scale
SHAP	Shapley Additive Explanations
SMILE	Systems Metabolomics using Interpretable Learning and Evolution
SPM	Statistical Parametric Mapping
SpO_2	Pulse Oximetry
SSI	Surgical Site Infection
ST	Skin Temperature
SVM	Support Vector Machine
T-CAV	Testing with Concept Activation Vectors
T-LSTM	Time-aware Long Short-Term Memory
t-SNE	t-distributed Stochastic Neighbor Embedding
TCN	Temporal Convolutional Network
TCoN	Time-aware and Co-occurrence-aware deep learning Network
TF-IDF	Term Frequency Inverse Document Frequency
TITV	Time-Invariant and Time-Variant
WCPH	Weibull Cox Proportional Hazards
XAI	Explainable Artificial Intelligence
XGBoost	eXtreme Gradient Boosting

Appendix B: Preliminary works

See Tables 10 and 11.

Table 10 Preliminary studies applying XAI to tabular data

References	# Cit.	Application	Input Data	AI model(s)	XAI method(s)	Dataset(s)
Davagdorj et al. (2021)	5	NCD prediction	EHR	DNN	SHAP	NHANES dataset[a]
Prentzas et al. (2019)	25	Stroke prediction	EHR	RF	InTrees, symbolic AI	ACSRS dataset Nicolaides et al. (2010)
Lisboa et al. (2020)	2	Cervical cancer diagnosis	EHR	MLP	PRN	Retrospective study
Leung et al. (2021)	11	COVID-19 diagnosis	EHR	RF, FSL–ANN	SHAP	COVID-19 dataset[b]
Cavaliere et al. (2020)	10	PD diagnosis	Handmade writings	Evolutionary algorithms	GE	HandPD dataset Pereira et al. (2015)
Ibrahim et al. (2020)	15	AMI prediction	ECG features	XGBoost	SHAP	ECG-ViEW II DB Sadhukhan et al. (2018)
Krishnakumar and Abdou (2020)	1	Arrhythmia detection	ECG features	stacked NB-k-NN-RF ensemble	SHAP	Arrhythmia dataset Guvenir et al. (1997)
Moreno-Sanchez (2020)	6	Survival prediction in HF patients	EHR	XGBoost	SHAP	HF dataset Ahmad et al. (2017)
Chmiel et al. (2021)	0	Hospital reattendance prediction	EHR	XGBoost	SHAP	Retrospective study
Chen et al. (2020)	15	Hospital readmission prediction in HF patients	EHR	MLP	Attention	Retrospective study
El-Bouri et al. (2020)	7	Hospital admission location prediction	EHR	ANN	Saliency maps	Retrospective study
Ochoa et al. (2021)	7	Therapy recommendation	EHR	LONN	Fuzzy rules	Pakistan database[c]
Costa et al. (2021)	1	Pain relief evolution assessment	EHR	RF-XGBoost ensemble	SHAP	Retrospective study
Dwivedi et al. (2021)	0	FHS detection	CTG/HRV features	LightGBM	feature permutation, SHAP	Kaggle Fetal Health dataset[d]
Zhai et al. (2020)	32	Sleep stage classification	Actigraphy, HR/HRV features	RF	SHAP	MESA sleep dataset Zhang et al. (2018)
Gullapalli et al. (2021)	1	Clinical opioid use tracking	HR/HRV, EDA, ST, 3-D Acc features	CTA-TCN	Attention	Pilot study

[a]https://www.cdc.gov/nchs/nhanes
[b]https://www.kaggle.com/einsteindata4u/covid19
[c]https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records
[d]https://www.kaggle.com/andrewmvd/fetal-health-classification

Table 11 Preliminary studies applying XAI to time series data

References	# Cit.	Application	Input Data	AI model(s)	XAI method(s)	Dataset(s)
Sun et al. (2021)	0	Mortality prediction	Longitudinal EHR	TCoN	Attention	MIMIC-III Johnson et al. (2016)
Thorsen-Meyer et al. (2020)	11	Mortality prediction	Longitudinal EHR	LSTM	SHAP	Retrospective study
Shamout et al. (2019)	39	Early detection of in-hospital deterioration	HR, HRV, BT, BP, SpO_2	Bi-LSTM ensemble	Attention	Retrospective study
Lauritsen et al. (2020)	124	Acute critical illness prediction	Longitudinal EHR	TCN	LRP	Retrospective study
Jiang et al. (2021)	0	Hospital readmission prediction	Longitudinal EHR	T-LSTM	Attention	MIMIC-III, New York State MDW
Luo et al. (2020)	45	Disease diagnosis	Longitudinal EHR	HiTANet	Attention	UCI COPD, HF, CKD datasets
Maweu et al. (2021)	16	Arrhythmia detection	ECG	CNN	Grad-CAM	MIT-BIH Arrhythmia database subset Plawiak (2017)
Ma et al. (2020)	0	AMI detection	MTS	CNN	PL	MIMIC-III
Bois et al. (2020)	2	BGC forecasting	longitudinal BGC, CHO intake, and insulin data	RETAIN Choi et al. (2016)	Attention	IDIAB dataset
Zhang et al. (2020)	36	Epileptic seizure detection	EEG	CNN	Attention	TUH-EEG database Obeid and Picone (2016)
Li and Sano (2020)	24	Stress detection	EDA, ST, 3-D Acc	LC-LSTM-DAE	Attention	Pilot study
He et al. (2021)	1	BP forecasting	Longitudinal EHR	MAC-LSTM	Attention	Retrospective study

Declarations

Competing interests The authors have no relevant financial or non-financial interests to disclose.

