# Explainable AI for clinical risk prediction: a systematic review of methods and modalities

Munib Mesinovic Peter Watkinson Tingting Zhu

August 2023

**Abstract**

Recent advancements in AI applications to healthcare have shown incredible promise in surpassing human performance in diagnosis and disease prognosis. With increasing complexity of AI models, however, come concerns regarding their opacity, potential biases, and the need for interpretability. To ensure trust and reliability in AI systems, especially in clinical risk prediction models, explainability becomes crucial. Explainability is usually refered to as an AI system's ability to provide robust interpretation of its decision-making logic or the decisions themselves to human stakeholders. In clinical risk prediction, other aspects of explainability like fairness, bias, trust, and transparency also represent important concepts beyond just interpretability. In this review, we address the relationship between these concepts as they are often used together or interchangeably. This review also discusses recent progress in developing explainable models for clinical risk prediction, highlighting the importance of quantitative and clinical evaluation and validation across multiple common modalities in clinical practice. It emphasizes the need for external validation and the combination of diverse interpretability methods to enhance trust and fairness. The adoption of rigorous testing, such as using synthetic datasets with known generative factors, can further enhance the reliability of explainability methods. Open access and code sharing resources are essential for transparency and reproducibility, enabling the growth and trustworthiness of explainable research. While challenges exist, an end-to-end approach to explainability in clinical risk prediction, incorporating stakeholders from clinicians to developers, is essential for success.

## 1 Introduction

### 1.1 Motivation

The advancement of artificial intelligence, and specifically machine learning and deep learning, in different areas of society has been staggering. Applied AI can now beat humans in a string of complex logic-based games, can produce literary-level text, and enable a priori modelling of complicated protein structures that has long eluded biochemists [1, 2, 3, 4]. This progress has not eluded the field of healthcare where AI has seen progress at outperforming clinicians at breast cancer screenings, improved drug discovery, and predictions of several diseases, including, more recently, COVID-19 outcomes [5, 6, 7, 8, 9]. With the rising capabilities of AI in patient care and prognosis come also rising risks of mismanagement, misclassification, misunderstanding (of the models themselves), and misuse.

Greater prediction capabilities often come with greater complexities of models which makes them more opaque and unclear to both their developers as well as potential users. As far as clinical risk prediction models are concerned, decisions about patient treatment or diagnosis affect people's lives profoundly, and AI systems are not without fault, sometimes leading to unreliable results or biased decision-making. For both clinicians and patients to understand and trust the decision-making process, especially when the impact of the decisions are significant, they need to be able to "evaluate and identify how [their personal] data is being used and whether the

outcome is correct" [10]. To do so, AI models are often checked for their explainability capacity which includes their interpretability and propensity to bias before being relied upon by clinicians.

Explainability, whose full definition will be addressed later, is an attribute of an (AI) automated decision system which describes the system's readiness to provide robust explanations of either its inner decision-making logic (inclusive) or the decisions themselves to human stakeholders [11]. In other words, an explainable AI model can provide sufficient justification of its decisions, making it easier to identify potential flaws in learning and sources of bias. Explainability can also aid in promoting positive aspects of machine learning by providing new insights into predictive patterns relevant to disease prognosis and outcomes [12]. As highlighted in previous work, explainable AI applications can make more accurate predictions as well as offer increased transparency and fairness over their human counterparts in clinical applications [13]. These insights can help generalise the models to different patient populations, increasing robustness and decreasing chances of unequal treatment for protected groups, while also learning more about the model, the data, and the problem itself.

Explainable AI models could, in addition, be a regulatory requirement. Existing regulatory frameworks like the General Data Protection Regulation (GDPR) might include a right to an explanation of automated decisions, depending on who one asks. Under Art. 15(1) the controller must provide "meaningful information about the logic involved" in AI systems, especially when explanations are needed to guarantee accuracy and to potentially challenge correctness [14, 15]. The interpretation of the word 'meaningful' leaves it to data protection authorities to decide on what information necessitates the enforcement of explainability. [16] are sceptical about legally enforced explainability since it seems that an explicit and legally sound right to an explanation was intentionally not included in the final version of the GDPR despite pressure from European lawmakers. Whatever the case may be, it is clear that future trajectories in AI applications to sensitive data areas like healthcare will lean towards robust explainability methods in clinical risk prediction. Finally, even when there is no legal obligation, it is important for clinicians to be able to both understand suggestions provided by clinical risk prediction models and be able to justify their own decision making to colleagues and patients.

This review hopes to present recent progress on multiple modality and methodology fronts in developing explainable models for clinical risk prediction. It also includes a summary assessment of these approaches with regards to quantitative and/or clinical evaluation and validation.

## 1.2 Contribution

We aim to provide a comprehensive view of applied explainability work in clinical risk prediction since 2019 which is inclusive of progress made across and within multiple modalities and which addresses topical concerns of reproducibility, external validation, and quantitative evaluation of explainability not attempted so far. Previous literature reviews, including [17, 18, 19, 20, 21] has either been focussed on general classification and diagnosis using AI which is too broad of an approach, a specific modality, a specific disease, or more philosophical rather than implementation-based approaches to validation, fairness, trustworthiness etc. but never combining all of the above for a comprehensive commentary on the current trajectories of explainable AI in healthcare. While we do not aim for a social and/or critical analysis of concepts like trust, we do comment on the role these ideas play in achieving truly explainable AI for clinical risk prediction. We will reveal the lack of common culture of clinically validating and quantitatively evaluating explainability methods in various modalities and applications for clinical risk prediction. We will also highlight patterns in the availability of reproducible code and results which undermines attempts at verifying research and increasing transparency of explainability work. In the end, we will highlight the need for a change in accepted research culture when developing and implementing existing or novel explainability frameworks that will answer a set of fundamental questions:

- Does the proposed machine learning model for clinical risk prediction provide added clinical benefit whether in knowledge gain or practical importance?
- Can similar* performance be achieved using a simpler or glass-box model instead of complex and costly architectures?
- Have the model prediction results been investigated by a clinician to some extent?
- Has the explainability method and its explanations for decisions been validated by a clinician?
- Has the explainability method been quantitatively evaluated in any manner, whether it be a metric or a comparative analysis?
- How will the explainable AI model benefit patients and how will that be clearly communicated?

The proposed framework for understanding the complex overlapping terms covered in this literature review can be seen in Figure 1 which encapsulates our reasoning of these topics in relation to explainability as a concept.

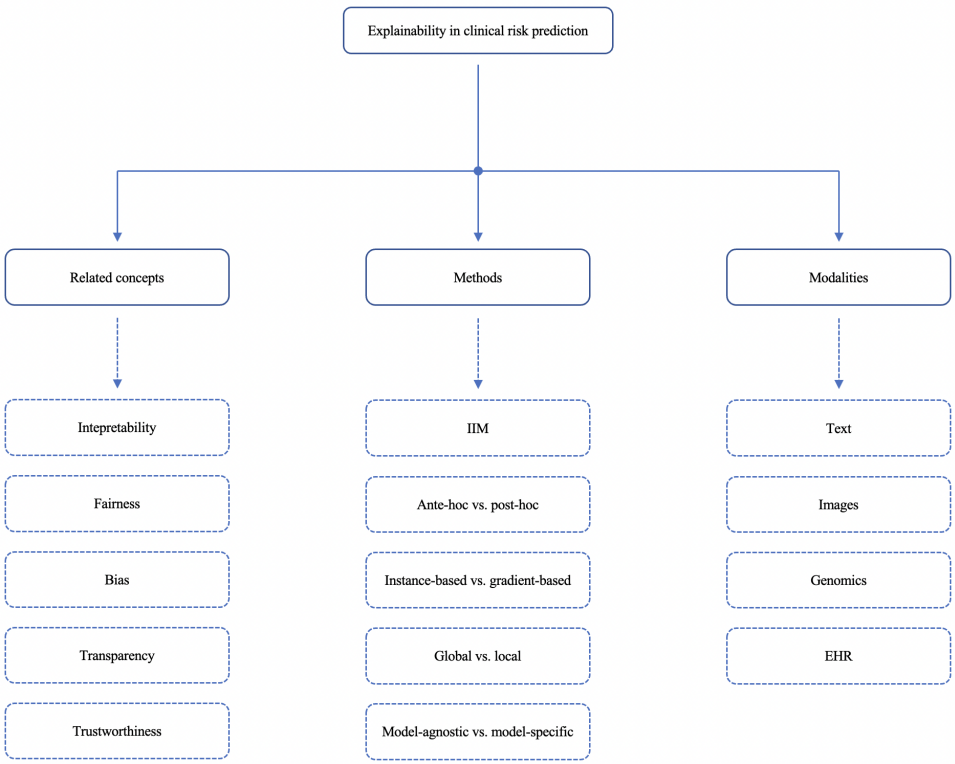


Figure 1: Framework for literature review structure showing the key terms, methods, and modalities investigated under XAI for clinical risk prediction

# 2 Background

## 2.1 Definitions

Explainability in the context of AI refers to the ability to understand and not just interpret the decisions or predictions made by an AI system. A key terminological difference exists that must

be made now between 'interpreting' and 'explaining' an AI model. To interpret a decision is to understand what it represents in the interpretable domain ie. in terms of the input features. In other words, which aspect of the data have contributed to making this prediction and this will be elucidated upon in the following paragraphs but suffice to say interpretations might 'explain' predictions without elucidating the mechanisms by which models work. Explaining a decision is not just functional but can also be deeper and more mechanistic in as far as it can concern the algorithmic underpinnings of the learning process of the model within [22]. To truly be able to do so involves providing transparent, trustworthy, unbiased, fair, and intelligible explanations for the reasoning behind the AI's output, enabling users or stakeholders to comprehend how the system arrived at a particular conclusion. Explainability aims to bridge the gap between the inherently complex inner workings of AI algorithms and the need for human comprehension that cannot be separated from concepts of fairness, bias, trustworthiness, transparency, and interpretability but which each does not fully capture on its own. Although we represent explainability here as a functional umbrella term encompassing these concepts, the boundaries are quite fluid as explainability can relate simultaneously to identifying and understanding biases in AI models, aiding the development of fairer systems, guaranteeing transparent model implementation, and providing intepretations of predictions.

As such, explainability relates to the deeper inner working of the model while also concerning the broader social and human interactions of the model in ways that none of the individual concepts within do. As [23] states, "explanations are social and involve conversations." From this philosophical perspective, interpretation is a subjective action, while explanation involves interactions [24]. These interactions do not just include functional interpretations of results but the meaning of the results and how they impact humans (bias and fairness) as well as the comprehensibility of the model predictions to relevant stakeholders (transparency and trustworthiness).

An important terminology note for the review is on the concept of clinical risk prediction versus diagnosis or classification. What we mean by AI for diagnosis or classification is in cases when the disease is usually already present and AI is used to help with faster, more accurate, or more insightful detection of the condition whereas risk prediction as we define it is more similar to prognostic clinical prediction, or predicting the risk of having the disease in the near- or long-term future. As such, prediction is then taken as a longitudinal instead of being a cross-sectional task and the patient population and associated data is more general with the first clinical starting point taken usually as inclusion criteria and the outcome occasionally being censored due to the existence of a prognostic time window and loss to follow-up [25]. This invites the use of survival analysis methods some of which are included in this review. Why that is important for explainability is that this allows for a greater focus on temporal trends and recurrent models for which some explainability methods have been developed but which under past systematic reviews have not received sufficient attention. For some modalities, however, like in the case of imaging, the boundary between clinical risk prediction and diagnosis is blurred especially since major explainability applications have been made on data rarely used for prognostic analysis. Inclusion of these research papers will be decided on a case-by-case basis even when explicit usage of clinical risk prediction is not specified so that a broader overview of explainability for medical imaging can be reliably ascertained. In other cases like for text data, clinical risk prediction is so rare that we decided to include applications to medical coding considering how sometimes medical coding applications can also be seen as examples of machine learning for phenotyping [26, 27]. These cases will be kept to a minimum and they will not impact the overall analyses and conclusions of the literature review but rather are included for the sake of methodological completion of XAI applications.

In common practice across a myriad of machine learning research applications, the terms interpretability and explainability are often used interchangeably. And while interpretability is a key concept underpinning explainability, it is not fully equivalent to it, in fact, "there is

no clear existing definition or evaluation criteria for interpretability" itself [28]. More recent work has attempted to use metrics such as similarity, bias detection, execution time, and trust for measuring different interpretability methods' performance but the criteria are often subject to bias and and might be measuring other aspects of the learning process not related to interpretability itself [29]. One way to attempt to define the concept is by describing its practical usefulness. For example, we can define interpretability as the degree to which humans can understand and comprehend the predictions made by machine learning models [29]. Here we go further and state that **interpretability of an AI system relates to how well the output predictions whether it be in clinical risk modelling or elsewhere can be interpreted by behaviour in the input features**. This can include concepts like feature importance and visualisations of learning [28]. Specifically, the usefulness of interpretability is in providing humans with an understanding of the most influential features that contribute to a specific prediction and the manner in which they do so. The end goal is to assist clinicians in perusing the reasoning behind risk modelling for their patients and thereby being able to both evaluate the prediction reasoning as well as learn relevant trends from features most important for prediction. Following this line of thought, the "more interpretable a machine learning system is, the easier it is to identify cause-and-effect relationships within the system's inputs and outputs" [30]. An example could include a model learning to predict clinical risk of acute conditions like heart attack from time-series data and being able to highlight specific trends in time for measurement of heart rate important to making those predictions. Explainability, as we have mentioned earlier and will elucidate more ahead, is a wider concept which includes interpretability as it satisfies understanding internal learning logic and pattern recognition but is concerned also with user experience of the system, data integration, problem motivation and definition, and bias and fairness.

The concept of bias has been increasingly mentioned in clinical risk prediction with AI influenced by similar challenges in other fields like law and insurance policy automation. Not all bias is necessarily problematic, as (machine) learning is often dependent on it, but certain patterns can be representative of underlying inequities and, in fact, propagate them further. Examples include exclusion of African-Americans in clinical studies which resulted in unfair and less accurate risk assessment both in research and in practice [31, 32]. The sources of bias can include skewed data collected through and representing a system that is often biased itself, such as using criminal records for automating sentencing in the United States, problematic feature engineering by making assumptions about the relevance or lack thereof of certain features, using a limited number of features that might bias inference, stark differences in sample size between vulnerable subpopulations which might be optimistic for outcomes concerning them, and using proxy features like, for example, housing codes (an indirect indicator of socioeconomic status) to make predictions of academic success for public school students [33, 34, 35]. [36] defines **bias as a kind of unfair systematic error that causes models to consistently wrongly predict for a certain subgroup of patients** and when those groups are vulnerable then bias can lead to unfair outcomes. Addressing bias is often done in a multi-faceted approach but it first requires detecting it. Explainability methods can help with that. [11] state that bias can be detected and corrected through explainability thereby implying that an explainable model would necessarily reduce bias or at the least expose it. [12] have studied the implications of using explainability to address bias in loan applications decisions with machine learning methods and evaluate the impact explainability (interpretability) methods like SHAP and LIME have on reducing bias with human-computer interaction. They also add synthetic bias to gender and age for control purposes. Their results show overriding of biased decisions by humans when provided with limited explanations compared to no interpretability albeit with no sufficient statistical significance. Therefore, an explainable model whose every aspect, including data collection and integration, is understandable to human users such as clinicians, would be vulnerable for bias exposure and would be understandable to a level which would allow accounting for the bias in the system itself.

Fairness is related to bias which is often used interchangeably but which [37] note is more

concerned with creating just usage of the systems rather than simply identifying bias in the data itself. They provide an example of an automated hiring system in which men are preferred over women. The system learns the criteria for hiring automatically from biased training data which makes it not intrinsically unfair, rather biased in a field with male prevalence. Once again, a sound and rigorous definition is lacking, but we can deduce that fairness for our purposes concerns making decisions or predictions for clinical risk which avoids delivering injustice, propagating prejudice, and inequitably serving different communities in clinical care. **Bias is thus an aspect of the data or model learning whereas fairness or the lack of it comes as a result of machine decision-making under the influence of bias on vulnerable communities** [38]. Indeed, explainability can similarly account for creating fair machine learning methods and applications in the domain of healthcare going beyond just accounting for bias in the data or parameter tuning [39]. As [40] note, unfairness in such systems appears when decisions get made after the introduction of bias into the system, a separate process. Unfairness can then be detected through counterfactual approaches where "a decision is fair towards an individual if it the same in (a) the actual world and (b) a counterfactual world where the individual belonged to a different demographic group" [40]. What this means in clinical risk modelling in practice is that both data need to be recognised for its potential biases as well as the feature processing and model inference stages, and the results need to be validated on various subpopulations either through counterfactual analysis by switching category groupings or stratified analysis etc. Thus, a fair machine learning system could still preserve certain biases which are unavoidable and aid predictive capacity while not propagating inequity in its decision-making. These interpretations and applications are up to clinicians and other users to make but they can only make them if the models are explainable enough to be able to understand potential presence of unfairness.