References

Ahmad T, Munir A, Bhatti SH, Aftab M, Raza MA (2017) Survival analysis of heart failure patients: a case study. PLoS ONE 12(7):0181001

Alvarez Melis D, Jaakkola T (2018) Towards robust interpretability with self-explaining neural networks. In: Advances in Neural Information Processing Systems, vol 31

Alvarez-Melis D, Jaakkola TS (2018) On the robustness of interpretability methods. arXiv preprint arXiv:1806.08049

Alves MA, Castro GZ, Oliveira BAS, Ferreira LA, Ramírez JA, Silva R, Guimarães FG (2021) Explaining machine learning based diagnosis of covid-19 from routine blood tests with decision trees and criteria graphs. Comput Biol Med 132:104335

Amann J, Blasimme A, Vayena E, Frey D, Madai VI (2020) Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 20(1):1–9

Ang ET, Nambiar M, Soh YS, Tan VY (2021) An interpretable intensive care unit mortality risk calculator. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp 4152–4158. IEEE

Antoniadi AM, Galvin M, Heverin M, Hardiman O, Mooney C (2021) Prediction of caregiver quality of life in amyotrophic lateral sclerosis using explainable machine learning. Sci Rep 11(1):1–13

Antoniadi AM, Du Y, Guendouz Y, Wei L, Mazo C, Becker BA, Mooney C (2021) Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: a systematic review. Appl Sci 11(11):5088

Apley DW, Zhu J (2020) Visualizing the effects of predictor variables in black box supervised learning models. J R Stat Soc Ser B 82(4):1059–1086

Arık SO, Pfister T (2021) Tabnet: attentive interpretable tabular learning. In: AAAI, vol 35, pp 6679–6687

Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, García S, Gil-López S, Molina D, Benjamins R et al (2020) Explainable artificial intelligence (xai): concepts, taxonomies, opportunities and challenges toward responsible ai. Inf Fusion 58:82–115

Arrotta L, Civitarese G, Bettini C (2022) Dexar: deep explainable sensor-based activity recognition in smart-home environments. Proc ACM Interact Mob Wear Ubiquitous Technol 6(1):1–30

Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE 10(7):0130140

Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473

Barakat NH, Bradley AP (2007) Rule extraction from support vector machines: a sequential covering approach. IEEE Trans Knowl Data Eng 19(6):729–741

Barda AJ, Horvat CM, Hochheiser H (2020) A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare. BMC Med Inform Decis Mak 20(1):1–16

Bau D, Zhou B, Khosla A, Oliva A, Torralba A (2017) Network dissection: quantifying interpretability of deep visual representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 6541–6549

Beebe-Wang N, Okeson A, Althoff T, Lee S-I (2021) Efficient and explainable risk assessments for imminent dementia in an aging cohort study. IEEE J Biomed Health Inform 25(7):2409–2420

Bennett DA, Schneider JA, Buchman AA, Barnes LL, Boyle PA, Wilson RS (2012) Overview and findings from the rush memory and aging project. Curr Alzheimer Res 9(6):646–663

Bjerring JC, Busch J (2021) Artificial intelligence and patient-centered decision-making. Philos Technol 34(2):349–371

Bois MD, El Yacoubi MA, Ammi M (2020) Interpreting deep glucose predictive models for diabetic people using retain. In: International Conference on Pattern Recognition and Artificial Intelligence, pp 685–694. Springer

Bonaz B, Sinniger V, Pellissier S (2020) Targeting the cholinergic anti-inflammatory pathway with vagus nerve stimulation in patients with covid-19? Bioelectron Med 6(1):1–7

Bruckert S, Finzel B, Schmid U (2020) The next generation of medical decision support: a roadmap toward transparent expert companions. Front Artif Intell 3:507973

Cavaliere F, Della Cioppa A, Marcelli A, Parziale A, Senatore R (2020) Parkinson's disease diagnosis: towards grammar-based explainable artificial intelligence. In: 2020 IEEE Symposium on Computers and Communications (ISCC), pp 1–6. IEEE

Chen P, Dong W, Wang J, Lu X, Kaymak U, Huang Z (2020) Interpretable clinical prediction via attention-based neural network. BMC Med Inform Decis Mak 20(3):1–9

Cheng F, Liu D, Du F, Lin Y, Zytek A, Li H, Qu H, Veeramachaneni K (2021) Vbridge: connecting the dots between features and data to explain healthcare models. IEEE Trans Vis Comput Gr 28(1):378–388

Chmiel F, Burns D, Azor M, Borca F, Boniface M, Zlatev Z, White N, Daniels T, Kiuber M (2021) Using explainable machine learning to identify patients at risk of reattendance at discharge from emergency departments. Sci Rep 11(1):1–11

Choi E, Bahadori MT, Sun J, Kulas J, Schuetz A, Stewart W (2016) Retain: an interpretable predictive model for healthcare using reverse time attention mechanism. In: Advances in Neural Information Processing Systems, vol 29

Cho S, Lee G, Chang W, Choi J (2020) Interpretation of deep temporal representations by selective visualization of internally activated nodes. arXiv preprint arXiv:2004.12538

Cinà G, Röber T, Goedhart R, Birbil I (2022) Why we do need explainable ai for healthcare. arXiv preprint arXiv:2206.15363

Clifford GD, Liu C, Moody B, Li-wei HL, Silva I, Li Q, Johnson A, Mark RG (2017) Af classification from a short single lead ecg recording: The physionet/computing in cardiology challenge 2017. In: 2017 Computing in Cardiology (CinC), pp 1–4. IEEE

Clifford GD, Liu C, Moody BE, Roig JM, Schmidt SE, Li Q, Silva I, Mark RG (2017) Recent advances in heart sound analysis. Physiol Meas 38:10–25

Costa ABD, Moreira L, Andrade DCD, Veloso A, Ziviani N (2021) Predicting the evolution of pain relief: ensemble learning by diversifying model explanations. ACM Trans Comput Healthcare 2(4):1–28

Cox DR (1992) Regression models and life-tables. breakthroughs in statistics. Stat Soc 372:527–541

Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y et al (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 486(7403):346–352

Das A, Rad P (2020) Opportunities and challenges in explainable artificial intelligence (xai): a survey. arXiv preprint arXiv:2006.11371

Dau HA, Bagnall A, Kamgar K, Yeh C-CM, Zhu Y, Gharghabi S, Ratanamahatana CA, Keogh E (2019) The ucr time series archive. IEEE/CAA J Autom Sin 6(6):1293–1305

Davagdorj K, Bae J-W, Pham V-H, Theera-Umpon N, Ryu KH (2021) Explainable artificial intelligence based framework for non-communicable diseases prediction. IEEE Access 9:123672–123688

Deng H (2019) Interpreting tree ensembles with intrees. Int J Data Sci Anal 7(4):277–287

Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

Diprose WK, Buist N, Hua N, Thurier Q, Shand G, Robinson R (2020) Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. J Am Med Inform Assoc 27(4):592–600

Dissanayake T, Fernando T, Denman S, Sridharan S, Ghaemmaghami H, Fookes C (2020) A robust interpretable deep learning classifier for heart anomaly detection without segmentation. IEEE J Biomed Health Inform 25(6):2162–2171

Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608

Drew BJ, Harris P, Zègre-Hemsey JK, Mammone T, Schindler D, Salas-Boni R, Bai Y, Tinoco A, Ding Q, Hu X (2014) Insights into the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients. PLoS ONE 9(10):110274

Du M, Liu N, Hu X (2019) Techniques for interpretable machine learning. Commun ACM 63(1):68–77

Duckworth C, Chmiel FP, Burns DK, Zlatev ZD, White NM, Daniels TW, Kiuber M, Boniface MJ (2021) Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during covid-19. Sci Rep 11(1):1–10

Duell J, Fan X, Burnett B, Aarts G, Zhou S-M (2021) A comparison of explanations given by explainable artificial intelligence methods on analysing electronic health records. In: 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), pp 1–4. IEEE

Dwivedi P, Khan AA, Mugde S, Sharma G (2021) Diagnosing the major contributing factors in the classification of the fetal health status using cardiotocography measurements: An automl and xai approach. In: 2021 13th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), pp 1–6. IEEE

El-Bouri R, Eyre DW, Watkinson P, Zhu T, Clifton DA (2020) Hospital admission location prediction via deep interpretable networks for the year-round improvement of emergency patient care. IEEE J Biomed Health Inform 25(1):289–300

Elshawi R, Al-Mallah MH, Sakr S (2019) On the interpretability of machine learning-based model for predicting hypertension. BMC Med Inform Decis Mak 19(1):1–32

ElShawi R, Sherif Y, Al-Mallah M, Sakr S (2019) Ilime: local and global interpretable model-agnostic explainer of black-box decision. In: European Conference on Advances in Databases and Information Systems, pp 53–68. Springer

Faruk MF (2021) Residualcovid-net: An interpretable deep network to screen covid-19 utilizing chest ct images. In: 2021 3rd International Conference on Electrical & Electronic Engineering (ICEEE), pp 69–72. IEEE

Filtjens B, Ginis P, Nieuwboer A, Afzal MR, Spildooren J, Vanrumste B, Slaets P (2021) Modelling and identification of characteristic kinematic features preceding freezing of gait with convolutional neural networks and layer-wise relevance propagation. BMC Med Inform Decis Mak 21(1):1–11

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232

Friedman JH, Popescu BE (2008) Predictive learning via rule ensembles. Ann Appl Stat 2(3):916–954

Ghorbani A, Wexler J, Zou JY, Kim B (2019) Towards automatic concept-based explanations. In: Advances in Neural Information Processing Systems, vol 32

Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE (2000) Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. Circulation 101(23):215–220

Goldstein A, Kapelner A, Bleich J, Pitkin E (2015) Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. J Comput Graph Stat 24(1):44–65