Another example of fairness building up on bias rather than being equivalent terms are ensemble systems that check for bias as well as address its understanding, measurement, diagnosis and poential steps for mitigation [41, 42]. Rather, it is not enough to state there is bias for a system to be evaluated for fairness, more steps need to be done to address it and communicate it. Several tools have been proposed to address these needs which consist of using several possible metrics for evaluating bias on the individual and global levels using distance, separation, and distortions in the feature and sample space. Feature correlations and perturbation analysis for sensitive clinical attributes can help identify measured biases. Mitigation can involve a range of steps including at the pre-, during, and post-processing stages. Machine learning models can be evaluated for fairness during learning with approaches like using counterfactual probabilities of two samples chosen from different groups being the same with respect to a given (non-sensitive) feature [40, 41]. The end outcome of such tools are visualisations which reveal group skewing either through heatmaps or clusters and the impact what-if style prodding questions have on the machine learning model as part of the fairness framework. Of course, these methods rely on the assumption that the vulnerable or sensitive groups of samples and features are known a priori which is easier to specify in clinical scenarios due to extensive literature on bias and discrimination across clinical care. Another limitation is implied dichotomy in sensitive groups which might not always hold in cases of hierarchical systems of discrimination.

A system is considered transparent, as defined by [43], if it is "human-understandable." Transparency does not necessarily pertain to any internal methodology of the model, per say, but rather refers to the broader system encompassing the model, its design, motivation, development, distribution, and validation as well. It is highly linked with the concept of trust, as the goal is for human users of the system to deal with a proper interface in the hopes of invoking trust and achieving desired human-machine teaming performance. Critical in the clinical realm, the end users are patients and/or clinicians whose trust must be depended on if these AI systems are to revolutionise healthcare practice [44]. Transparency, as a requirement, thus, cannot just be achieved through the use of computational methods, rather, it relies on involving humans into the model creation pipeline end-to-end as they, and their experience,

is what defined the system as transparent or not. Stakeholders do not need to understand in detail the inner working of the system to achieve transparency, but the nature of the predictions made by the model should be clear (risks or deterministic predictions). For purposes of ethical evaluation of systems, especially in clinical care, transparency would answer questions regarding motivations, interests, and needs achieved through the system for the different stakeholders, and thus be beneficial in elucidating potential conflicts of interest [45]. Transparency, when satisfied, would allow patients, developers, and clinicians, as well as other stakeholders, to evaluate potential conflicts of interest and other negative consequences by opening up the system creation process outside of just developers' oversight [46, 47, 48].

A common difference between transparency and interpretability, most of all, is that transparency pertains to parts outside the machine learning system itself. Interpretability concerns itself with interpreting the model prediction outputs as a function of the inputs. Transparency is focused on the user experience with the system as a whole [11]. Per [46], a transparent AI system for healthcare needs to have its objective and clinical scenario elucidated and verified with clinicians and other stakeholders, including end users, before any system design has taken place. A key defining characteristic of a transparent AI system in healthcare is that its relative transparency relies on the experience of the clinical end users and not on any particular method. While interpretability might be achieved without external non-expert verification of the machine learning models, transparency necessitates that each design choice meets a certain level of evidence prior to development through empirical target user feedback. Furthermore, it should be accessible either in documentation or presentation how this evidence was collected and what the achieved levels of transparency are. Setting out a cohesive strategy for successful communication of these requirements with stakeholders might be a large bottleneck in operations but can drastically affect how users perceive the transparency, and explainability, of the system they are using [49]. For these purposes, it is vital there is some measurement of user experience with the system, either through clinical validation or human-computer interaction testing [50]. In other words, **transparency is the characteristic of an AI system to have its motivation, design, and implementation elucidated to stakeholders whether it be through documentation or user-level testing or some other means**. Recognising the limitations of end user involvement in design of systems in healthcare, [46] propose an evaluation framework using INTRPRT guidelines with user research, prior experience, empirical user testing, and need assessment in a detailed format for developers to peruse. Without explicitly stating how their published research has attempted to address the transparency requirements, researchers risk their proposals being inadequate for real-world implementations with clinical end users.

As already mentioned, trust or trustworthiness is closely intertwined with the concept of transparency, as users tend to be more willing to rely on systems that have been developed end-to-end and applied in a transparent manner [51]. Out of all of the concepts underpinning explainability, trustworthiness is the most difficult to separate, mostly because it is usually explained through a system already being fair, accountable, and transparent [52]. Sometimes, trustworthiness is also defined as satisfying certain confidence levels in prediction results and the methods being interpretable [53]. Whether a model is trustworthy, thus, might depend on individual interpretation. One model might be trusted to make certain predictions but not with all tasks. Trustworthiness plays a key balancing role between increased interpretability and transparency, on the one hand, and increased predictive performance on the other. For clinical risk prediction, a model must be sensitive, well-calibrated, and achieve good prediction results to achieve better trustworthiness, factors which are not accounted for in the other explainability sub-concepts. Therefore, to achieve integrated explainability, trustworthiness is the most linked concept with predictive performance, necessary for advancing change and implementing models in real-world clinical care [54].

## 2.2 Evaluation Criteria

The previous introductory section highlighted the different faces of explainability that each on their own involved significant further study and analysis. In the case of clinical risk prediction, all of them can be causes of concern and evaluation when investigating machine learning applications. To that end, here we propose the first-of-its-kind comprehensive evaluation review for peer-reviewed and published clinical risk prediction research which will comment on each of the paper's attempts or lack thereof of addressing the needs for explainability and what methods, validations, and open-access practices they followed in achieving the same specifically focussed on clinical applications. To help aid the evaluation of this research we define some guiding questions in reviewing the papers:

- Interpretability: Is there any attempt to comment on the most important features identified by the model during learning? Did the authors provide any analysis of the model learning so as to make the system more interpretable? What interpretability methods were applied and to what extend were they clinically validated?
- Fairness: Has bias been evaluated, diagnosed, visualised, communicated, or addressed?
- Bias: Did the authors comment on potential sources of bias for their system (including the data)? Are there any reflections on future risks of bias in deployment or applications of the proposed model and its envisioned work scenarios?
- Transparency: Have the authors consulted clinicians or other experts as well as patients at any stage of the project process? Have they commented on their needs, motivations, and limitations as well as constraints regarding the model and its applications? Do the authors document and communicate these findings in their work?
- Trust: Have the model results been checked for random perturbations and noise (including the interpretability results)? What statistical or power tests were completed and do the authors indicate confidence intervals or standard errors in their results?

While each paper might not be evaluated for these questions, we will show how a simple check for clinical, statistical, and reproducibility tests show the current culture of explainable AI research for clinical risk prediction. In the review, different modalities for clinical usage will be considered, and each section will correspond to a separate modality under which a table will be included containing the summary information for each reference. Each row of the table will correspond to a specific reference and the underlying model, interpretability approach, dataset size, topic, and the existence of a clinical validation or open access check. quantitative evaluation will correspond either to simple statistical tests done on interpretability results and checking for noise as well as using several interpretability methods to confirm consistency of explanations.

Clinical relevance is highly important in explainable AI for clinical risk prediction precisely because the ultimate goal of utilizing AI in healthcare is to improve patient outcomes and enhance clinical decision-making. Explainable AI models should not only be accurate and reliable but also provide insights and explanations that are meaningful and actionable for healthcare professionals. By emphasizing clinical relevance, explainable AI ensures that the generated explanations align with medical knowledge, guidelines, and the specific needs of healthcare practitioners. It enables clinicians to understand how the AI system arrives at its predictions or recommendations, thus building trust and confidence in the technology. This understanding allows clinicians to integrate AI outputs into their existing knowledge and expertise, resulting in more informed decision-making.

Moreover, clinical relevance in explainable AI helps in the adoption and acceptance of AI systems within the healthcare domain. Healthcare professionals are more likely to trust and

embrace AI technologies when the explanations provided by these systems are relevant and align with their clinical intuition. If the explanations are not clinically meaningful or comprehensible, clinicians may disregard or be skeptical of the AI system's recommendations, potentially leading to a lack of adoption and underutilization of valuable AI tools. Additionally, clinical relevance in explainable AI is crucial for patient-centered care. Patients and their families rely on healthcare professionals to make informed decisions about their health. When AI models provide explanations that are clinically relevant, it becomes easier for healthcare providers to communicate the reasoning behind their recommendations to patients. This transparency and shared decision-making process can enhance patient understanding, trust, and engagement in their own care.

Explanations have no performance guarantees and, as such, the fidelity and value of their explanations is often insufficiently investigated, but in few cases of external human validation [28]. Presenting a limitation in explainability work, inadequate external testing of explainability methods makes proposed explainable AI models for clinical risk prediction merely approximations to the model's decision protocol and not truly describe its underlying reasoning. Relying solely on using post-hoc explanations, as such, to assess the quality of model decisions could be just another source of error. For example, taking saliency maps which will be introduced later, true explainable decisions do not just reveal the answer to what part of the image the model was learning from but also whether it was reasonable that the model was looking in this region [55]. It is, thus, of high importance that while presenting an application or development of explainability methods to different modalities in clinical risk prediction, we also highlight their (non-)existent evaluations whether they be human in the form of clinical validation or algorithmic as the implementation of some quantitative metric. A relatively more common strategy that some could pursue is applying several different methods and comparing the explanations of each to guarantee either consistency or, under certain noise constraints, robustness.

## 2.3 Taxonomy of Methods

### 2.3.1 Inherently Interpretable Models (IIM)

IIMs or, how they are often termed, glass-box or white-box models due to their relative decision-making transparency are among the simplest methods to achieve explainable prediction in clinical risk modelling. Their transparency is usually a consequence of internal structure of the model itself being open to analysis and the relationship between input and output can be clearly represented either through equations or visualisations. Another key attribute for identifying IIMs is their lack of needing an external model or method to render the prediction model explainable [56]. Entire groups of these methods have been proposed for various applications throughout time with some, as in the case of linear regression, preceding the development of machine learning itself.

One of the most common group of methods included here are sparse (linear) classifiers. They include a collection of models from linear and logistic regressions to wider generalised additive models. In linear or logistic regression models, the output is modelled as a linear combination of the features or assuming a Gaussian distribution for the outcome conditional distribution thus making assumptions that might be violated in a non-linear or non-Gaussian setting. To overcome this, generalised linear models (GLMs) assume a non-Gaussian conditional distribution for the outcome whose mean would be a non-linear function of the same linear combination of features used in a standard regression. Generalised additive models (GAMs) do something similar but instead of changing the outcome distribution change the combination of features from a linear function to a combination of arbitrary non-linear functions of each feature [57]. A summary equation for this set of methods can be seen below:

$$g\left(E_Y(y \mid x)\right) = \beta_0 + f_1\left(x_1\right) + f_2\left(x_2\right) + \ldots + f_p\left(x_p\right) \tag{1}$$

$g$ is often referred to as a link function which helps transform the output for certain problems (classification or regression). $E_Y(y \mid x)$ is the conditional distribution for the outcome that is either assumed to be Gaussian (linear regression) or non-Gaussian (GLM). $\beta_0$ is the intercept while the functions are either numeric weights in the case of a linear combination of features (weighed sum) or arbitrary functions that can be non-linear (GAMs). By extracting the weights of the fitted model which represent relative weighing (or feature importance), one can obtain a quantitative and definite interpretation of the model's predictions [57]. A limitation to interpretability for these methods could be the choice of link function which breaks the simple relationship between the weighted sum of inputs and the output as well as relying on GAMs implying that interpretability is no longer as simple as using a single weight for a feature contribution to prediction. These disadvantages are often overcome through the use of decision tree and random forest models while achieving superior predictive performance.

Before moving on to discretisation methods that address some of the limitations of linear models, an important aspect of sparse linear models is, in fact, their sparsity. Using a large number of features makes prediction problems less tractable and the models prone to overfitting (even when regularisation methods are applied), and having to interpret too many features reduces readability of the explained results [58]. Supersparse linear integer models (SLIMs) have been developed to address these challenges and tested on medical applications. SLIMs use highly predictive integer scoring mechanisms for risk with a smaller number of features than regularised linear models. They "use a linear form that helps users to gauge the influence of each input variable with respect to the others" while encoding interpretability requirements through operational constraints that can be represented in a rule-based format [59]. The way it achieves this is through a balanced loss between higher prediction accuracy and higher sparsity while using restricted and discrete coefficients $\lambda$ for sample $X_i$:

$$\begin{array}{ll} \min_\lambda & \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[y_i \boldsymbol{\lambda}^T X_i \leq 0\right] + C_0\|\boldsymbol{\lambda}\|_0 + \epsilon\|\boldsymbol{\lambda}\|_1 \\ \text{s.t.} & \lambda \in \mathcal{L}. \end{array} \tag{2}$$

SLIM minimises the $0-1$ loss $\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[y_i \lambda^T X_i \leq 0\right]$ and $\ell_0$-norm $\|\lambda\|_0 := \sum_{j=1}^{P} \mathbb{1}\left[\lambda_j \neq 0\right]$. Coefficients are restricted to $\mathcal{L} = \{-10, \ldots, 10\}^{P+1}$, and can include operational constraints like $\|\lambda\|_0 \leq 10$. SLIM uses a regularisation $\ell_1$ factor $\epsilon\|\lambda\|_1$ in the objective to restrict coefficients to coprime values. The parameter $\epsilon$ is set to a negligible value to avoid $\ell_1$-regularization effects [59].

Discretisation methods reduce a problem and its data into a set of discrete subsets thereby transforming the continuous solution space into one defined at discrete points. Decision trees and their ensembles resort to this strategy when learning by evaluating each feature value and selecting the cut-off to maximize the class separation as measured through entropy or Gini impurity (classification) or minimize the variance of the outcome (regression). At each step of the partition of the feature, a decision gets made to split the partition further until the leaf nodes are reached and instances are classified based on the average outcome in that leaf node. Due to the discrete nature of the decision-making, these methods provide straightforward intepretability frameworks. One just has to follow the splits for each feature and evaluate how much it has reduced the variance or Gini index. Thus, each importance can be interpreted as share of the overall model importance [60]. With its simple structure, it arguably provides a more intuitive interpretability than sparse linear models whose outcomes are products of several layers of possibly non-linear functions in a multi-dimensional space. A limitation is that in cases of linear relationships between input and output, it approximates the behaviour through a series of discrete step function which might not be of high fidelity. Lastly, these methods have been shown to be unstable by changing their feature importance drastically based on small changes in the feature space due to the possibility an entirely different tree emerges from slight

changes in the earlier and parent nodes.

Another example of discretisation methods are rule-based or expert-based systems, one of the earliest examples of domain adaptation of 'machine learning' to healthcare practice [61]. These are methods that, similarly to decision trees, break down the outcome modelling into a series of discrete steps or rather decisions of an IF-THEN nature whose cut-off criteria are not determined by optimising a certain metric but from domain expertise or, in the case of Bayesian rule lists, from data mining the features for patterns [62]. As long as the decision rules are understandable, these models are radically interpretable, simple, and often times more compact than their decision tree cousins. Through domain expertise, they also apply a certain type of feature selection by *a priori* defining what features are of importance from expert knowledge and only including those in the decision rule structure whereas linear models would assing weights to all features and learn a less sparse model. Similarly to decision trees, they are not adapted to linear modelling problems while, additionally, only being able to solve classification problems with discrete features.

As compared to the previous methodologies, k-Nearest Neighbors provides an example- or instance-based interpretability whereas the earlier models account for global feature importance. Since it uses no modular learning of any global structure to make predictions there are also no global weights or decisions computer to interpret in the first place. The algorithm, in the case of classification, classifies the instance depending on the most common class in the neighbourhood of the data point in the feature space. Neighbourhood then depends on the number of neighbours and the distance metric one uses to identify the nearest k neighbours. To explain a specific prediction, one relies on the nearest neighbours used and their common feature patterns. Some work has used linear models to automate finding these patterns once clusters are formed [63]. This implies, however, that in cases of higher dimensionality, seeing those patterns becomes intractable and, consequently, uninterpretable.

Static or tabular data are not the only type of data encountered in clinical machine learning applications. Often, images or timeseries require a different *modus operandi*, and deep learning, a machine learning paradigm relying on neural networks with a large number of layers (convolutional, recurrent, or otherwise), has been a great fit for these data formats and corresponding problems. Traditional machine learning models require feature engineering to create the input vectors from data, especially when they are images, whereas deep learning, through representation learning, allows the method to learn from raw data directly the optimal representation of internal features for learning [64]. When learning from images or time measurements, these models look for spatio-temporal patterns in the data, and, as such, the interpretability of these methods is a different and more complex challenge. Causality and fuzzy logic can help make combined systems with deep learning *inherently* interpretable. While it might seem that adding fuzzy logic together with deep learning to make it more explainable is an external addon rather than an approach for IIM, it can be argued that the combined system is now a new method on its own which does have inherent explainability characteristics. This is the approach taken here. Traditional machine learning methods have seen their own implementations in the fuzzy logic realm but the focus here is on a more interesting approach of transforming inherently *un*interpretable deep learning otherwise by fuzzy rules. For decision tree, support vector machine, and nearest neighbor fuzzy implementations, consult the following literature: [65, 66, 67, 68, 69]. The assumption in fuzzy logic is that instead of assuming input or output is deterministic, i.e. fixed value, one assigns uncertainty to both ends of the learning system. An example would be the decisions that a human makes while driving a car as provided by [70]. The decision-making of "if the distance to the car ahead is less than 2.5 m and the road is 10% slippery then reduce car speed by 25%" is approximated with numerical uncertainty reflecting the imprecise language of *If* the distance to the car ahead is low and the road is slightly slippery *Then* slow down. The numerical meanings of "low", "close", and "slow down" can vary and are, thus, fuzzy.