Goyal Y, Feder A, Shalit U, Kim B (2019) Explaining classifiers with causal concept effect (cace). arXiv preprint arXiv:1907.07165

Guidotti R, Monreale A, Ruggieri S, Pedreschi D, Turini F, Giannotti F (2018) Local rule-based explanations of black box decision systems. arXiv preprint arXiv:1805.10820

Gullapalli BT, Carreiro S, Chapman BP, Ganesan D, Sjoquist J, Rahman T (2021) Opitrack: a wearable-based clinical opioid use tracker with temporal convolutional attention networks. Proc ACM Interact Mob Wear Ubiquitous Technol 5(3):1–29

Gulum MA, Trombley CM, Kantardzic M (2021) A review of explainable deep learning cancer detection models in medical imaging. Appl Sci 11(10):4573

Gupta A, Jain J, Poundrik S, Shetty MK, Girish M, Gupta MD (2021) Interpretable ai model-based predictions of ecg changes in covid-recovered patients. In: 2021 4th International Conference on Bio-Engineering for Smart Technologies (BioSMART), pp 1–5. IEEE

Guvenir HA, Acar B, Demiroz G, Cekin A (1997) A supervised machine learning algorithm for arrhythmia analysis. In: Computers in Cardiology 1997, pp 433–436. IEEE

Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems, vol 29

Hartl A, Bachl M, Fabini J, Zseby T (2020) Explainability and adversarial robustness for rnns. In: 2020 IEEE Sixth International Conference on Big Data Computing Service and Applications (BigData-Service), pp 148–156. IEEE

Hatwell J, Gaber MM, Atif Azad RM (2020) Ada-whips: explaining adaboost classification with applications in the health sciences. BMC Med Inform Decis Mak 20(1):1–25

He L, Liu H, Yang Y, Wang B (2021) A multi-attention collaborative deep learning approach for blood pressure prediction. ACM Trans Manag Inf Syst 13(2):1–20

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778

Holzinger AT, Muller H (2021) Toward human-ai interfaces to support explainability and causability in medical ai. Computer 54(10):78–86

Holzinger A, Carrington A, Müller H (2020) Measuring the quality of explanations: the system causability scale (scs). KI-Künstliche Intell 34(2):193–198

Horsak B, Slijepcevic D, Raberger A-M, Schwab C, Worisch M, Zeppelzauer M (2020) Gaitrec, a large-scale ground reaction force dataset of healthy and impaired gait. Sci Data 7(1):1–8

Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861

Hsieh T-Y, Wang S, Sun Y, Honavar V (2021) Explainable multivariate time series classification: a deep neural network which learns to attend to important variables as well as time intervals. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, pp 607–615

Ibrahim L, Mesinovic M, Yang K-W, Eid MA (2020) Explainable prediction of acute myocardial infarction using machine learning and shapley values. IEEE Access 8:210410–210417

Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS (2008) Random survival forests. Ann Appl Stat 2(3):841–860

Ivaturi P, Gadaleta M, Pandey AC, Pazzani M, Steinhubl SR, Quer G (2021) A comprehensive explanation framework for biomedical time series classification. IEEE J Biomed Health Inform 25(7):2398–2408

Jian J-Y, Bisantz AM, Drury CG (1998) Towards an empirically determined scale of trust in computerized systems: distinguishing concepts and types of trust. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol 42, pp 501–505. SAGE Publications Sage CA, Los Angeles, CA

Jiang J, Hewner S, Chandola V (2021) Explainable deep learning for readmission prediction with tree-glove embedding. In: 2021 IEEE 9th International Conference on Healthcare Informatics (ICHI), pp 138–147. IEEE

Johnson AE, Pollard TJ, Shen L, Lehman L-wH, Feng M, Ghassemi M, Moody B, Szolovits P, Anthony Celi L, Mark RG (2016) Mimic-iii, a freely accessible critical care database. Sci Data 3(1):1–9

Jung J-M, Kim Y-H, Yu S, Kyungmi O, Kim CK, Song T-J, Kim Y-J, Kim BJ, Heo SH, Park K-Y et al (2019) Long-term outcomes of real-world Korean patients with atrial-fibrillation-related stroke and severely decreased ejection fraction. J Clin Neurol 15(4):545–554

Kapcia M, Eshkiki H, Duell J, Fan X, Zhou S, Mora B (2021) Exmed: an ai tool for experimenting explainable ai techniques on medical data analytics. In: 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI), pp 841–845. IEEE

Khodabandehloo E, Riboni D, Alimohammadi A (2021) Healthxai: collaborative and explainable ai for supporting early diagnosis of cognitive decline. Futur Gener Comput Syst 116:168–189

Kim S-H, Jeon E-T, Yu S, Oh K, Kim CK, Song T-J, Kim Y-J, Heo SH, Park K-Y, Kim J-M et al (2021) Interpretable machine learning for early neurological deterioration prediction in atrial fibrillation-related stroke. Sci Rep 11(1):1–9

Kim L, Kim J-A, Kim S (2014) A guide for the utilization of health insurance review and assessment service national patient samples. Epidemiology and health **36**

Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F, et al (2018) Interpretability beyond feature attribution: quantitative testing with concept activation vectors (tcav). In: International Conference on Machine Learning, pp 2668–2677. PMLR

Kindermans P-J, Hooker S, Adebayo J, Alber M, Schütt KT, Dähne S, Erhan D, Kim B (2019) The (un) reliability of saliency methods. In: Explainable AI: interpreting, explaining and visualizing deep learning, pp 267–280. Springer

Knaus WA, Harrell FE, Lynn J, Goldman L, Phillips RS, Connors AF, Dawson NV, Fulkerson WJ, Califf RM, Desbiens N et al (1995) The support prognostic model: objective estimates of survival for seriously ill hospitalized adults. Ann Intern Med 122(3):191–203

Kok I, Okay FY, Muyanli O, Ozdemir S (2022) Explainable artificial intelligence (xai) for internet of things: a survey. arXiv preprint arXiv:2206.04800

Kovalchuk SV, Kopanitsa GD, Derevitskii IV, Matveev GA, Savitskaya DA (2022) Three-stage intelligent support of clinical decision making for higher trust, validity, and explainability. J Biomed Inform 127:104013

Krishnakumar S, Abdou T (2020) Towards interpretable and maintainable supervised learning using shapley values in arrhythmia. In: Proceedings of the 30th Annual International Conference on Computer Science and Software Engineering, pp 23–32

Kumarakulasinghe NB, Blomberg T, Liu J, Leao AS, Papapetrou P (2020) Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models. In: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), pp 7–12. IEEE

Kwon BC, Choi M-J, Kim JT, Choi E, Kim YB, Kwon S, Sun J, Choo J (2018) Retainvis: visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. IEEE Trans Vis Comput Gr 25(1):299–309

Lauritsen SM, Kristensen M, Olsen MV, Larsen MS, Lauritsen KM, Jørgensen MJ, Lange J, Thiesson B (2020) Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nat Commun 11(1):1–11

Lemeshow S, May S, Hosmer DW Jr (2011) Applied survival analysis: regression modeling of time-to-event data. Wiley, New York

Leung CK, Fung DL, Mai D, Wen Q, Tran J, Souza J (2021) Explainable data analytics for disease and healthcare informatics. In: 25th International Database Engineering & Applications Symposium, pp 65–74

Li B, Sano A (2020) Extraction and interpretation of deep autoencoder-based temporal features from wearables for forecasting personalized mood, health, and stress. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 4(2):1–26

Linardatos P, Papastefanopoulos V, Kotsiantis S (2020) Explainable ai: a review of machine learning interpretability methods. Entropy 23(1):18

Lin J, Keogh E, Lonardi S, Chiu B (2003) A symbolic representation of time series, with implications for streaming algorithms. In: Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pp 2–11

Lisboa PJ, Ortega-Martorell S, Olier I (2020) Explaining the neural network: A case study to model the incidence of cervical cancer. In: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, pp 585–598. Springer

Looveren AV, Klaise J (2021) Interpretable counterfactual explanations guided by prototypes. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp 650–665. Springer

Lu J, Jin R, Song E, Alrashoud M, Al-Mutib KN, Al-Rakhami MS (2020) An explainable system for diagnosis and prognosis of covid-19. IEEE Internet Things J 8(21):15839–15846

Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 30

Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I (2020) From local explanations to global understanding with explainable ai for trees. Nat Mach Intell 2(1):2522–5839

Luo J, Ye M, Xiao C, Ma F (2020) Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 647–656

Markus AF, Kors JA, Rijnbeek PR (2021) The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. J Biomed Inform 113:103655

Ma D, Wang Z, Xie J, Guo B, Yu Z (2020) Interpretable multivariate time series classification based on prototype learning. In: International Conference on Green, Pervasive, and Cloud Computing, pp 205–216. Springer

Maweu BM, Dakshit S, Shamsuddin R, Prabhakaran B (2021) Cefes: a cnn explainable framework for ecg signals. Artif Intell Med 115:102059

Mikalsen KØ, Bianchi FM, Soguero-Ruiz C, Skrøvseth SO, Lindsetmo R-O, Revhaug A, Jenssen R (2016) Learning similarities between irregularly sampled short multivariate time series from ehrs

Mishra S, Dutta S, Long J, Magazzeni D (2021) A survey on the robustness of feature importance and counterfactual explanations. arXiv preprint arXiv:2111.00358

Mohseni S, Block JE, Ragan ED (2018) A human-grounded evaluation benchmark for local explanations of machine learning. arXiv preprint arXiv:1801.05075

Moncada-Torres A, van Maaren MC, Hendriks MP, Siesling S, Geleijnse G (2021) Explainable machine learning can outperform cox regression predictions and provide insights in breast cancer survival. Sci Rep 11(1):1–13

Mondal AK, Bhattacharjee A, Singla P, Prathosh A (2021) xvitcos: Explainable vision transformer based covid-19 screening using radiography. IEEE J Transl Eng Health Med 10:1–10

Moreno-Sanchez PA (2020) Development of an explainable prediction model of heart failure survival by using ensemble trees. In: 2020 IEEE International Conference on Big Data (Big Data), pp 4902–4910. IEEE