Breaking down the inference problem into fuzzy logic if-then steps aids interpretability and can be combined with high-performing opaque deep learning models through, for example, a fusion layer at the end of a neural network, one part of which is deep representational learning and the other having fuzzy rules [71]. The weights in the fuzzy part of the networks are themselves fuzzy and can aid learning the joint probability distribution which, consequently, improves prediction performance. Another approach is to switch between neural and fuzzy layers or add the fuzzy logic system as a module with autoencoder structures through pre-trained layer replacement, often referred to as a deep type-2 fuzzy logic system (D2FLS) [72, 73]. Whatever the strategy, using fuzzy if-then rules holds great promise in overcoming the black-box nature of deep learning while keeping the superior predictive performance. There are limitations, however, as to how well suited these methods are for interpretability itself. Since they were built for uncertainty integration, they have not been yet optimised for explainability applications especially in higher dimensions [74]. Future research can help address these concerns by implementing these methods in different case scenarios from static data processing to image classification and prediction while addressing the higher dimensionality through possible grouping of features.

The topic of using logic-based rules to interpret complex models without external add-ons necessarily also includes work on boolean rules generation with column generation methods (BRCGs). These methods attempt to reduce binary classification problems into a set of compact Boolean clauses similar to the ones seen earlier in rule-based models and fuzzy logic but with a set structure of more compact and easier to explain rules [75]. Each conjunction in these unordered rule sets is considered an individual rule and a positive prediction occurs when at least one of the rules is satisfied. These rules can be mined from the data (with decision trees for example) after which they can be selected through a myriad of methods including integer programming albeit which often comes with a limited search space [76, 77]. Rule selection through column generation searches over an exponential number of all possible clauses optimised in a greedy manner with the combined reduced cost objective. This approach also allows a quantitative way of measuring intepretability by just looking at the size of the rules set itself. In cases of higher dimensionality, this approach becomes less tractable and is addressed by formulating an approximate Pricing Problem by randomly selecting a limited number of features and samples [76]. The classification accuracy and computational costs for these binary classifiers are, however, a large limitation to their widespread acceptance as general IIM frameworks. Recent work has attempted to use optimised column generation to aid with the computational constraints and has included fairness into the objective evaluation through equality of opportunity and equalized odds constraints [78].

Besides relying on fusion layers and rule-based methods to turn opaque models inherently interpretable by some type of combination, another example that should be investigated more is incorporating transfer learning into explainability. The concept of transfer learning rests on pre-training or learning weights for one task and 'transferring' that knowledge and the learned weights to another problem, sometimes with a limited amount of re-learning and often to a different domain [79]. By using transfer learning to preserve the high predictive performance of black-box models and transferring them to interpretable simpler models a more balanced trade-off between deep learning and transparency can be obtained. An abstract representation of this process can be seen in Figure 2. The limitation, however, is on how transferrable the features learned from the complex model are to the simpler model [80]. Assuming sufficiently, through this approach, one can possibly simultaneously use a simpler and more reliable model that clinicians would be familiar with while preserving higher performance, apply a model trained on a large amount of data to a smaller dataset thereby avoiding overfitting, and reducing computational costs by simply 'cut-and-pasting' a pre-trained model to the problem at hand [20, 81]. All of these can help achieve more explainable systems that are not just intepretable, but easier to check for fairness and bias, especially considering the smaller and more accessible datasets that can now be used for deployment.

Methods like ProfWeight transfer the learning process from a pre-trained deep neural network

with a high predictive capability to a simpler interpretable model or a very shallow network of low complexity and a priori low test accuracy. The method relies on flattened intermediate representations used to generate confidence scores of linear probes for weighting of samples during training of the simpler models. This approach has not yet been sufficiently evaluated in real-world applications such as clinical risk prediction but could hold great promise in achieving a stable balance between higher test accuracies and white-box models [82]. Possible extensions could include new and more challenging data domains and applications, experimenting with a combination of different models of varying complexity, and optimising for inference time and reducing costs since the initial weights still need to be computed from a complex deep learning model.

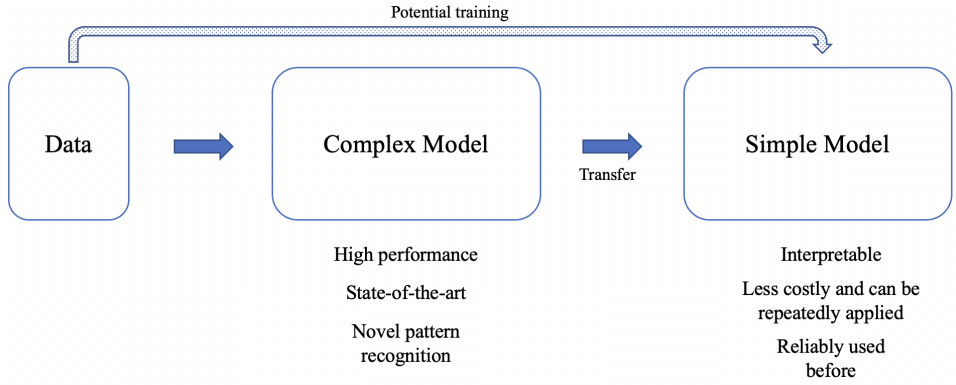


Figure 2: Using transfer learning to optimise predictive performance of complex models including deep learning and transformers and combine them with simpler interpretable models that can be more reliably presented in clinical care

Inherently interpretable models also have their limitations when their explanations are often burdened by possible and unidentified confounders. Recent work on human agency with explainable AI has found that the added user reliance on using an IIM with less features and cleared decision-making is not significant. Additionally, researchers found that using an IIM limited users' "abilities to detect when the model had made a sizable mistake, seemingly due to information overload caused by the amount of detail in front of them" [83]. A key factor in usability then is the number of features included in the model itself. The tendency for outliers to break the explainability framework is to be addressed through collaborative explainability mechanisms between human users and automated systems by inviting users to give their own predictions before seeing the model. In this way, users might be more likely to notice any outlier values. Thus, while users might be more likely to rely on simpler and clearer models, the benefits to prediction accuracy are negligible, meaning that using IIMs with clearer features of a smaller number has first and foremost a psychological effect in building trust even if not a practical one.

Recent advances, however, have been in the field of explainability for models and methods that do not fit the IIM criteria. Large, complex, and often opaque models that deal with learning patterns from text, images, genomics, and timeseries measurements present unique challenges when applied to clinical risk score prediction. Often times, separate explainability methodologies of similar complexity as the models they purport to explain have to be developed and tested. That will be the focus of the next section.

### 2.3.2 Ante-hoc vs. Post-hoc

We have previously described several paradigms and methods underpinning inherently interpretable models in AI, including classic models in machine learning, such as decision tree and rule-based models. Often, because these models do not require any post-prediction or post-inference interpretability, they get classified as ante-hoc methods. In cases of models that are not IIM, such as deep learning models or large ensembles, a post-hoc explanation framework needs to be applied after the fact, post-model-learning that is separate from the model itself. Under the post-hoc category, we can define two subroups, those post-hoc methods that only work for specific data modalities or models, and model-agnostic explainability methods. The post-hoc approach attempts to address the trade-off between high interpretability and high predictive performance by providing methods to add-on interpretability to complex but high-performing models. Thus, instead of developing an IIM which, as we have seen, suffers under certain constraints either as part of the assumptions or model prediction capabilities, post-hoc explanations can be paired with a wide swath of complex AI models to extract interpretability without understanding the inherent reasoning of the model itself.

There has been recent work on making deep neural models ante-hoc interpretable including approaches like self-explainable neural networks (SENNs) and CoFrNets. SENNs rely on staggered generalisation of linear models as a function of locally interpretable functions and representations of the input. The example of the model structure can be seen in Figure 3 [84]. So far, SENNs have only been tested on a few tabular and extracted imaging data. CoFrNets, on the other hand, rely on the mathematics of continued fractions (CFs), typically represented as a ladder-like sequence: $a_0 + \frac{b_1}{a_1 + \frac{b_2}{a_2 + \cdots}}$ which can represent any real number and any analytic function [85]. Using the repeating structure and linear activations in between layers (which can be shown are sufficient to achieve universal approximation), the input is passed to each layer and a linear number of weights is learned for each layer instead of the quadratic amount in classic neural networks [86]. Per-example feature importance is then computed using the gradients throughout the ladder as a function of the input and global attributions taking advantage of the ladder being a representation of a multivariate power series with the coefficients of the two forms mapping one-to-one. Further work is needed, however, to establish the high predictive performance compared to other black-box deep learning models.

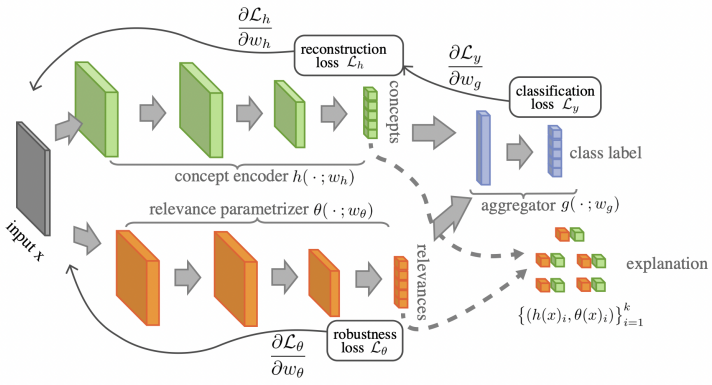


Figure 3: SENN components: a concept encoder (green) that transforms the input into a small set of interpretable basis features; an input-dependent parametrizer (orange) that generates relevance scores; and an aggregation function that combines to produce a prediction. The robustness loss on the parametrizer encourages the full model to behave locally as a linear function, yielding immediate interpretation of both concepts and relevances [84]

Both of the previous methods touch upon conceptual learning, that is, when deep learning

models learn from data, they, in fact, extract key concepts important for making the predictions. If one could extract these internal learning concepts, then one would be a lot closer to understanding the internal reasoning of the models, thus making them more explainable. Recently published work builds on this approach by appending a concept generation module (encoder) to jointly optimise for interpretable concepts on top of a latent encoder used for image classification. The concepts are then passed to a decoder to estimate reconstruction error and thus provide an indirect way to measure how "interpretable" the concepts are by how well they reconstruct the input. The concepts learned are individually informative by enforcing a fidelity loss. One does not have to rely on the concepts being directly supervisable (i.e. ground truth annotations are not available) because using self-supervision through incorporation of a separate loss function as an auxiliary task allows extension to more learning problems. Even though the framework adds components to existing deep learning models, most can be discarded after training and explanation-generating module can be used at prediction time [87]. More experiments need to be done to show good explainability and predictive performance on datasets with varying sizes, including those that are not just image-based, but the conceptual learning paradigm offers a motivating approach to ante-hoc explainability of black-box neural networks.

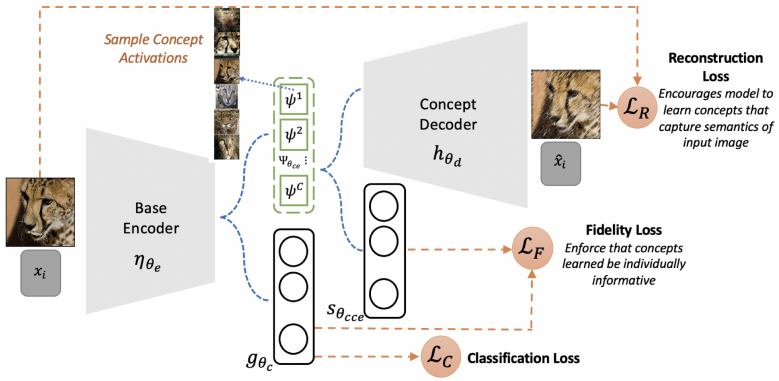


Figure 4: Conceptual learning modular framework for appending explainability to deep learning models [87]

A more popular field of inquiry has been post-hoc methods which provide interpretability and limited explainability after model training, inference, and deployment. These methods can be model-specific or model-agnostic and more will be said on this distinction further in the text. For now, it is sufficient to mention widespread post-hoc explainability methods that have made great inroads in clinical risk prediction applications. These include feature attribution (global and local) methods like Shapley values which adapt concepts from game theory to investigate feature importance and contributions to predictions. A Shapley value is calculated for each feature $i$ with feature value $X_i$, usually based only on the test set for risk prediction, by using sets of all possible unions with $n$ features except feature $i$. The value is the difference between the results of the characteristic function $v$ on $N$ (the set of all features) and $S$ (the subset of $N$ without feature $i$). The Shapley value is then averaged across the marginal contributions of all possible combinations of the feature unions:

$$\varphi_i(v) = \sum_{S \subset N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} \left( v\left(S \bigcup \{X_i\}\right) - v(S) \right)$$

Shapley values are simultaneously useful for both finding the contribution of each feature to an individual prediction as well as providing global explainability across all samples [88, 89]. Calculating the value itself is usually computationally expensive even on small test datasets, at which point kernel sampling estimators are used to summarise the process. By defining a

specific kernel in KernelSHAP allows for a smaller number of necessary samples by a combination of sampling and penalized linear regression. The deep learning extension to the method, DeepSHAP, relies on layer-wise composition of approximate Shapley values which [90] propose is approximated through uncertainty propagation in polynomial time [91]. Furthermore, the method is adaptable to a variety of data modalities making it widely useful across clinical applications. The method does not, however, take feature dependence into account, especially in its kernel version leading to over-weighing outlier data points. The tree version of SHAP addresses this but its reliance on conditional expected predictions is known to produce noise as non-intuitive feature importance values [60]. Another assumption or constraint of the method is using the same weight for all marginal contributions, thus giving the same importance when a large number of features are given versus a small number of features. Recent work like WeightedSHAP generalises the Shapley value by relaxing the efficiency axiom and learns which marginal contributions to focus on directly from data and their relative signal to a prediction. It has shown more robust performance on identifying the most influential features of various grouping sizes. Future work could include domain adaptation of the weighting approach and more testing on various data modalities and scenarios.

A slightly earlier proposed method, local interpretable model-agnostic explanations (LIME), focuses only on explaining local or individual predictions. LIME relies on introducing perturbations on the training data and observing how they affect the corresponding predictions given an interpretable model like some of the IIMs mentioned earlier all the while approximating the black-box model in the neighborhood of the sample [53]. In applications to clinical risk prediction, the method has been found to produce unstable feature importance due to random perturbations. [92] propose DLIME (D standing for deterministic), which uses hierarchical clustering to group the training data together and kNN to select the relevant cluster of the new instance being explained instead of relying on randomness. After finding the relevant cluster, a linear model is trained over the selected cluster to generate the explanations. Evaluated on clinical risk problems, DLIME presents a more stable version of instance-wise interpretability than its predecessor.

LIME and SHAP present foundational interpretability methods due to their widespread adoption and relatively clear heuristic. Local Interpretable Visual Explanations (LIVE) build on the LIME paradigm by prioritising the visualisation of the interpetability results while keeping the random data perturbations approach and local point variation. While LIVE does not require an interpretable input space since it approximates the black box model directly from the data, it has no theoretical foundation like Shapley values. BreakDown uses a non-local approach to post-hoc, model agnostic interpretability that permutes the data samples based on relaxed loss effect constraints instead of relying on random sampling [93]. The space of features is searched to look for a set of variables where the model prediction will be close to the average expected value across all model predictions. To reduce the computational complexity of the search, a greedy approach is taken. These methods, however, decompose final prediction into additive attribution components and usually do not work well for models with significant inter-feature correlations.

Manipulating feature input and evaluating its effects on the outcome is a common set-up of many interpetability methods. The contrastive explanations method (CEM) finds what features should be minimally and sufficiently present to justify its classification and analogously what should be minimally and necessarily absent. This method was evaluated on deep learning models for applications including brain activity prediction but generating explanations of the form "an input $x$ is classified in class y because features $f_i, \cdots, f_k$ are present and because features $f_m, \cdots, f_p$ are absent" [94]. As compared to other methods like LIME, CEM highlights what features need to be present for a specific classification and not just list positively or negatively relevant features that may not be necessary or sufficient to justify the classification. Finding this set of features is defined as an optimisation problem where the modified data is

sampled close to the original by using reconstructions from an autoencoder. These types of explanations might be more intuitive to many clinical applications especially in diagnosis as highlighted by [95]. Finding the contrastive explanations can be aided using an adversarial model, for example, but the method's limitations of defining what contrastive examples would represent in complex data cases like timeseries has not been sufficiently explored. An image pixel can be made into a constrastive case by being set to 0 but greyscale examples would encounter challenges while a similar approach for timeseries would not be robust [96].

Post-hoc methods are generally model agnostic and apply to a myriad of black box models by directly manipulating the input either through finding a subset of data points or examples that are more "informative" or explainable to the prediction or a subset of the feature space that does the same. ProtoDash is a method that generated prototypes from the data and assigns non-negative weights to each of them depending on their importance to the prediction problem. Since each set of prototypes can be weighted, ProtoDash offers a robust response in cases of class imbalance which have been shown to affect the fidelity of earlier interpretability methods [97]. The method also addressed covariate shift between datasets because it not only finds prototypical examples for a dataset $X$, but it can also identify (weighted) prototypical examples from $X^{(2)}$ that best represent another different dataset $X^{(1)}$ [98]. Effectively, ProtoDash is an instance-based interpretability method as the prototypes it samples are examples of samples that would be more important for a prediction rather than features as such. Other earlier instance-level methods like K-medoids and MMD attempt to do the same but ProtoDash provides more general rule explanations with more robust kernel performance that help identify cases of low test accuracy while being relatively constant in its interpretability [99, 100].