Morris MD (1991) Factorial sampling plans for preliminary computational experiments. Technometrics 33(2):161–174

Mousavi S, Afghah F, Acharya UR (2020) Han-ecg: an interpretable atrial fibrillation detection model using hierarchical attention networks. Comput Biol Med 127:104057

Muller H, Mayrhofer MT, Van Veen E-B, Holzinger A (2021) The ten commandments of ethical medical ai. Computer 54(07):119–123

Nicolaides AN, Kakkos SK, Kyriacou E, Griffin M, Sabetai M, Thomas DJ, Tegos T, Geroulakos G, Labropoulos N, Doré CJ et al (2010) Asymptomatic internal carotid artery stenosis and cerebrovascular risk stratification. J Vasc Surg 52(6):1486–1496

Nori H, Jenkins S, Koch P, Caruana R (2019) Interpretml: A unified framework for machine learning interpretability. arXiv preprint arXiv:1909.09223

Oba Y, Tezuka T, Sanuki M, Wagatsuma Y (2021) Interpretable prediction of diabetes from tabular health screening records using an attentional neural network. In: 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA), pp 1–11. IEEE

Obeid I, Picone J (2016) The temple university hospital eeg data corpus. Front Neurosci 10:196

Ochoa JGD, Csiszár O, Schimper T (2021) Medical recommender systems based on continuous-valued logic and multi-criteria decision operators, using interpretable neural networks. BMC Med Inform Decis Mak 21(1):1–15

Okay FY, Yıldırım M, Özdemir S (2021) Interpretable machine learning: a case study of healthcare. In: 2021 International Symposium on Networks, Computers and Communications (ISNCC), pp 1–6. IEEE

Oviedo F, Ren Z, Sun S, Settens C, Liu Z, Hartono NTP, Ramasamy S, DeCost BL, Tian SI, Romano G et al (2019) Fast and interpretable classification of small x-ray diffraction datasets using data augmentation and deep neural networks. NPJ Comput Mater 5(1):1–9

Pal A, Sankarasubbu M (2021) Pay attention to the cough: early diagnosis of covid-19 using interpretable symptoms embeddings with cough sound signal processing. In: Proceedings of the 36th Annual ACM Symposium on Applied Computing, pp 620–628

Pan SJ, Tsang IW, Kwok JT, Yang Q (2010) Domain adaptation via transfer component analysis. IEEE Trans Neural Networks 22(2):199–210

Pang X, Forrest CB, Lê-Scherban F, Masino AJ (2019) Understanding early childhood obesity via interpretation of machine learning model predictions. In: 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), pp 1438–1443. IEEE

Pataky TC (2010) Generalized n-dimensional biomechanical field analysis using statistical parametric mapping. J Biomech 43(10):1976–1982

Payrovnaziri SN, Chen Z, Rengifo-Moreno P, Miller T, Bian J, Chen JH, Liu X, He Z (2020) Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. J Am Med Inform Assoc 27(7):1173–1185

Penafiel S, Baloian N, Sanson H, Pino JA (2020) Predicting stroke risk with an interpretable classifier. IEEE Access 9:1154–1166

Pereira CR, Pereira DR, Da Silva FA, Hook C, Weber SA, Pereira LA, Papa JP (2015) A step towards the automated diagnosis of parkinson's disease: Analyzing handwriting movements. In: 2015 IEEE 28th International Symposium on Computer-based Medical Systems, pp 171–176. IEEE

Pereira T, Ding C, Gadhoumi K, Tran N, Colorado RA, Meisel K, Hu X (2019) Deep learning approaches for plethysmography signal quality assessment in the presence of atrial fibrillation. Physiol Meas 40(12):125002

Perez E, Strub F, De Vries H, Dumoulin V, Courville A (2018) Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 32

Plawiak P (2017) Ecg signals (1000 fragments). Mendeley Data, v3

Pölsterl S, Navab N, Katouzian A (2016) An efficient training algorithm for kernel survival support vector machines. arXiv preprint arXiv:1611.07054

Prentzas N, Nicolaides A, Kyriacou E, Kakas A, Pattichis C (2019) Integrating machine learning with symbolic reasoning to build an explainable ai model for stroke prediction. In: 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), pp 817–821. IEEE

Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pre-training

Rashed-Al-Mahfuz M, Haque A, Azad A, Alyami SA, Quinn JM, Moni MA (2021) Clinically applicable machine learning approaches to identify attributes of chronic kidney disease (ckd) for use in low-cost diagnostic screening. IEEE J Transl Eng Health Med 9:1–11

Reyna MA, Josef C, Seyedi S, Jeter R, Shashikumar SP, Westover MB, Sharma A, Nemati S, Clifford GD (2019) Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. In: 2019 Computing in Cardiology (CinC), p 1. IEEE

Ribeiro MT, Singh S, Guestrin C (2016) "Why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 1135–1144

Ribeiro MT, Singh S, Guestrin C (2018) Anchors: high-precision model-agnostic explanations. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 32

Rojat T, Puget R, Filliat D, Del Ser J, Gelin R, Díaz-Rodríguez N (2021) Explainable artificial intelligence (xai) on timeseries data: a survey. arXiv preprint arXiv:2104.00950