So far we have only discussed methods that manipulate the data space in some form to evaluate feature or example importance for prediction, but other post-hoc paradigms exist that more directly measure model learning. In deep learning, computed gradients are used to update weights through the learning procedure. In cases such as images or timeseries with spatio-temporal dimensionality, visualising the absolute weights of the gradient updates across those dimensions can indicate what part of the example image or timeseries is more influential for a prediction averaged across all examples [101]. Examples include using visual saliency maps to highlight sections of ECG signal important for arrhythmia classification or x-ray images with lung cancer detection which can aid in learning while providing clinicians with more explainable model predictions [102, 103]. These saliency maps are often sensitive to noise in the input and several approaches have been proposed to address these limitations including by adding noise to the data multiple times and then taking the average of the resulting saliency maps for each example like with SmoothGrad [104]. Newer methods like saliency guided training have features iteratively masked with low gradient values and then minimise a combined loss function of the KL divergence between model outputs from the original and masked inputs, and the model's own loss function. Experiments have shown the latter maintains higher predictive performance while simultaneously reducing noise sensitivity [105]. General limitations of saliency maps in clinical risk prediction have been their tendency to represent clinically non-useful information as important as evaluated by physicians and simply localising the area of the example does not add further intepretability on why that area is important for model learning which can appear misleading to domain experts [106]. They have also been vulnerable to adversarial attacks when explanations remain unchanged even after significant adversarial perturbation in the input has changed model behaviour [107].

Visualising weights from the learning process to highlight which aspects of the data the model focuses on for training is the underlying heuristic of attention mechanisms which became a highly popular deep learning model-specific method to achieve better predictive performance while providing more interpretable results. An attention function maps a query and a set of key-value pairs where the output is computed as a weighted sum of the values and the weights are computed by a compatibility function of the query with the corresponding key. Multi-head

attention parallelises the attention layers to project multiple mappings simultaneously. A comparison of the two frameworks can be seen in Figure 5. Compared to recurrent layers for learning sequences of data requiring $O(n)$ operations, a self-attention layer connects all positions with a constant number of sequentially executed operations, thus being computationally cheaper. In the example of timeseries data, the first step is to initialise the attention vectors mapped to the features such that for each feature $i$ an attention vector $a_i$ of length $T$ (number of time-steps) is learned with $|a_i| = 1$:

$$X_{\text{new}} = A \odot X$$

where $X$ represents the $T \times m$ input data ($m$ number of variables). Once sofmax normalisation is applied, $a_{it}$ can be interpreted as contribution of feature $i$ within a fixed time step $t$. [108] extend this for global interpetability by applying the softmax to the transposed input with the same notation:

$$a_t = \text{softmax}\left(x_t W_t\right)$$

By aggregating values $a_{i1}, \ldots, a_{iT}$ through time one can get the global contribution of the $i$-th feature with feature value $x_t$ and weights $W_t$ at time $t$. Not only is an absolute ranking of features provided, but one can use the softmax activation matrix to extract patient-level feature importance as well as visualise the attention weights using heatmaps (attention maps) both in timeseries and imaging problems, similar to the saliency maps scenario despite early criticism of the lack of explainability of such maps for standard attention mechanisms [109]. Other criticism of attention mechanisms not being sufficiently interpretable come from their propensity to highlight less meaningful tokens in text problems and lack of support from gradient-based methods. Whether attention weights actually capture feature importance as such was addressed by [110] who show that combinatorial shortcuts could be a source of this problem where attention weights themselves could be carrying extra information used by downstream layers instead of just the most important aspects of the input. They propose using a combination of random attention layer pretraining with mask-neutral learning with instance weighting where instead of fitting a biased expectation one can recover a mask-neutral distribution unrelated to the output. The downstream parts of attention layers become less biased, and combinatorial shortcuts can be partially mitigated. Some results do indicate better interpretability of the attention plots but further experiments are still needed to evaluate the fidelity of this approach in achieving the proposed changes in interpretability.

Since being originally proposed in 2017 as an alternative to convolutional and recurrent behaviour in deep learning pattern recognition, transformers (models based solely on attention mechanisms) have become robust prediction models in various applications for clinical risk prediction, including ICU outcome prediction, ECG signal diagnosis, and blood pressure response [111, 112, 113, 114, 115]. Some limitations include inability to distinguish between positive and negative associations among features and output, lack of propagated relevancy through the attention layers partially addressed through a new layer-propagation strategy, and unstable interpretability with covariate shift [116, 117].

We have seen in this section a combination of methods that try to address the interpretability aspects of explainability using both innovative ante-hoc glass-box approaches to deep learning as well as what have become off-the-shelf methods like LIME and SHAP. Most of the methods share limitations on their robustness to random noise, permutation and covariate shift, as well as a lack of clinical evaluation of their inherent interpretability qualities.

### 2.3.3 Global vs. Local

Another axis of separation for explainability methods is on the global (feature-space) versus local (sample-space) levels. A local or instance-based (used synonymously for some methods) explanation deals with explaining the prediction of an individual sample, whereas a global

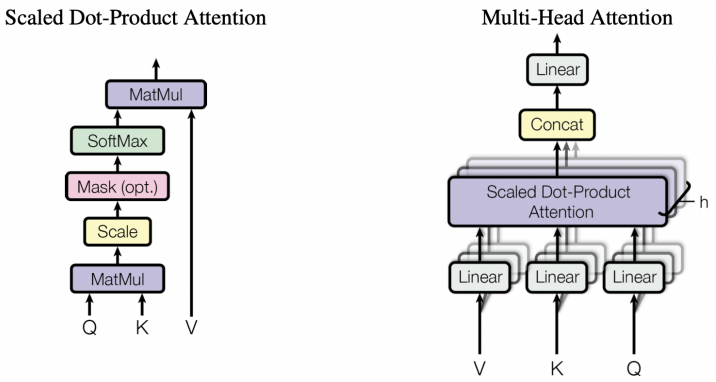


Figure 5: The standard attention function calculation is on the left while multi-head attention with its parallel structure is on the right [111]

explanation attempts to provide a summarised view of the entire model. In the clinical risk prediction realm, both can be valuable despite the latter being more popular in recent years of research. Local explanations can be used for explaining individual patient trajectories and predictions all the way to their grouping in phenotype clusters, for example, while global explanations can give one a more holistic view of the impacts of certain risk factors on a population-size patient cohort [29]. Local interpretability methods usually provide more stable and linear results for explanations which is partially why methods in deep learning like guided backpropagation, Grad-CAM, and SmoothGrad saliency maps (described in gradient-based section later on) have remained popular over time.

From the interpretability methods covered earlier, LIME and SHAP also support local explanations by providing a result for each instance in the sample space. The instance is then perturbed according to different criteria in the case of LIME while SHAP uses a more robust generation of individual explanations because of its guarantee of a fair distribution effect among the features coming from a more established theoretical background for the latter method. SHAP provides high efficiency of feature values computation especially in the presence of multi-collinearity among the features while providing a theoretical framework to prove, under certain assumptions, the unique existence of such local models [88]. An extension of LIME which resolves the significant dependence on weights assigned to different perturbed samples is ILIME (I standing for influence-based). ILIME uses the influence of perturbed instances on the instance of interest as well as their Euclidean distance to estimate the weights with influence functions [118].

CEM also supports local interpretability as does ProtoDash which produces as an explanation representative instances or prototypes from the data that estimate the underlying distribution. As mentioned earlier, ProtoDash generates these samples by assigning non-negative weights of importance. The generation depends on the property of sub-modularity of its scoring function to find samples efficiently. The computational cost of such guarantees are significant, thus a class of approximate sub-modular functions are used [119].

MAPLE or model agnostic supervised local explanations method uses random forests to achieve neighbourhood selection in local linear modelling. The neighbourhood for each instance is defined by each instance's occurrence in the same leaf node as the instance to be explained. A score is then assigned to each feature based on its frequency of splitting a node at the root of the trees in the forest. The set of most important features is then used for approximating the linear regression coefficients in the final explanations. The great advantage of MAPLE is that it uses random forests which suggests better predictive performance while still being interpretable [29, 120].

Counterfactual explanations mentioned earlier in the introduction as a relatively recent proposal for a more general and abstract way to explain models while addressing fairness and bias concerns also can be classified as a local interpretability method. Counterfactual explanations instruct how an instance can be changed to alter a prediction. By generating counterfactual instances, similar to prototypes, we can explore the inner workings of a model and explain instance-level predictions. Unlike prototypes, however, these samples do not have to be existing in the training data. Counterfactuals rely on conceptual thinking of hypothetical situations that go against the established facts. In practice this sums up to perturbing the feature values of a sample before making the predictions and analysing how the prediction changes. In short, "a counterfactual explanation of a prediction describes the smallest change to the feature values that changes the prediction to a predefined output" as explained by [60]. Creating these explanations for machine learning models has been first proposed by [121] who suggest minimising the loss:

$$L\left(X, X', y', \lambda\right) = \lambda \cdot \left(\hat{f}\left(X'\right) - y'\right)^2 + d\left(X, X'\right)$$

The loss consists of the quadratic distance between the model prediction for the counterfactual $X'$ and the desired outcome $y'$ and the distance between the sample $X$ to be explained and the counterfactual $X'$. $\lambda$ is a parameter over which one maximises by iteratively solving for $X'$ and increasing $\lambda$ until a sufficiently close solution is found. The distance function $d$ is the Manhattan distance weighted with the inverse median absolute deviation (MAD) of each feature $i$ in feature set $n$.

$$d\left(X, X'\right) = \sum_{i=1}^{n} \frac{|X_i - X'_i|}{MAD_i}$$

The total distance is the sum of all absolute differences of feature values between sample $X$ and counterfactual $X'$. The usual advantage of Manhattan distance over Euclidean for robustness to outliers stands as usual.

A potential algorithm for computing counterfactuals based on [60] is:

1. Propose a sample $X$ to be explained and the desired outcome $y'$.
2. Sample a random sample as an initial counterfactual.
3. Optimize the loss with the initially sampled counterfactual as starting point.
4. While $\left|\hat{f}\left(X'\right) - y'\right| > \epsilon$:
    - Increase $\lambda$.
    - Optimize the loss with the current counterfactual as starting point.
    - Return the counterfactual that minimizes the loss.
5. Repeat steps 2-4 and return the list of counterfactuals or the one that minimizes the loss.

The method does not handle categorical features with many different levels well. The authors of the method suggested running the method separately for each combination of feature values of the categorical features, but this will lead to a combinatorial explosion if you have multiple categorical features with many values. For example, six categorical features with ten unique levels would mean one million runs.

One of the limitations of counterfactuals generally is their lack of uniqueness, i.e. multiple counterfactual explanations can be provided for the same prediction. The method above cannot efficiently handle categorical features with many different levels since it depends on running the method separately for each combination of feature values. An approach was proposed by

[122] using four-objective loss that allows for multi-category counterfactual optimisation with a nondominated sorting genetic algorithm. The approach, however, still suffers from the general limitations of non-uniqueness. One can even consider adversarial examples as a sub-type of counterfactuals i.e. counterfactuals that are meant to make the model give false predictions. Doing so can help in fact unpack how a model makes decisions without explicitly providing an interpretability framework. Some approaches include exploiting the duality between adversarial counterfactuals and explanations to compute one from the other like those proposed by [123]. Extended work on applications in healthcare like for glaucoma detection have shown better explainability than classical methods like saliency maps [124].

K-nearest neighbours (kNN) is a popular distance-based machine learning method which assigns the class of each point based on its k closest neighbours. Because it is a nonparametric model using all of the available data to make a prediction for each sample, it is in-effect an instance-based interpretable machine learning model. The lack of learned parameters implies no global interpretability as such but using the k neighbours that were used for the prediction for each sample, one can provide limited interpretability for that prediction. This becomes true in cases of constrained feature space where heuristic or simple analysis of neighbouring points' features is manageable but not in cases where a large number of features is used [60]. This capability was in an indirect way used in the methodology for DLIME when the 'important' and 'explainable' clusters are identified using kNN in combination with hierarchical clustering first used to generate an intermediate clustering stage [92]. Further work on integrating local explainability propensity of kNN methods in combination with other methods like, for example, clustering Shapley values in local regions around a specific sample have not been sufficiently studied. They might provide an avenue in addressing instability in local explanations due to random perturbations as was done in the case of DLIME.

Other methods mentioned earlier like ProfWeight, Boolean rules, CoFrNets, and others all provide, as of now, global explainability exclusively which means they provide feature importance and explanations across the entire sample set without resolution to the sample level. One can then say that such methods provide an average estimate explanation for the model and the data and are useful in understanding general behaviours of the system.

### 2.3.4 Model-agnostic vs. Model-specific

The division of explainability methods also includes specifications of the levels of model nature the methods can apply to. Some methods are model-agnostic, meaning that they can be used to provide interpretable results for a variety of models. This is the case for Shapley values, for example, which can be used to provide interpretability for extreme gradient boosted decision trees, deep learning models, as well as some clustering algorithms [125]. An early prototype for model-agnostic explainability is leave-one-covariate-out (LOCO) in which one uses the prediction error of removing one covariate and comparing it against using it. The prediction error is estimated using confidence intervals for each possible value of the feature's domain. If this range is above significance levels, then the feature is decided to be important [126, 127]. A more general method for achieving something similar is permutation importance because, as we mentioned, this method relies on manipulating the input feature sets rather than having anything to do with explaining the internal model dynamics, so it can be applied across different machine learning models as long as a relatively distinct subgroup can be formed for the feature space [128]. An obvious problem with these approaches is the computational cost explosion resulting in cases of higher dimensionality.

A subgroup of model-agnostic methods can be described as visualisation methods because of their focus on creating curves and plots describing the output's associations to input features after prediction. Examples like partial dependence plots (PDPs) rely on using a partial dependence function of the output with respect to subsets of the feature space. Values in the

input space are varied slightly with respect to their marginal distribution with the partial dependence function approximated by a statistical model from which plots of associations between covariates and output are created [129]. The assumption is that because this is done for a single predictor at a time, that predictor is relatively independent of the remaining set of features on average which might not stand. This limitations combined with PDP's weakness of dealing with extrapolations in the feature space have led to the introduction of individual conditional expectation (ICE) plots. In ICE, instead of plotting the feature average partial effect on the output, one plots the estimated conditional expectation curves, each reflecting the output as a function of the feature effectively disaggregating divergent effects. The PDP curve in the plots is then simply the average of N ICE curves behaving as a local interpretability model that plots a curve for each instance of the data [130]. Some limitations of ICE plots include their overcrowding due to curves created for each sample instance in the plot, difficulty detecting outliers and overfitting extrapolations from view. Furthermore, the complexity of a higher feature space makes this method practically infeasible. Some recent advances like ICE feature impacts address these by allowing comparisons between different ICE plots and defining a metric, takeing into account all samples as well as outliers over which feature contributions are averaged to identify the most impactful features (not the same as PDP where the curves and expectations are themselves averaged for each covariate contribution), thus severely decluterring the plots [131, 132].

In some cases there is additional flexibility in the types of explanations generated as well. Instead of getting just feature importance rankings or plots, some model-agnostic methods can provide other types of explanations in parallel like linear formulations or graphic interfaces [133]. The great benefit of this approach is the ability to use a variety of complex, high performance, and uninterpretable models which can be rendered more interpretable without having to make sacrifices on performance by choosing simpler and more transparent models.

Other model-specific methods tend to only work on a certain subset of machine learning models and methods. Cases include saliency maps developed for elucidating image regional weigting in the learning process of deep learning models similar to attention maps. Another example described earlier includes generalised linear rule models (GLRMs) which learn a linear combination of conjunctions using link functions like logit in the case of logistic regression but which only work on this set of model respectively [134]. Depending on the intended usage, these two approaches offer different advantages. If the user in clinical risk prediction wants to focus on more predictive and high-performing models, then model-agnostic methods often can aid in final interpretability design if existing methods have not been proposed and sufficiently evaluated for that model before. On the other hand, some might take the route of choosing a simpler yet more explainable model which they can couple with a model-specific method to provide more clinical insight and explainability all the while giving satisfactory predictive performance. In fact, it is possible to take the approach of using different intepretability methods and evaluating their comparative explainability and performance [135]. There is no one size fits all for this type of explainability challenges and the different context for each problem should be investigated robustly before the implementation stage.

### 2.3.5 Perturbation- vs. Gradient- vs. Instance-based

Some, albeit rare, classifications of explainability methods include their grouping based on how the predictions and associated interpretability results are calculated with respect to the pairing of input and output. Such groupings include gradient- and perturbation-based (sometimes confusingly called instance-based) approaches. Perturbation-based methods perturb the model around the prediction to infer feature importance of the input-output pairing [136, 137]. Perturbations of images, for example, are performed by removing or inserting pixel or patch values to generate saliency maps at different levels of occlusion. Pixel-wise perturbations tend to be more spatially discrete and represent saliency more accurately in terms of location, yet

their higher granularity leads to worse representation of the semantics of salient objects [138]. Patch-wise methods provide more smooth saliency as the boundaries in the maps correspond better to object boundaries. Morphological fragmental perturbation pyramid (MFPP) is an example of such a method which uses morphological fragmentation to divide the input images into multiscale fragments and produce a perturbation mask by random masking [139]. Finding the right set of perturbations to apply remains a challenge across all cases.