Rubini LJ, Eswaran P (2015) Uci chronic kidney disease. School Inf. Comput., Sci. Univ, California, Irvine

Sadhukhan D, Pal S, Mitra M (2018) Automated identification of myocardial infarction using harmonic phase distribution pattern of ecg data. IEEE Trans Instrum Meas 67(10):2303–2313

Sahakyan M, Aung Z, Rahwan T (2021) Explainable artificial intelligence for tabular data: a survey. IEEE Access 9:135392–135422

Sakr S, Elshawi R, Ahmed A, Qureshi WT, Brawner C, Keteyian S, Blaha MJ, Al-Mallah MH (2018) Using machine learning on cardiorespiratory fitness data for predicting hypertension: the henry ford exercise testing (fit) project. PLoS ONE 13(4):0195344

Saltelli A, Annoni P, Azzini I, Campolongo F, Ratto M, Tarantola S (2010) Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. Comput Phys Commun 181(2):259–270

Schalk G, McFarland DJ, Hinterberger T, Birbaumer N, Wolpaw JR (2004) Bci 2000: a general-purpose brain-computer interface (bci) system. IEEE Trans Biomed Eng 51(6):1034–1043

Schölkopf B, Locatello F, Bauer S, Ke NR, Kalchbrenner N, Goyal A, Bengio Y (2021) Toward causal representation learning. Proc IEEE 109(5):612–634

Seedat N, Aharonson V, Hamzany Y (2020) Automated and interpretable m-health discrimination of vocal cord pathology enabled by machine learning. In: 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), pp 1–6. IEEE

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp 618–626

Sha C, Cuperlovic-Culf M, Hu T (2021) Smile: systems metabolomics using interpretable learning and evolution. BMC Bioinform 22(1):1–17

Shafer G (2016) Dempster's rule of combination. Int J Approx Reason 79:26–40

Shamout FE, Zhu T, Sharma P, Watkinson PJ, Clifton DA (2019) Deep interpretable early warning system for the detection of clinical deterioration. IEEE J Biomed Health Inform 24(2):437–446

Shankaranarayana SM, Runje D (2019) Alime: autoencoder based approach for local interpretability. In: International Conference on Intelligent Data Engineering and Automated Learning, pp 454–463. Springer

Shashikumar SP, Josef CS, Sharma A, Nemati S (2021) Deepaise-an interpretable and recurrent neural survival model for early prediction of sepsis. Artif Intell Med 113:102036

Shrikumar A, Greenside P, Kundaje A (2017): Learning important features through propagating activation differences. In: International Conference on Machine Learning, pp 3145–3153. PMLR

Siddiqui SA, Mercier D, Munir M, Dengel A, Ahmed S (2019) Tsviz: demystification of deep learning models for time-series analysis. IEEE Access 7:67027–67040

Simonyan K, Vedaldi A, Zisserman A (2013) Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034

Slijepcevic D, Horst F, Lapuschkin S, Horsak B, Raberger A-M, Kranzl A, Samek W, Breiteneder C, Schöllhorn WI, Zeppelzauer M (2021) Explaining machine learning models for clinical gait analysis. ACM Trans Comput Healthcare 3(2):1–27

Song X, Yu AS, Kellum JA, Waitman LR, Matheny ME, Simpson SQ, Hu Y, Liu M (2020) Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction. Nat Commun 11(1):1–12

Spildooren J, Vercruysse S, Desloovere K, Vandenberghe W, Kerckhofs E, Nieuwboer A (2010) Freezing of gait in parkinson's disease: the impact of dual-tasking and turning. Mov Disord 25(15):2563–2570

Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M (2014) triving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806

Su Z, Figueiredo MC, Jo J, Zheng K, Chen Y (2020) Analyzing description, user understanding and expectations of ai in mobile health applications. In: AMIA Annual Symposium Proceedings, vol. 2020, p 1170. American Medical Informatics Association

Sun C, Dui H, Li H (2021) Interpretable time-aware and co-occurrence-aware network for medical prediction. BMC Med Inform Decis Mak 21(1):1–12

Sun Z, Dong W, Shi J, He K, Huang Z (2021) Attention-based deep recurrent model for survival prediction. ACM Trans Comput Healthcare 2(4):1–18

Sundararajan M, Taly A, Yan Q (2017) Axiomatic attribution for deep networks. In: International Conference on Machine Learning, pp 3319–3328. PMLR

Tahmassebi A, Martin J, Meyer-Baese A, Gandomi AH (2020) An interpretable deep learning framework for health monitoring systems: a case study of eye state detection using eeg signals. In: 2020 IEEE Symposium Series on Computational Intelligence (SSCI), pp 211–218. IEEE

Thimoteo LM, Vellasco MM, Amaral J, Figueiredo K, Yokoyama CL, Marques E (2022) Explainable artificial intelligence for covid-19 diagnosis through blood test variables. J Control Autom Electr Syst 33(2):625–644

Thorsen-Meyer H-C, Nielsen AB, Nielsen AP, Kaas-Hansen BS, Toft P, Schierbeck J, Strøm T, Chmura PJ, Heimann M, Dybdahl L et al (2020) Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. Lancet Digit Health 2(4):179–191