An example of the gradient-based approach are integrated gradients that can help visualise input importance by calculating the integral of the gradients of the output prediction with respect to the input image pixels without any modification to the original network. The computation is done by using the average gradient as the input varies along a linear path from a baseline to provided input. The method can be used on various models such as image, text, or structured data [140]. The intuition for the original proposal came from using analogues of model coefficients like in classic regression models to describe attributions to output from input as pairs. As [141] note: "gradients (of the output with respect to the input) is a natural analog of the model coefficients for a deep network, and therefore the product of the gradient and feature values is a reasonable starting point for an attribution method." More formally, we attempt to explain a function $F : \mathrm{R}^n \rightarrow [0, 1]$ i.e. the output of the neural network with $X \in \mathrm{R}^n$ as input, and $X' \in \mathrm{R}^n$ as baseline (like a black image for imaging problems).

Integrated gradients are defined as the path intergral of the gradients along the straightline path from the baseline $X'$ to the input $X$. The integrated gradient along the $i^{th}$ dimension (can be understood as $i^{th}$ feature) for an input $X$ and baseline $X'$ with a straightline slope of $\alpha$ is defined as follows:

$$\mathrm{IG}_i(X) ::= (X_i - X_i') \times \int_{\alpha=0}^{1} \frac{\partial F\left(X' + \alpha \times (X - X')\right)}{\partial X_i} d\alpha \tag{3}$$

Another sub-domain of gradient-based methods are saliency-based methods which use gradient values either raw or normalised to infer salient features like in saliency maps described earlier. [136] highlight the weakness of such methods under robustness checks despite their advantages like simple formulations and lack of reliance on model pliability. Saliency or heatmaps have also been criticised for insufficient correlation with the network which they are meant to interpret thereby undermining their reliability as interpreters of the model. To address this concern, [142] propose Integrated-Gradients Optimized Saliency (I-GOS), a method which optimises a heatmap with the objective of maximally decreasing classification scores on the masked image. This is done by computing descent directions based on integrated gradients thus avoiding local optima and speeding up convergence. Gradient-based methods are often, however, not invariant under simple transformations of the input, and are very sensitive to the choice of reference point.

Further well-known examples of gradient-based methods include Deep Learning Important FeaTures (DeepLIFT) and deconvolution approaches. Instead of perturbing neurons which can be computationally inefficient due to seperate forward propagation for each perturbation, DeepLIFT explains the prediction of a neural network on a specific instance by using backpropagation of all neurons' contributions in the network to each feature. DeepLIFT determines scores in a single backward pass from the difference between the activation of each neuron to its 'reference activation' instead of relying on the gradient values themselves which tend to be unstable. It separately takes into account positive and negative contributions and thereby can reveal dependencies which are missed by methods like deconvolutions and guided backpropagation [143]. Deconvolution approaches like the deconvolutional network (Deconvnet) rely on deconvolving the network, going from neuron activations in the given layer back to the input. The reconstruction highlights part of the input that is most strongly activating the neuron [144]. The limits of this approach is that some backpropagated signals can be zero'd through ReLU activations. In Guided Backpropagation, where this only happens if either the input to the ReLU during the forward pass or the importance signal during the backward pass is

negative, is similar to computing gradients with the difference being that any gradients that become negative during the backward pass are discarded at ReLUs. Due to the zero-ing out of negative gradients, both approaches tend to not capture inputs that negatively contribute to the output prediction [143].

Layer-wise Relevance Propagation (LRP) is another attribution method that propagates the prediction backward in the neural network and compared to its perturbation- and gradient-based competitors it does not involve multiple computationally expensive neural network evaluations. It instead utilises the graph structure of the deep neural network to generate explanations [145]. The relevance score that is propagated is estimated for neurons $j$ and $k$, for example, based on those above them:

$$R_j = \sum_k \frac{z_{jk}}{\sum_j z_{jk}} R_k \tag{4}$$

The quantity $z_{jk}$ (neuron activation) models the extent to which neuron $j$ has contributed to make neuron $k$ relevant (denominator guaranteeing that the relevance is propagated downwards equally). The last layer relevance score is the pre-activation value corresponding to the class for which relevance scores are desired, and are obtained from the forward pass of an input. The propagated values to each layer are conserved and their sum is constant. LRP has been applied in many settings to offer interpretability to deep learning models which will be highlighted later under the applications section. LRP has shown to give better relevance or attributions than those by DeepLIFT without relying on the need for a reference input [146]. LRP suffers from its dependence on ReLU activations resulting in non-negative attribution maps thereby limiting the interpretability of the application, later addressed by an extension of the transformer using LRP-type of relevance score to compute relative relevance of heads in the attention mechanism using GELU instead. Conservation of the relevance score in the self-attention is then maintained through a normalisation step of the scores at the layers of the model. Class-specific interpretability can be obtained by extensions like Contrastive-LRP (CLRP) and Softmax-Gradient- LRP (SGLRP) where results of the class to be visualized are contrasted with the results of all other classes, to emphasize the differences and produce a class-dependent heatmap [116].

Other methods rely on manipulating the layers of the neural network itself similar to attention layers intuitively. Class Activation Mapping (CAM) effectively removes the last fully connected layers and replaces the MaxPooling layer by a global average pooling (GAP) layer after which the weighted average of features is extracted to form a per-class activation map (overlayed on the input like a heatmap) [147]. This makes it a non-gradient-based method but extensions of it like Gradient-weighted CAM (Grad-CAM) and Grad-CAM++ which propagate gradients through the network in obtaining the final heatmaps for input regional relevance. While the original CAM model depended on having a particular convolutional architecture of the model, the extensions can be applied to all CNN models and, in fact, use pixel-wise weighting of the gradients to optimise for multi-class scenarios in creating feature relevance maps [148]. Despite early proposals in image classification, these methods have been applied to semantic image segmentation and even time-series classification tasks. Other alterations of the method includes methods like Eigen-CAM which extracts the principle components from the convolutional layers and helps against classification errors made by fully connected layers in CNNs without needing to backpropagate gradients or use feature weighting. Early experiments have shown promising results albeit further evaluation on different datasets is needed alongside clearer analysis of the relative interpretability compared to other mask methods [149, 150].

Both perturbation- and gradient-based methods can be described as attribution-based as they establish a link of attribution between some area or form of the input to the output prediction. A disadvantage of the perturbation-based methods is the large number of possible combinations if one blindly attempted at going through all possible ways of perturbing the input without relying on approximations. Additionally, for different samples in the same dataset

belonging to the same class contradicting explanations can be generated resulting in decreased user trust. In a way, gradient-based approach address this by using the gradient as a proxy for these changes. relying on evaluating numerous feature subsets or solving an optimiza- tion problem for each instance of data. While gradient-based methods provide faster explanations, they also tend to be less accurate [106]. An advantage of perturbation-based methods, on the other hand, is their ability to query models repeatedly and their model-agnostic nature [138].

Another group of instance-based methods seeks to remedy the disadvantages of perturbation- and gradient-based methods by optimising for the fidelity of instance-level explanations. An example of such methods are amortized explanation methods (AEMs) which generate explanations by learning a global selector model that efficiently extracts locally important features in an instance of data with a single forward pass using an objective that measures the fidelity of the explanations. AEMs assess these selections with a predictor model for the output. L2X and INVASE fit the predictor and selector models jointly, often referred to as joint amortized explanation methods (JAMs) [151]. Since they are instance-based methods, JAMs have been previously deployed to explain mortality predictions to a patient-level [152]. Some problems with JAMs include their propensity to encode predictions with the learned selector and failure to select features involved in control flow (features involved only in branching decisions/nodes of tree structured generative process). An example of this could be mortality predictions for patients with chest pain using EHRs as [151] point out. Blood troponin levels can be a control flow feature where abnormal values indicate that cardiac imagining should be used to assess severity whereas normal values indicate that the chest pain may be non-cardiac. In this case, using JAMs to interpret the prediction will not represent the true role of troponin in patient mortality. To address these issues, [151] propose REAL-X and EVAL-X, a framework that combines efficient instance-level explanations with a single forward pass while also detecting when predictions are encoded in explanations without making out-of-distribution queries using approximations of the true data generating distribution. Other examples of instance-based methods include ProtoDash (and earlier iterations like influential instances and MMD-Critic) and counterfactuals described in detail in earlier sections. This section has included a myriad of different methods which have often developed as a response to weaknesses present in each other. There is no perfect explaianbility method as we still also lack a robust explainability metric to reliably compare them by. The section was meant to highlight diverging methodological pathways in approaching the same problem of using perturbations of the input or gradients of complex deep learning models in extracting some form of discernible explanations from machine learning problems. Attempts doing so in clinical risk prediction will be described in the next sections.

### 2.3.6 Concept-based

Compared to previously described methodologies, concept-based explainability relies on using internal conceptual learning of the models in extracting semantically-meaningful latent variables or concepts. Prior mentioned methods use coefficients, input values like pixel values and gradients or layer activation maps that do not correspond to human-level conceptual understanding. Work based on conceptual learning with concept activation vectors (CAVs) first and foremost seeks to address this gap by having deep learning models learn a latent space that would correspond more closely to human-level insight. Sometimes, these concepts are obtained through PCA dimensionality reduction (PCANet) or by using representative training patches to explain a prediction or even by using human-labeled concepts in estimation of concept scores [153, 154, 155, 156]. Conditional VAE models have also been proposed to model the causal effect between concepts and output predictions [157]. Going back to CAVs, these are vectors that are in the direction of the values of that concept's set of examples. The vector is normal to a hyperplane separating examples without a concept and examples with a concept in the activations. Testing with CAVs (T-CAV) uses directional derivatives to measure the 'conceptual sensitivity' to a high-level concept as learned by a CAV i.e. the sensitivity of the output to

changes in the input towards the direction of a concept at activation. The score (per concept per class) is the fraction of inputs whose l-layer activation vector was positively influenced by the concept (the magnitude of this influence is not taken into account). To check that the selected CAVs are meaningful, statistical significance testing with generating hundreds of CAVs and comparing their scores for consistency. The contributions of this approach are its global explainability not relying on a single sample for visualisation like with saliency maps and humans can be included in the conceptual selection of the models [158]. A major limitation of this approach is relying on, possibly biased, human-labeled examples of that concepts and the space of possible concepts is not necessarily limited a priori.

Automated Concept-based Explanation (ACE) is meant to address some of these limitations by automating the selection of meaningful concepts. A concept as such can be described as meaningful, coherent, and important. The first relates to its semantic and human meaningfulness, the second describes the necessity of different examples for the same concept to be perceptively similar, and the third touches on the concept's usefulness in the prediction task. The steps involved in applying ACE were detailed in the original paper and can be seen in Figure 6. Several resolutions are used to achieve more complex concepts from the image like parts and objects. The last step is assigning T-CAV scores to each concept based on its importance following the earlier described method of approximating the average positive effect of the concept on the class prediction [159]. The automation plays a crucial role in selecting the most relevant concepts based on several layers of segmentation coupled with a similarity metric from the activation layer distances computer in a large deep learning model. Extensions of the method to applications in the time-series domain are still, sadly, missing, but work on automated concept-based explanations has been done in text domains [160]. Some limitations include the reliance on clustering with image segmentation in identifying concepts from the representation space which might not always provide the most complex concepts. Work on more diverse data is still to be evaluated but a promising automated framework for conceptual explainability makes great inroads in alternative interpretability of opaque large deep learning models.

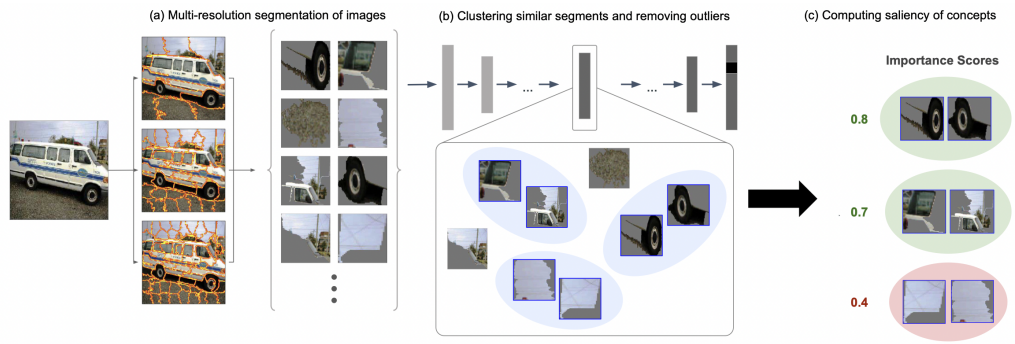


Figure 6: The ACE method consists of 4 steps: (a) A set of images from the same class is given. Each image is segmented with multiple resolutions resulting in a pool of segments all coming from the same class. (b) The activation space of one bottleneck layer of the model is used as a similarity space. After resizing each segment to the standard input size of the model, similar segments are clustered in the activation space and outliers are removed to increase coherency of clusters. (d) For each concept, its TCAV importance score is computed given its example segments [159]

An example of a concept-based interpretability method that does not require generative modelling, retraining, or reliance on clustering some latent space of a large deep learning model, is the Shapley value extension ConceptSHAP. In ConceptSHAP, subsets of concepts are

selected based on completeness as measured by a concept's score to reconstruct the output. The completeness score is then just the average accuracy of predicting the label using the concept scores done by some MLP. In other words, measuring the accuracy achieved by the concept score means measuring how "complete" the concepts are without relying on a score implicitly assuming a first-order relationship between the concepts and the outputs like ones does with T-CAV and ACE. These methods also make no assurances of the completeness of the concepts in explaining the model such as how many are needed for sufficient interpretability and their saliency scores may fail to capture concepts that have non-linear effects. Both of these limitations are addressed by ConceptSHAP where concepts are extracted in an unsupervised manner without these assumptions and where relevance is determined by Shapley values. However, a small MLP is involved in extracting concept embeddings, which makes the method not fully interpretable and self-contained [161, 162].

The role of causality has been mentioned earlier in implementing some interpretability methods for machine learning models. An entirely separate review can be written for causality and causal machine learning's role in explainability research but here we will present a method intersecting the former with concept-based explanations called Causal Concept Effect (CaCE). CaCE explores the causal effects of the presence or absence of a human-interpretable concept on the model prediction. This method as well as some of the previous ones are global i.e. providing average interpretability across all instances of the data provided and as such are more vulnerable to confounding of concepts. Investigating the presence or absence of concepts relies on using conditional generative models (VAEs for example) to generate counterfactuals and approximate the causal effect of concept explanations. For each sample in the test set, the difference in the model outputs is calculated for the original sample and for the counterfactual as generated. CaCE, by definition, is the average of these differences. T-CAV as we have mentioned remains vulnerable to collinearities between concepts and class labels whereas CaCE learns to measure the causal effects of the explanations. In some cases, CaCE can be computed directly if one has access to the generation process of the data (ground truth CaCE). One limitation of this method is that it is rather a metric for estimating the causal effects of pre-determined concepts rather than identfying the same for added explainability. As such, it would be best combined with other explainability methods as an added metric for concept-based expanation evaluations. Furthermore, relying on the data generating process to perturb concepts is unsustainable both because of the myriad of combinations, biases, and costs associated as well as the vulnerability to the weaknesses of the generative models used. Further work needs to be done in exploring the relationship between causal effects, causality, explainability, and conceptual learning for more human-friendly yet automated ways to provide explanations.

Concept-based explanations remain a source of insufficiently explored research especially compared to other areas of explainability. Moving away from image-based examples, concept-based methods have also been applied to text with INLP, CausaLM, and S-Learner [163, 164]. Each makes different assumptions on the model structure and their approach to estimating the causal effects of changes in the concept level. In CausaLM, for example, causal model explanations are generated using counterfactual language representation models based on fine-tuning of deep contextualized embedding models (like BERT) with auxiliary adversarial tasks derived from a causal graph [165]. Only recently has there been work on proposing a real-world dataset for benchmarking approaches and metrics to compare concept-based explanation methods which shows how limited current approaches are. Causal Estimation-Based Benchmark (CEBaB) is a benchmark dataset for text source concept-based explanations providing original and counterfactual input as well as real-world higher level concepts to study the causal effects on model behaviour. The causal effects of specific features in a causal graph are estimated and each explanation method is evaluated as a causal estimator of these measurements. CEBaB provides counterfactual examples that allow one to estimate these causal effects on an individual as well as conceptual level by directly comparing the actual change in model predictions with the change that a concept-based explanation method predicts [166]. Their approximate counterfactual

baseline outperforms all earlier mentioned methods at capturing both the direction and magnitude of causal effects showing just how far the field has to go to create robust concept-based explainability methods.

### 2.3.7 Data Modalities

**Imaging**
With the rise of deep learning models like convolutional neural networks (CNNs) in achieving success for image-based learning in the later half of 2010s, applications to medical imaging seemed like the next step. As early as 2017, the first Food and Drug Administration (FDA) approved application of AI in medicine saw success with *Arterys* for cardiac magnetic resonance images and since expanding to include liver and lung imaging, chest and musculoskeletal x-ray images, and noncontrast head CT images [167]. Within different sub-fields and specialisations in medicine, there exist different potential sources of image data including colonoscopies and the EUS platform in diagnosing malignant colon polyps or pancreatic cancer [168]. As such, a large component of AI in medical imaging applications includes (early) diagnosis rather than risk prediction whether that be detection, analysis, image denoising in the case of MRI, generating clinical text description, or even just segmentation and visualisations [169, 170, 171, 172]. Thus, a review of methods for clinical risk prediction using image data might be more sparse but the fact remains that a significant discussion of explainability in health care on image data cannot avoid mentioning these important applications as they contribute significantly to recent discussions of XAI in healthcare.