Tjoa E, Guan C (2020) A survey on explainable artificial intelligence (xai): toward medical xai. IEEE Trans Neural Networks Learn Syst 32(11):4793–4813

Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25(1):44–56

Van der Maaten L, Hinton G (2008) Visualizing data using t-sne. J Mach Learn Res 9(11):2579–2605

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Conference on Neural Information Processing Systems, vol 30

Wachter S, Mittelstadt B, Russell C (2017) Counterfactual explanations without opening the black box: automated decisions and the gdpr. Harv. JL & Tech. 31:841

Waitman LR, Aaronson LS, Nadkarni PM, Connolly DW, Campbell JR (2014) The greater plains collaborative: a pcornet clinical research data network. J Am Med Inform Assoc 21(4):637–641

Wang G, Zhou Y, Huang F-J, Tang H-D, Xu X-H, Liu J-J, Wang Y, Deng Y-L, Ren R-J, Xu W et al (2014) Plasma metabolite profiles of Alzheimer's disease and mild cognitive impairment. J Proteome Res 13(5):2649–2658

Wang D, Yang Q, Abdul A, Lim BY (2019) Designing theory-driven user-centric explainable ai. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp 1–15

Ward IR, Wang L, Lu J, Bennamoun M, Dwivedi G, Sanfilippo FM (2021) Explainable artificial intelligence for pharmacovigilance: what features are important when predicting adverse outcomes? Comput Methods Programs Biomed 212:106415

Weiss SM, Indurkhya N, Zhang T (2010) Fundamentals of predictive text mining. Springer, New York

Wexler J, Pushkarna M, Bolukbasi T, Wattenberg M, Viégas F, Wilson J (2019) The what-if tool: interactive probing of machine learning models. IEEE Trans Vis Comput Gr 26(1):56–65

Wickstrøm K, Mikalsen KØ, Kampffmeyer M, Revhaug A, Jenssen R (2020) Uncertainty-aware deep ensembles for reliable and explainable predictions of clinical time series. IEEE J Biomed Health Inform 25(7):2435–2444

Wu M, Hughes M, Parbhoo S, Zazzi M, Roth V, Doshi-Velez F (2018) Beyond sparsity: tree regularization of deep models for interpretability. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 32

Yan L, Zhang H-T, Goncalves J, Xiao Y, Wang M, Guo Y, Sun C, Tang X, Jing L, Zhang M et al (2020) An interpretable mortality prediction model for covid-19 patients. Nat Mach Intell 2(5):283–288

Yeh C-K, Kim B, Arik S, Li C-L, Pfister T, Ravikumar P (2020) On completeness-aware concept-based explanations in deep neural networks. Adv Neural Inf Process Syst 33:20554–20565

Ye L, Keogh E (2009) Time series shapelets: a new primitive for data mining. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 947–956

Zafar MR, Khan NM (2019) Dlime: a deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems. arXiv preprint arXiv:1906.10263

Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: European Conference on Computer Vision, pp 818–833. Springer

Zeiler MD, Taylor GW, Fergus R (2011) Adaptive deconvolutional networks for mid and high level feature learning. In: 2011 International Conference on Computer Vision, pp. 2018–2025. IEEE

Zeng X, Yu G, Lu Y, Tan L, Wu X, Shi S, Duan H, Shu Q, Li H (2020) Pic, a paediatric-specific intensive care database. Sci Data 7(1):1–8

Zeng X, Hu Y, Shu L, Li J, Duan H, Shu Q, Li H (2021) Explainable machine-learning predictions for complications after pediatric congenital heart surgery. Sci Rep 11(1):1–11

Zhai B, Perez-Pozuelo I, Clifton EA, Palotti J, Guan Y (2020) Making sense of sleep: multimodal sleep stage classification in a large, diverse population using movement and cardiac sensing. Proc ACM Interact Mob Wear Ubiquitous Technol 4(2):1–33

Zhang G-Q, Cui L, Mueller R, Tao S, Kim M, Rueschman M, Mariani S, Mobley D, Redline S (2018) The national sleep research resource: towards a sleep data commons. J Am Med Inform Assoc 25(10):1351–1358

Zhang X, Yao L, Dong M, Liu Z, Zhang Y, Li Y (2020) Adversarial representation learning for robust patient-independent epileptic seizure detection. IEEE J Biomed Health Inform 24(10):2852–2859

Zhang O, Ding C, Pereira T, Xiao R, Gadhoumi K, Meisel K, Lee RJ, Chen Y, Hu X (2021) Explainability metrics of deep convolutional networks for photoplethysmography quality assessment. IEEE Access 9:29736–29745

Zhang Y, Yang D, Liu Z, Chen C, Ge M, Li X, Luo T, Wu Z, Shi C, Wang B et al (2021) An explainable supervised machine learning predictor of acute kidney injury after adult deceased donor liver transplantation. J Transl Med 19(1):1–15

Zheng K, Cai S, Chua HR, Wang W, Ngiam KY, Ooi BC (2020) Tracer: a framework for facilitating accurate and interpretable analytics for high stakes applications. In: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, pp 1747–1763

Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2921–2929