Although deep learning models have continued to achieve high levels of accuracy in diagnosing conditions from medical images, concerns remain regarding their sensitivities as well as lack of explainability for clinical relevance. Since this data modality is more closely tied to opaque and black-box deep learning models as compared to the other data modalities, the medical and clinical AI community has had to discuss the specific needs for a meticulous assessment of these models [173]. As image data is more easily analysed by humans in its original structure like in the case of text data, the patterns learned by AI might not be readily amenable to human identification as, for example, a conventional radiographic analysis would be [174]. This further highlights the importance in having explainabilty as an integral component of medical imaging AI applications.

The structure of medical images is represented by a numeric matrix containing values of either 0 or 1 in the case of greyscale, or between 0 and 255, in the case of a colour image input. The matrix can be respectively processed to switch between these modes of colour but depending on the model this might not present an important difference. There are no specific requirements for medical image analysis for deep learning as compared to other images as data augmentations, normalisations, and text association approaches remain similarly relevant for them as for other types of images. A key limitation remains also the need for extensive labelling, an expensive and slow process usually limited by human and time constraints, resulting in data augmentation techniques being even more important in increasing the sample size with the same amount of available labels [175, 176]. One must be aware that the aforementioned steps do affect potential explainability approaches as data that has been augmented and normalised might not be as easily visualised for prediction explanations and steps need to be taken in explainability methods when addressing the same. Medical image like other patient data needs to be properly de-identified and linked to the record, transferred, checked for quality control, and structured in a standardised manner in addition to eventual labelling. Figure 7 highlights these steps in the usual order of the data handling process. Another important consideration is that in most health care systems there is not widespread sharing of these patient data. Medical image data is often stored and analysed separately from other related patient data like electronic health records or genomics as well as data from different sites being in disparate silos, which is not optimal for research. When addressing generalisability and trustworthiness of algorithms developed

on these datasets, implementing limited explainability does not go far enough and being able to externally and clinically validate the models through multi-centre and interdisciplinary collaborations is a necessity [177]. Explainability in combination with multi-centre validation is, thus, crucial to achieving high-impact clinically meaningful AI algorithms.

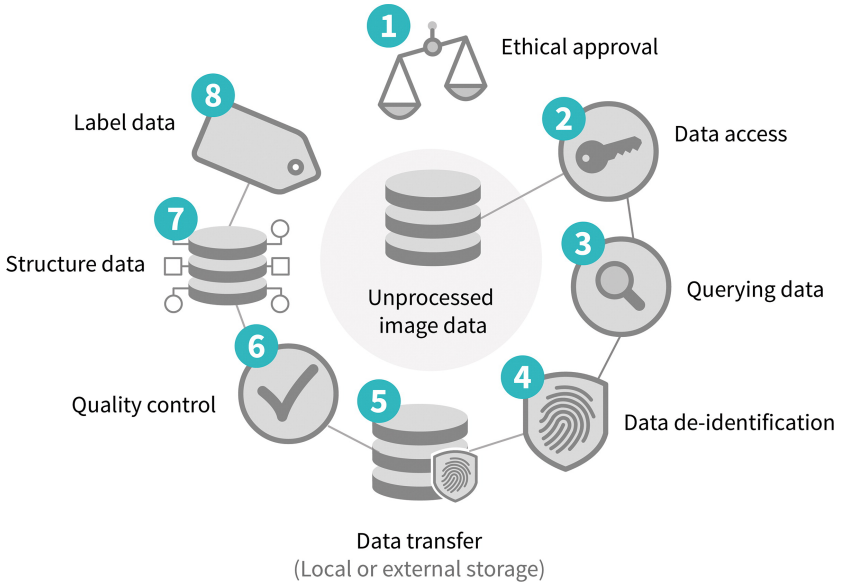


Figure 7: Diagram showing the data handling process for medical image data but which can be extended for other modalities as well [177]

In some cases, other data modalities can be represented as images to take advantage of the deep learning models specifically designed for this data type. In cases of electrocardiogram (ECG) or other signal analysis, occassionally during the data processing pipeline the signal will be converted to a spectral image which is then given to the deep learning model as input [178]. This, of course, means that the explainability analysis of such input remains limited with respect to the original input as the only image can be explained either in frequency or spectral domain and not the raw signal itself. But considering the noise and complexity of the original raw signal this can be seen as an improvement in creating human-understandable explanations for the predictions. In the following sections, we will detail approaches in AI for clinical risk prediction on medical images and later investigate the same with multi-modality in question, especially when combined with textual data from clinical notes.

**Text**

Following from the previous discussion on medical image data, some images like radiology slides often come with associated clinical notes and text. The application of AI to clinical text is multi-faceted as it can apply to analysing a physician's notes on diagnosis and tests and extracting relevant patient characteristics, using medical image annotations by specialists for labelling or phenotyping, and even using chatbots, survey and feedback forms as a source for estimating patient well-being [179, 180, 181, 182]. Besides radiology reports, other clinical annotations of image data are also possible like electroencephalography reports when studying epilepsy or echocardiography reports in cardiovascular medicine [183, 184]. When it comes to reading text from radiology reports, for example, methods from natural language processing (NLP) provide an avenue to extract relevant information and apply machine learning to relatively unstructured text data, data stored without a pre-determined or standardized format [181]. These unstructured data can contain patient details not captured in the regular electronic

record. Physicians often depend on this type of data to synthesise a previous clinician's notes and gain a better understanding of a patient's condition or treatment effects. For some health conditions like sepsis, early symptoms might be difficult to identify so having access to previous clinical notes describing the patient state is especially useful in clinical AI risk modelling [181, 185]. NLP is often used to extract diagnostic information either from ICD coding schemes or clinical notes directly through, for example, discharge summaries [186, 187]. One often needs to, however, analyse extensive amounts of clinical free text to extract relevant phrases (can be a challenge to identify) for diagnostic purposes [26]. The variety of such NLP applications, thus, comes with sometimes unique and sometimes shared challenges specific for this data modality for clinical AI applications.

And similarly again to medical image data, a large bottleneck in these applications are the annotations. Besides the costs and error vulnerabilities of manual annotations, in NLP, the task specifics often means that the training data annotations can rarely be reused. Some attempts in addressing this includes using active learning by involving human experts with relatively small annotation effort [188]. Other approaches include using pair-wise cosine similarity to compare sentences and prioritize those least similar to annotated sentences for annotation [26]. Multiple proposals use word embeddings and ontologies to create domain-specific mappings highlighting potential improved performance compared to more conventional techniques for concept normalisation [179]. Ideally, like in the previous modality case, unsupervised or semi-supervised learning paradigms can offer more robust answers to useful utilisation of clinical text for patient diagnosis.

Generalisability and external multi-centre validation is also a potential solution to clinical text analysis when there are concerns of data representativeness between different health centres and their clinical text data. Most studies are limited by using only their local hospital or medical centre data which has been shown to lead to biased, overfitted, and disparate model solutions in clinical NLP [26, 189]. The value of openly available and relatively large datasets like Medical Information Mart for Intensive Care (MIMIC) is, thus, high allowing them to be used for standardised testing of models albeit relying on only a few of the same datasets without updates for developing and testing models over the years might eventually lead to similar problems. Using multiple centres for external validation of models in clinical text which can differ greatly between health centres should be the optimal avenue to pursue [190]. Even in cases when data is available from different centres, standardising and preprocessing it consistently with large differences present still remains a challenge.

The models often developed for text applications are often commonly shared with time-series implementations as well which we will discuss further in the electronic health records (EHR) section. Suffice to say that compared to modelling based on static or image data, the sequential pattern of text data requires special attention and the underlying model paradigm is to develop dynamic learning frameworks capable of internal updates as the model learns to "read" the sequence of text. Traditional rule-based and count-based models also exist which rely on the numbers and distributions of specific phrases in the text but they are obviously limited in their learning capacity compared to more flexible deep learning approaches [179, 191]. Large parallelised and parametrised approaches built on the transformer backbone have received increased attention in text generation and prompt responses with the generative pre-trained transformer (GPT) models and their variations. Pre-training such large models and applying them in clinical setting includes cases like mental health support but the inability of such models to adjust for tone, context, and body language of the situation remains an important limitation in clinical integration [192]. Another major limitation is semantic repetitiveness, incoherence in long conversations, and internal contradictions [193]. As of now, using these relatively opaque models in healthcare with high-stakes interactions and emergencies comes with serious considerations that despite the overall popularity of the models in the general public leaves much to be desired in healthcare applications.

**Genomics**
Genomic data consists of a sequence of nucleotides with the respective labels being A, T, C, or G belonging to different bases present in human DNA. As such, the sequence of these letters in a description of the genome or its specific parts, often called genes, can be understood almost as text data. The expression of these sequences into proteins often leads down the line into the development of specific traits, susceptibilities or vulnerabilities to comorbidities and diseases. An important difference between human language text and genomic sequences being that the underlying sequential context and structure present in human language and sentences is not replicated in a genomic sequence. The ordering of the nucleotides is of importance, but there are no larger structures like phonems or words constructed for additional layered meaning. To this end, methods using sequential modules like recurrent units in NLP and time-series analysis have been applied in detecting single nucleotide mutations (called SNPs) for the purposes of diagnosing so-called Mendelian diseases from a person's genomic profile [194]. Due to the large amount of data that could be present (hundreds of millions of nucleotides up to billions for a whole genome sequence of a human DNA), deep learning has been a popular methodology to explore in detecting minor changes and patterns in a human-unreadable format. Studies have explored detecting variants causing complex eye diseases or early onset Alzheimer's disease often using either whole genome sequences or combining genomic data with other modalities like electronic health records and clinical notes and annotations [195, 196, 197]. In a lot of these implementations, the genomic profile sequence is represented in matrix form by one-hot encoding for the four possible nucleotide labels [198]. Genomic data represented in matrix format can then be also viewed as image data. DeepVariant, a deep learning model, detects single-nucleotide variants and Indels from sequences given a fixed reference sequence [194, 199]. DeepVariant relies on the dissimilarities in input images, in this case the input and reference sequence, to perform the (image) classification task for genetic variant calling. Sometimes, however, features are extracted from the genomic profile instead of learning based on the sequence directly. Examples can include describing genomic data by measuring the presence of regulatory elements in the sequence, k-mer counts, methylation, number of certain nucleotide pairs, and presence of histone modifications [194]. In either case, deep learning methods like recurrent and convolutional and graph units (and their combinations) have become the standard implementation for these problems thus carrying their opaque and unexplainable nature with them to another data modality realm.

Lack of explainability often comes as a cost of increased predictive performance in big data applications on genomic datasets. Attempts have been bade to introduce feature importance methods like permutation approaches or gradient backpropagation to deep learning models for genomics and thus produce attribution scores which can highlight the parts of a given input that are most influential for the model prediction. In DNA sequencing this amounts to highlighting which specific nucleotides and changes affect the model prediction the most as illustrated in Figure 8, and in most cases they correspond to instance-level explanations for each sequence input respectively. This is usually acceptable because in applications of variant identification or SNP detection, a single sample is usually the target of interest anyway instead of predictions for an entire dataset. A limitation of saliency maps as recognised by [200] is their weakness to neuron saturation. If there are two identical patterns in the data such as a particular motif being repeated, erasing one would not affect the model prediction. In the case of perturbation-based gradients or input-masked gradients, the importance scores would be low for both motifs, as they are individually not necessary for the prediction. DeepLIFT and integrated gradients solve this problem by comparing the input features with their 'reference' values such as a shuffled version of the original sequence. As in most cases, therefore, several explainability methods should be investigated when developing XAI applications in healthcare and their limitations and results respectively compared.

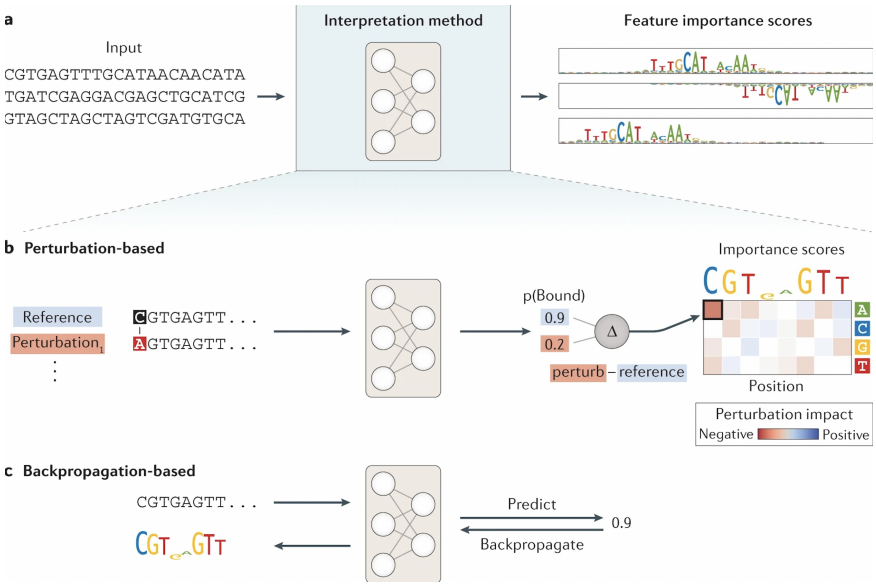


Figure 8: Different paradigms for providing explanations in AI for genomic data predictions. a) Feature importance scores proportional to letter heights state which bases of the sequence were more influential. Obtaining these scores can be done through: b) Perturbation-based approaches with single base changes and c) Backpropagation-based approaches using gradients with methods such as DeepLIFT [200]

**EHR**

In the field of clinical risk prediction with AI, barring imaging, electronic health records (EHR) data remain the most popular source of big data in healthcare. With rising rates of digitisation across the health services sectors globally, advanced machine learning models for disease prediction can finally be developed and evaluated on more or less longitudinal, large, complex, and real-world patient data. Decades ago, promises were made by leading international organisations in the revolution of patient care provided by clinical information systems part of which EHRs would play a critical role [201]. Unfortunately, due to financial and infrastructure challenges, EHR records have often stymied in their contributions towards the gap in access to better care with countries like the United States and the United Kingdom being far ahead of other countries and developing economies, including those facing the greatest health challenges [202]. This disparity can lead to a worsening of first-world-centred healthcare AI research which might not generalise to a majority of the world and those who need it the most. In the US, as of 2017, over 95% of hospitals were using EHR with a small difference between big hospitals and rural clinics. In the UK, more recently, EHRs played an important role in COVID-19 studies with millions of patients integrated into big data models and thus informing public guidance on health services [203]. For clinical risk prediction, EHRs remain an indispensible source of information to track both patient origins, progress, and departure from clinical care alongside a larger overview of disease progression and treatment than one would usually get from discrete slides or images.

EHR data usually consists of patient information like demographics, medical and surgical history, allergies and medications, diagnoses and procedures, vital sign measurements (heart rate, temperature, etc.), and laboratory blood test results. Some of these might be missing depending on the specific hospital wards the data originates from. For example, if working on intensive care unit (ICU) data, more regularly sampled clinical features would be present like vital sign and blood test measurements, but if operating on primary care data like longitudinal physician or clinician sources (like general practitioners in the UK), then one effectively works with static

demographic and anamnesis variables with the occasional physiological measurement. The data can also be relatively structured and unstructured, sometimes requiring text-processing and natural language processing methods to extract codified diagnoses and treatment information from clinical notes [204]. Besides inherent complexities one should not make the mistake of assuming EHR data is solely focused on generating information for research. In fact, at least in the case of the US, most EHR data is actually collected for hospital billing purposes which presents later challenges in the processing pipeline like inaccuracies, missingness, and bias [205]. These limitations are important but they are not lethal to research and as such should be recognised in studies and appropriately addressed. In clinical risk prediction, this usually amounts to detailing specific attributes of the data like sampling rate in case of signal measurements, amounts of missingness and reproducible and sensible approaches to addressing the same (like with simple or advanced imputation methods), and apparent sources of bias such as patient characteristics, income distributions, and socio-racial diversity.

Furthermore, EHR data, as defined here, is often of a mixed nature consisting of two main modalities: static or tabular and time-series data as seen in Figure 9. Static or tabular data usually consists of numeric or categorical variables such as patients' age, sex, ethnicity, comorbidities, and they can usually be one-hot encoded in the case of multi-level variables. Time-series data usually refers to numeric measurements captured over time instead of just having an instantaneous single measurement for a feature. For example, a patient might have several measurements for heart rate over time, hours in the case of ICU or months and years in the case of primary care. If the sampling frequency for this clinical signal is small enough, these features can be then effectively treated as static variables. Some time-series features are often raw waveforms like electrocardiogram (ECG) signals from wearable sensors usually used to monitor a patient's cardiovascular response and the heart's electrical activity. ECGs are often used for monitoring and diagnosis of relevant conditions [206]. In some cases, instead of using the raw physiological signal, features are extracted and used in algorithmic modelling. That is true in cases like ECG with parameters describing the QRS complex but also in cases like estimating clinical features of pulse pressure, mean arterial pressure (MAP), and heart rate variability [207]. When working with general time-series data that is neither ECG nor photoplethysmography (PPG) signal, different pre-processing techniques exist. Either features are extracted like the extremes, mean, variance or standard deviation and then treated as static variables used as model input or special recurrent models are used to process the time-series for machine learning. In the case of the latter, examples include using long-short-term-memory (LSTM) units or recurrent neural networks or gated recurrent units (GRUs) to capture the long from characteristics of a time-varying feature [208]. Other methods used to be more popular in time-series processing such as using Gaussian processes for standardising nonuniform data in both the time- and frequency-domain or even integrating the Gaussian process as part of an interpretable deep learning framework [209]. In sum, EHR records present a myriad of unique characteristics which makes the models proposed for their analysis similarly diverse. Explainability research has produced interesting advances balancing both static feature importance methods (post-hoc, global or local) as well as dynamic feature importance for time-series input. More detailed discussion on these approaches is included in the following section detailing the studies and their findings.

# 3 Search Methodology

We used IEEExplore and PubMed to define the starting database search using the key terms identified in Table 1. For all of the databases, we considered only publications in the last 5 years which corresponds to the overwhelming majority of research on the topic due to its relatively recent development as can be seen in Figures 10 and 11. Only papers published in English were reviewed. The final reference list was generated on the basis of originality and relevance to the scope of this Review by screening the titles and abstracts of the publications. For PubMed,

| Tabular | Categorical | Continuous |
|---|---|---|
| Patient ID | Sex | Heart Rate |
| 0001 | M | 99 |
| 0002 | F | 101 |
| … | … | … |

| Time-series | | Categorical | Continuous |
|---|---|---|---|
| Patient ID | Time (hr) | Mechanical Ventilation | Heart Rate |
| 0001 | 1 | 0 | 101 |
| 0001 | 3 | 1 | 99 |
| … | … | … | … |

Figure 9: Two most common data formats for EHR include tabular (top) and time-series (bottom) data. Usually, one will have several measurements over time for one patient implying multi-indexing with both the patient identifier and the time-stamp for each feature for each sample in the data handling stages

we included publications classified as classical article, comparative study, evaluation study, mutlicenter study, observational study, technical report, and validation study. For IEEExplore, we considered only journal publications. There were 95 publications identified by PubMed and 76 by IEEExplore. Google Scholar was also perused for additional publications not included in the PubMed and IEEExplore search.

The publications were split into different groups depending on the data modalities concerned. A summary can be seen in Figure 12. One can see that text and EHR data occupy the largest portion of applications probably due to their relatively prevalent use in clinical risk prediction applications whereas imaging is more prevalent in diagnosis or classifications cases. The exclusion criteria included:

- overlapping journal articles
- using the word interpretability in a clinical setting, for example when interpreting tumours from images without any relevance to machine learning interpretability per say
- works focused on classification and not clinical risk prediction
- journals not in the Q1 category of impact ranking

After applying the exclusion criteria we were left with a total of 89 publications to include in the review.

# 4 Applications of XAI on Healthcare Modalities

## 4.1 Imaging

In this section we will describe the broad strokes of XAI progress in medical imaging applications for risk prediction. Table 2 contains a detailed listing of the identified publications. We see

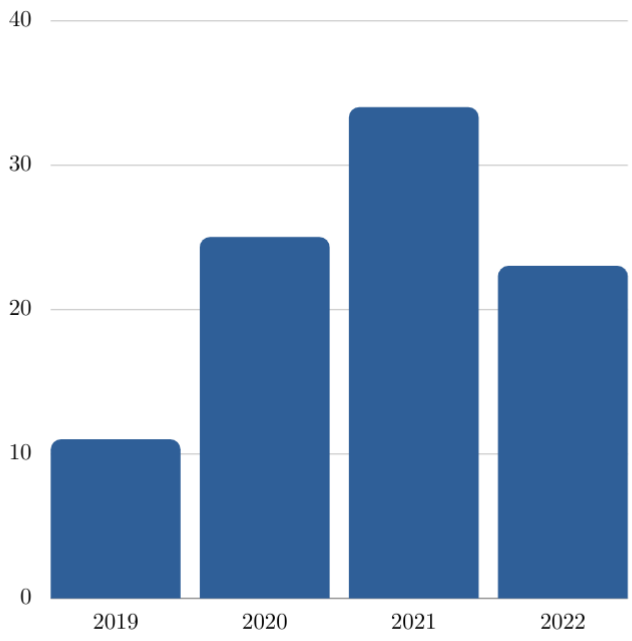


Figure 10: Frequency distribution of publications on XAI for clinical risk prediction since 2019 on PubMed database

that the vast majority of models are based on deep learning, mostly CNN models with some autoencoder structures as well. In terms of XAI methods, there is quite a variety of applications ranging from attention, SHAP/LIME, and CAM and its extensions. Interestingly, compared to the other modalities, medical imaging applications are the most likely to be clinically validated with a team of clinicians on board for added analysis of the explanations. In terms of quantitative evaluation, a few have resorted to statistical tests to account for sampling bias but no significant evaluations were found to add legitimacy to the explanations or the applications of the methods. Reproducibility and open-access standards are lacking as a majority of the references do not include a clear link to working code-hosting services. As medical imaging is usually collected as slides by specific procedures, the dataset sizes are a lot smaller and consist of up to thousands of samples maximum and can be in the low hundreds. Some methods like Grad-CAM being able to produce relatively robust explanations across both of these extremes showcases the variety of explainability methods available right now but, sadly, insufficiently evaluated otherwise.

## 4.2 Text

As we already mentioned, progress on XAI in clinical text analysis has been limited but key applications must be mentioned. Some of these are included in Table 4 with a detailed description of the key characteristics for each publication article. Text modality is the most common modality present in XAI for clinical risk prediction, mostly due to the presence of medical coding applications which we take as phenotyping prediction problems. A vast array of models have been applied to these problems ranging from deep learning models, through random forests, to regression models. The vast majority of XAI techniques are either inherent due to the usage

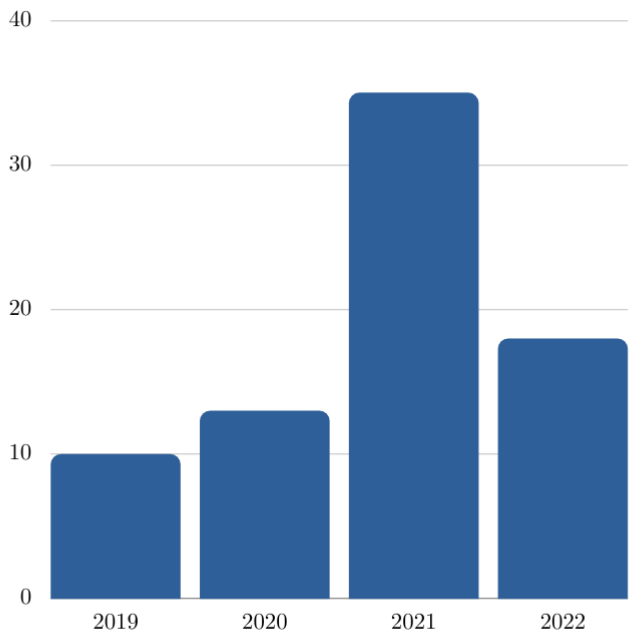


Figure 11: Frequency distribution of publications on XAI for clinical risk prediction since 2019 on IEEExplore database

of IIM or attention which was first proposed for textual data anyway. Dataset sizes also vary in orders of magnitude, and applications include anything from ICU outcome prediction to mental health risk prediction. Most applications did not include clinical validation or quantitative evaluation for the explainability methods and few had reproducible and easily accessible code repositories for their work. We see that the pattern present in the other modalities is also present here with textual data albeit at a larger scale due to there being a lot more applications to analyse.

## 4.3 Genomics

Genomics is by far the least represented of the modalities in XAI for clinical risk prediction as most of the work concerns pattern classifications or dimensionality analysis and not disease prediction per say. Table 5 shows the few papers that have done so for mortality, phenotyping, and COVID-19 severity prediction. Vast majority use some neural network architectures or IIM approaches like regressions. One of the positives is that most of these applications are indeed openly accessible with clear guidelines on how to access the code from the main paper. Some quantitative evaluation has been done including statistical tests and in cases of regression models, using external post-hoc methods to verify what has been claimed by extracting the regression coefficients. The most used post-hoc interpretability method is LIME and the dataset sizes vary considerably whether they be gene pairs or protein encodings. As far as clinical validation is concerned, most applications have not consulted clinicians or medical literature on the findings of the explainability results and that remains a potential avenue of further contribution for future applications.

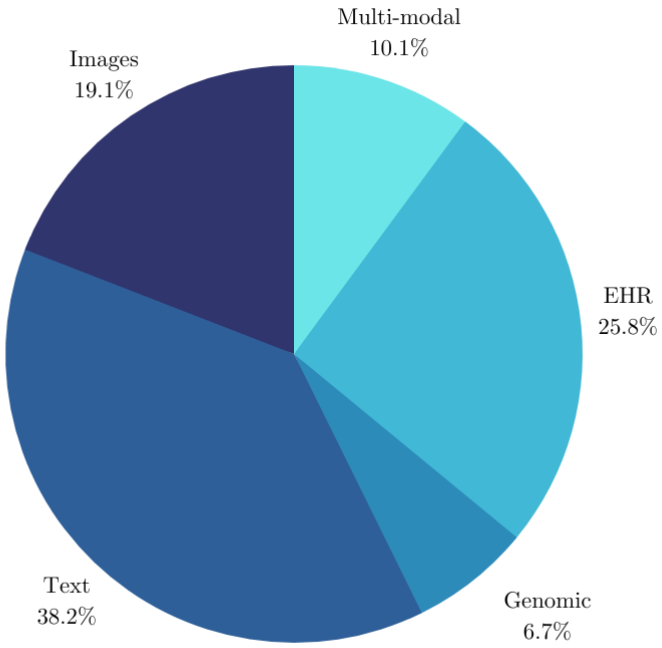


Figure 12: Pie chart of relative proportions of different modalities present in the explainability for clinical risk prediction applications literature

## 4.4 EHR

As far as EHR applications of explainable AI are concerned, most applications rely on tabular or static data with a smaller amount risk prediction problems dealing with time-series or sequential data. Overall, we can see that XGBoost and Shapley values are the most popular model and interpretability method used with a diverse range of dataset sizes ranging from a few hundred to hundreds of thousands of samples. In cases of time-series measurements, attention and deep learning adaptations of Shapley values are the most common implementations of interpretability in clinical risk prediction. To avoid inflating references, we did not include research papers that use the same model and interpretability method as ones already included to focus more on the overall trend and not including as many references as possible for its own sake. In terms of clinical tasks addressed, we see that mortality, COVID-19 outcomes (especially since 2020), and acute conditions like AKI and stroke are the most common application problems. Most of the papers did not clinically validate the interpretability results to guarantee consistency with medical literature to a significant extent, hence only including the interpretability plots or rankings as more of an afterthought rather than an end in and of itself. This stimulates the research culture in clinical risk prediction where interpretability is seen as a post-hoc step for checkmark purposes and not a significant aspect of the experimental design contributing to the research aims of the investigation at hand. A positive note is the use of some inherent models like regressions or causal models to address the need for explainability in cases of acute condition prediction following our earlier comments on some early push in this direction. The performance costs are then sometimes offset by carefully tuning models like XGBoost in tabular data examples which can achieve superior performance compared to most other, including advanced deep learning, models. Lastly, the vast majority of the papers have not made their code easily accessible from the main page either of the journal publication or the .pdf of the paper which undermines

Table 1: Search criteria for databases

| | | |
|---|---|---|
| "explainability" | "interpretability" | "XAI" |
| **AND** | | |
| "machine learning" | "artificial intelligence" | "AI" |
| **AND** | | |
| "clinical" | "medicine" | "healthcare" |
| **AND** | | |
| "prediction" | | |

the open access culture of the field. As compared to genomics and multi-modality sections of papers, this percentage of easily accessible code implementations is a lot smaller in the EHR case.

As can be seen from Table 6, most research applications in clinical risk prediction do not present any form of quantitative evaluation for their applications of interpretability methods when claiming explainability. Those that have been marked as having done so usually have only consulted their on-board clinical experts whereas only a few have more systematically used clinical feedback in evaluating the importance results from their applied interpretability methods. That remains the standard in the field of clinical AI for risk prediction using EHR records especially as it concerns any attempts at quantitative evaluation which none of the included papers have attempted to do with their interpretability results. This patterns stems as a consequence partially of the lack of research impetus on such requirements in these applications as well as a lack of existing reliable methods to quantitatively evaluate interpretability methods themselves. Very recent work by [266] highlights that in high-risk fields that are highly regulated like healthcare, there need to be robust quantitative evaluation frameworks for explainability of AI. Usual approaches such as using humans as evaluators, occluded datasets, and retraining models are costly, unreliable, and in some cases suffer under distribution shift. The authors propose corrupting the samples at increasing increments with the top-k and bottom-k important features as identified by the interpretability method. The change in the predictive score for each of these combinations is then plotted as two curves, one for the top-k features being corrupted, and another for the bottom-k. In a reliable interpretability method, the top-k curve measuring change in predicted score should rise dramatically as the most important features are corrupted and hence the prediction score should change significantly more than with the bottom-k curve. An AUROC and modified F1 score can be used to measure optimal behaviour between these curves and then used to compare interpretability methods. Future work will hopefully extend these results and check whether indeed, as the authors find on their few real-world and synthetic datasets, Shapley approaches remain superior methods for interpretability, at least in EHR time-series applications.

## 4.5 Multi-modalities

Multi-modal data is usually a combination of either text and image data or in some cases EHR and image, or EHR and text data. Few have used genomics in combination with the other modalities in explainable AI applications to clinical risk and this area presents a major potential field of growth. The models that get respectively applied are often related to the modalities, LSTMs in cases of time-series and text inclusion, CNNs if images are involved, and even XGBoost when text can be successfully summarised into a tabular format. The XAI methods consist of a plethora of mostly attention mechanisms and CAM when it comes to image data. There are examples of clinical validation where physicians were directly involved in the projects but most do not include sufficient analysis of the medical implications of the

Table 2: XAI for clinical risk prediction for medical images for different methods, as well as evaluation criteria they might satisfy. C.V. stands for clinical validation, ie. whether the explainability was at all clinically evaluated, Q.E. for quantitative evaluation of the explainability process, and O.A. whether the application is open access

| Reference | Year | Problem | | Dataset | XAI | | | |
|---|---|---|---|---|---|---|---|---|
| | | Task | Model | Size | Method | C.V. | Q.E. | O.A. |
| [210] | 2020 | myeloma | CBAM | 152 | attention | X | X | X |
| [211] | 2020 | cardiomyopathy | VAE | 10,000 | concepts | ✓ | X | X |
| [212] | 2020 | colorectal cancer | CNN | 5,000 | fuzzy | ✓ | X | X |
| [213] | 2020 | Alzheimer's | CNN | 642 | LIME | ✓ | X | X |
| [214] | 2021 | breast cancer | SVM | >1000 | LRP | ✓ | X | ✓ |
| [215] | 2021 | Alzheimer's | RL | 1,349 | attention | ✓ | X | X |
| [216] | 2021 | lung cancer | capsule network | 1,018 | correlation | ✓ | X | X |
| [217] | 2021 | COVID-19 recovery | CNN | 2,530 | FSR | X | ✓ | ✓ |
| [218] | 2021 | COVID-19 mortality | XGBoost | 3,028 | SHAP | ✓ | X | X |
| [219] | 2021 | MVI | CNN | 309 | Grad-CAM | ✓ | X | X |
| [220] | 2021 | breast cancer | CNN | 10,815 | Grad-CAM | ✓ | X | X |
| [124] | 2021 | glaucoma | CNN | 6,430 | adversarialism | ✓ | ✓ | X |
| [221] | 2021 | brain age | CNN | 2,639 | regression | ✓ | X | ✓ |
| [222] | 2021 | osteoarthritis | CNN | 4,796 | masking/CAM | ✓ | ✓ | ✓ |
| [223] | 2021 | colorectal cancer | CNN | 343 | regression | ✓ | ✓ | X |
| [224] | 2021 | lung cancer | XGBoost | 211 | SHAP | X | X | X |
| [225] | 2022 | endometrial cancer | CNN | 2,751 | attention | ✓ | X | ✓ |

Table 3: XAI for clinical risk prediction for text data for different methods, as well as evaluation criteria they might satisfy. C.V. stands for clinical validation, ie. whether the explainability was at all clinically evaluated, Q.E. for quantitative evaluation of the explainability process, and O.A. whether the application is open access

| Reference | Year | Problem | | Dataset | XAI | | | |
|---|---|---|---|---|---|---|---|---|
| | | Task | Model | Size | Method | C.V. | Q.E. | O.A. |
| [226] | 2019 | deterioration | CNN | 600,000 | attention | X | X | X |
| [227] | 2019 | mortality | regression | 23,310 | correlation | X | ✓ | X |
| [228] | 2019 | rapid response | Cox | 776,849 notes | inherent | X | ✓ | X |
| [229] | 2019 | readmission | regression | 136,963 | inherent | X | X | X |
| [230] | 2019 | AKI | regression | 16,558 | inherent | X | X | X |
| [231] | 2019 | medical screening | regression | 27,665 | inherent | X | ✓ | X |
| [232] | 2019 | mechanical restraint | random forest | 5,050 | inherent | ✓ | X | X |
| [233] | 2020 | dementia | random forest | 13,747 | inherent | X | X | X |
| [234] | 2020 | heart failure | RNN | 4,682 | attention | X | X | X |
| [235] | 2020 | LOS | regression | 313 | inherent | X | X | X |

Table 4: XAI for clinical risk prediction for text data for different methods, as well as evaluation criteria they might satisfy (continued). C.V. stands for clinical validation, ie. whether the explainability was at all clinically evaluated, Q.E. for quantitative evaluation of the explainability process, and O.A. whether the application is open access

| Reference | Year | Problem | | Dataset | XAI | | | |
|---|---|---|---|---|---|---|---|---|
| | | Task | Model | Size | Method | C.V. | Q.E. | O.A. |
| [236] | 2020 | SSI | CNN | 21,611 | attention | X | X | X |
| [237] | 2020 | LOS | TF-IDF | 12,962 | inherent | X | ✓ | X |
| [238] | 2020 | ICU admission | regression* | 120,649 | inherent | X | X | X |
| [239] | 2020 | mortality | XGBoost | 235,826 | inherent | X | X | X |
| [240] | 2020 | dementia | LightGBM | 207,416 | inherent | X | X | X |
| [241] | 2020 | ICU admission | decision tree | 10,504 | inherent | ✓ | X | X |
| [242] | 2020 | deterioration | Cox | 61,740 | inherent | X | X | X |
| [243] | 2020 | multiple | Bayesian latent topic | 80,000 | inherent | ✓ | ✓ | ✓ |
| [244] | 2020 | ED admission | CatBoost | 499,853 | inherent | X | X | X |
| [245] | 2020 | preterm birth | CatBoost | 3,611 | inherent/SHAP | X | X | ✓ |
| [246] | 2020 | ED admission | random forest | 89,459 | inherent | X | X | X |
| [247] | 2020 | cervical cancer | random forest | 1,321 | inherent | X | ✓ | X |
| [248] | 2020 | adherence | regression* | 791 | inherent | ✓ | X | X |
| [249] | 2021 | coding | Bi-GRU | 36,998 | attention | X | ✓ | X |
| [250] | 2021 | coding | CNN | 36,998 | attention | X | X | X |
| [251] | 2021 | readmission | XGBoost | 291 | inherent | X | X | X |
| [252] | 2021 | suicide | regression | 1,232 | inherent | ✓ | X | X |
| [253] | 2021 | ICU admission | XGBoost | 412,858 | MI | X | X | X |
| [254] | 2021 | mortality | LSTM | 50,000 | attention | X | X | X |
| [255] | 2022 | coding | CNN | 52,729 | attention | ✓ | ✓ | X |
| [256] | 2022 | coding | transformer | 1,300 | attention | X | X | X |
| [257] | 2022 | mental health | Bi-LSTM | 15,044 | attention | X | X | X |
| [258] | 2022 | depression | LSTM | 277,552 | LIME | X | X | X |
| [259] | 2022 | heart disease | EBM | 5,390 | inherent | ✓ | X | X |

Table 5: XAI for clinical risk prediction for genomic data for different methods, as well as evaluation criteria they might satisfy (continued). C.V. stands for clinical validation, ie. whether the explainability was at all clinically evaluated, Q.E. for quantitative evaluation of the explainability process, and O.A. whether the application is open access

| Reference | Year | Problem | | Dataset | XAI | | | |
|---|---|---|---|---|---|---|---|---|
| | | Task | Model | Size | Method | C.V. | Q.E. | O.A. |
| [260] | 2020 | AMR | regression | 1,595 | inherent | ✓ | ✓ | ✓ |
| [261] | 2020 | mortality | neural network | 7,803 | LIME | X | ✓ | X |
| [262] | 2021 | COVID-19 severity | regression | 12,965 | inherent | ✓ | ✓ | X |
| [263] | 2021 | phenotyping | neural network | 11,214 | inherent | X | X | ✓ |
| [264] | 2021 | macular degeneration | neural network | 32,215 | LIME | X | X | ✓ |
| [265] | 2021 | mortality | neural network | 3,431 | inherent | X | X | ✓ |

explainability results and quantitative evaluations of the XAI applications are even more limited. There are cases where open access approaches have been followed in terms of code transparency but those remain, sadly, a minority. Interestingly, attention seems to be the most popular way explainability is being incorporated into multi-modal applications for clinical risk prediction but it is still early days and this could be a limitation of existing deep learning methods applied in multi-modality applications being a more natural fit for attention (like LSTMs) for example.

# 5 Challenges and Future Outlook

It is important to note that many of the included methods are, in fact, practically limited. They are sometimes vulnerable to failures, redundant explanations, and wrong indications which makes interpretability on its own an insufficient attribute for achieving explainable and reliable AI clinical risk models [55]. Some work like [274, 283] and others has in recent years addressed the need for external validation if not as a companion to increased explainability then in lieu of it. These strategies for XAI in healthcare applications should be taken with a grain of salt as models externally validated on similar patient cohorts might not add additional trust or evaluate fairness more robustly than intuitive explainability methods. Ideally, a combination of both external validation with diverse stratified sub-population cohorts based on socio-economic and comorbidity groups and robust interpretability methods whether they be inherent or post-hoc is recommended to achieve explainability. An important extension should be to develop more robust explainability frameworks that can work across modalities and methodological contexts, as well as provide evidence for external multi-centre validation of proposed prediction models.

From the overview of explainable AI applications for clinical risk prediction across multiple modalities, it is clear that only rarely have the benefits of the models or explanations been evaluated for the clinician or patient reception. Some studies have included tests conducted with clinicians and patients for the generated explanations but sadly claims of explainable AI are greatly exaggerated in most cases. Furthermore, the added performance benefits of using opaque deep learning models at the cost of explainability are not convincing and the community seems to acknowledge this in modalities such as text and EHR where the majority of models are IIM or decision-based classifiers like XGBoost. The approach taken, thus, is to not have to compromise significantly between the trade-off but rather make inherently interpretable models that can also be high-performing. It is difficult to beat XGBoost and its cousins in tabular machine learning, even with deep learning models like TabNet and NODE. This approach should be extended to other modalities as well.

Table 6: XAI for clinical risk prediction for EHR data for different methods, as well as evaluation criteria they might satisfy. C.V. stands for clinical validation, ie. whether the explainability was at all clinically evaluated, Q.E. for quantitative evaluation of the explainability process, and O.A. whether the application is open access

| Reference | Year | Problem | | Dataset | XAI | | | |
|---|---|---|---|---|---|---|---|---|
| | | Task | Model | Size | Method | C.V. | Q.E. | O.A. |
| [267] | 2019 | hypertension | random forest | 23,095 | SHAP/LIME | X | ✓ | X |
| [268] | 2019 | obesity | XGBoost | 860,510 | SHAP | X | X | X |
| [269] | 2020 | COVID-19 | XGBoost | 485 | SHAP | X | X | X |
| [270] | 2020 | heart anomaly | CNN | 220,188 | SHAP | X | X | X |
| [271] | 2020 | AKI | XGBoost | 153,821 | inherent | X | X | ✓ |
| [272] | 2020 | mortality | random forest | - | SHAP | X | X | X |
| [273] | 2020 | stroke | Dempster-Shafer | 27,876 | inherent | ✓ | ✓ | X |
| [274] | 2020 | multiple | AdaBoost | - | WHIPS | X | ✓ | ✓ |
| [91] | 2020 | acute illness | TCN | 3,764 | DTD | X | X | ✓ |
| [275] | 2021 | dementia | XGBoost | 9,103 | SHAP | X | X | X |
| [260] | 2021 | Alzheimer's | LGP | 172 | inherent | X | ✓ | ✓ |
| [276] | 2021 | stroke | LightGBM | 3,213 | SHAP | ✓ | X | X |
| [277] | 2021 | CKD | random forest | 400 | SHAP | ✓ | ✓ | X |
| [278] | 2021 | post-op complications | XGBoost | 2,858 | SHAP | X | X | X |
| [279] | 2021 | AKI | XGBoost | 894 | SHAP | ✓ | X | X |
| [280] | 2021 | ECG changes | TabNet | 150 | attention | ✓ | X | X |
| [281] | 2021 | COVID-19 | random forest | 5,644 | SHAP/LIME | X | ✓ | X |
| [282] | 2021 | mortality | XGBoost | 36,658 | SHAP | X | X | X |
| [283] | 2021 | ED admission | XGBoost | 82,402 | SHAP | ✓ | ✓ | X |
| [284] | 2021 | QOL | XGBoost | 186 | SHAP | ✓ | X | X |
| [285] | 2021 | ACS | XGBoost | 278,608 | SHAP/LIME | X | ✓ | X |
| [286] | 2021 | PPG abnormality | CNN | 3,764 | attention | X | ✓ | X |
| [287] | 2022 | COVID-19 | XGBoost | 1,500 | SHAP | X | X | X |

Table 7: XAI for clinical risk prediction for multi-modal data for different methods, as well as evaluation criteria they might satisfy (continued). C.V. stands for clinical validation, ie. whether the explainability was at all clinically evaluated, Q.E. for quantitative evaluation of the explainability process, and O.A. whether the application is open access

| Reference | Year | Problem | | Dataset | XAI | | | |
|---|---|---|---|---|---|---|---|---|
| | | Task | Model | Size | Method | C.V. | Q.E. | O.A. |
| [288] | 2019 | mortality | CNN | 2220 | inherent | ✓ | X | X |
| [243] | 2020 | multiple | Bayesian latent topic | 80,000 | inherent | ✓ | ✓ | ✓ |
| [197] | 2021 | Alzheimer's | CNN | 2,220 | occlusion | X | X | ✓ |
| [289] | 2021 | pneumonia | Bayesian network | 35,389 | inherent | X | ✓ | X |
| [215] | 2021 | Alzheimer's | RL | 1,349 | attention | X | X | ✓ |
| [220] | 2021 | breast cancer | CNN | 10,815 | CAM | ✓ | X | X |
| [251] | 2021 | post-op complications | XGBoost | 291 | inherent | X | X | X |
| [254] | 2021 | mortality | LSTM | 50,000 | attention | X | X | X |
| [290] | 2022 | multiple cancers | AMIL/SNN | 6,592 | attention | ✓ | X | ✓ |

Some possible ideas for adoption of more rigorous testing of applied explainability in clinical risk prediction modelling is to test the methods on synthetic datasets with known underlying generative factors. There are already some examples of such datasets extracted from existing and in-use clinical datasets like the Clinical Practice Research Datalink (CPRD) containing EHRs of millions of patients in the United Kingdom or UK Biobank which similarly contains a large amount of relatively complex and diverse patient data [291]. Testing that the implemented XAI methods correctly identify most of the known factors is an important reliability test before being proposed to be used in clinical decision-making systems. It would add further trust when negotiating with clinicians regarding uptake. In our review of EHR time-series applications, we highlighted recent developments in implementing quantitative evaluations using a combination of corruption and ranking solutions based on adapted AUROC and F1 scores. More work still needs to be done on suggesting further frameworks for robust evaluation and comparison of interpretability or explainability more broadly in EHR and other clinical risk prediction domains.

While explainable AI holds great potential for revolutionizing clinical risk prediction, it is essential to recognize that it is not a panacea. Human expertise and domain knowledge remain indispensable in healthcare decision-making. Explainable AI should be seen as a tool to assist and augment healthcare professionals' judgment, providing them with transparent insights into the underlying factors contributing to predictions. If researchers seek to use explainability to guarantee the trust and reliability of their models for clinicians, patients, and other stakeholders, their implementations not being transparent or easily accessible sends the wrong message and stifles research growth. Since the papers that do have open access links on their publications are from high-ranking journals that demand open access and code sharing when submitting, a possible way to mitigate this is to have more journals demand accessible code sharing resources from authors, especially in cases of explainability research which, as we have seen, cannot be separated from concepts around trustworthiness, the key to which is transparency and reproducibility in the case of software.

We should not, however, be overly critical or pessimistic about the outcomes of explainability research in clinical risk prediction. The medical field is more than familiar with using black-box technology as many drugs' mechanisms have still not been elucidated for their health benefits.

Paracetamol is a commonly cited example of a popular over-the-counter drug whose mechanisms of action have not been revealed for a much longer time than we have had deep learning [55, 292]. Several methods for achieving interpretability in clinical risk prediction were explored, including rule-based models, feature importance techniques, and post-hoc methods but a key point has been made of interpretability on its own being an insufficient attribute of a truly explainable clinical risk prediction AI model. Each approach, thus, presented its own advantages and limitations, emphasizing the need for a careful selection and integration of multiple techniques to ensure a comprehensive understanding of AI-driven predictions. The progress forward has to be focused on an end-to-end approach to explainability in clinical risk prediction, no longer being enough to simply apply an explainability method to a model, but rather to clinically and quantitatively measure its success while also including different stakeholders from clinicians, and patients, to developers into the process. Clear regulations, guidelines, and research culture practices will help make this transition smoother and for the benefit of a larger group of stakeholders.

# References


[1] F.-Y. Wang, J. J. Zhang, X. Zheng, X. Wang, Y. Yuan, X. Dai, J. Zhang, and L. Yang, "Where does alphago go: From church-turing thesis to alphago thesis and beyond," *IEEE/CAA Journal of Automatica Sinica*, vol. 3, no. 2, pp. 113–120, 2016.

[2] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel *et al.*, "Mastering atari, go, chess and shogi by planning with a learned model," *Nature*, vol. 588, no. 7839, pp. 604–609, 2020.

[3] R. Dale, "Gpt-3: What's it good for?" *Natural Language Engineering*, vol. 27, no. 1, pp. 113–118, 2021.

[4] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[5] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. S. Corrado, A. Darzi *et al.*, "International evaluation of an ai system for breast cancer screening," *Nature*, vol. 577, no. 7788, pp. 89–94, 2020.

[6] C. Leibig, M. Brehmer, S. Bunk, D. Byng, K. Pinker, and L. Umutlu, "Combining the strengths of radiologists and ai for breast cancer screening: a retrospective analysis," *The Lancet Digital Health*, vol. 4, no. 7, pp. e507–e519, 2022.

[7] J. Deng, Z. Yang, I. Ojima, D. Samaras, and F. Wang, "Artificial intelligence in drug discovery: applications and techniques," *Briefings in Bioinformatics*, vol. 23, no. 1, 2022.

[8] G. Chassagnon, M. Vakalopoulou, E. Battistella, S. Christodoulidis, T.-N. Hoang-Thi, S. Dangeard, E. Deutsch, F. Andre, E. Guillo, N. Halm *et al.*, "Ai-driven quantification, staging and outcome prediction of covid-19 pneumonia," *Medical image analysis*, vol. 67, p. 101860, 2021.

[9] M. Van Smeden, G. Heinze, B. Van Calster, F. W. Asselbergs, P. E. Vardas, N. Bruining, P. De Jaegere, J. H. Moore, S. Denaxas, A. L. Boulesteix *et al.*, "Critical appraisal of artificial intelligence-based prediction models for cardiovascular disease," *European Heart Journal*, vol. 43, no. 31, pp. 2921–2930, 2022.

[10] D. Branley-Bell, R. Whitworth, and L. Coventry, "User trust and understanding of explainable ai: exploring algorithm visualisations and user biases," in *International Conference on Human-Computer Interaction*. Springer, 2020, pp. 382–399.

[11] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020.

[12] A. Malhi, S. Knapic, and K. Främling, "Explainable agents for less bias in human-agent decision making," in *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*. Springer, 2020, pp. 129–146.

[13] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[14] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, vol. 10, no. 3152676, pp. 10–5555, 2017.

[15] A. Selbst and J. Powles, ""meaningful information" and the right to explanation," in *Conference on Fairness, Accountability and Transparency*. PMLR, 2018, pp. 48–48.

[16] S. Wachter and B. Mittelstadt, "A right to reasonable inferences: re-thinking data protection law in the age of big data and ai," *Colum. Bus. L. Rev.*, p. 494, 2019.

[17] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, "Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)," *Computer Methods and Programs in Biomedicine*, p. 107161, 2022.

[18] M. Nazar, M. M. Alam, E. Yafi, and M. M. Su'ud, "A systematic review of human–computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques," *IEEE Access*, vol. 9, pp. 153 316–153 348, 2021.

[19] K. Rasheed, A. Qayyum, M. Ghaly, A. Al-Fuqaha, A. Razi, and J. Qadir, "Explainable, trustworthy, and ethical machine learning for healthcare: A survey," *Computers in Biology and Medicine*, p. 106043, 2022.

[20] S. Dey, P. Chakraborty, B. C. Kwon, A. Dhurandhar, M. Ghalwash, F. J. S. Saiz, K. Ng, D. Sow, K. R. Varshney, and P. Meyer, "Human-centered explainability for life sciences, healthcare, and medical informatics," *Patterns*, vol. 3, no. 5, p. 100493, 2022.

[21] F. Giuste, W. Shi, Y. Zhu, T. Naren, M. Isgut, Y. Sha, L. Tong, M. Gupte, and M. D. Wang, "Explainable artificial intelligence methods in combating pandemics: A systematic review," *IEEE Reviews in Biomedical Engineering*, 2022.

[22] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital signal processing*, vol. 73, pp. 1–15, 2018.

[23] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[24] J. Zhong and E. Negre, "Ai: To interpret or to explain?" in *Congrès Inforsid ((INFormatique des ORganisations et Systèmes d'Information et de Décision) 2021*, 2021.

[25] M. van Smeden, J. B. Reitsma, R. D. Riley, G. S. Collins, and K. G. Moons, "Clinical prediction models: diagnosis versus prognosis," *Journal of clinical epidemiology*, vol. 132, pp. 142–145, 2021.

[26] I. Spasic, G. Nenadic *et al.*, "Clinical text data in machine learning: systematic review," *JMIR medical informatics*, vol. 8, no. 3, p. e17984, 2020.

[27] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis," *IEEE journal of biomedical and health informatics*, vol. 22, no. 5, pp. 1589–1604, 2017.

[28] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 2018, pp. 80–89.

[29] R. ElShawi, Y. Sherif, M. Al-Mallah, and S. Sakr, "Interpretability in healthcare: A comparative study of local machine learning interpretability techniques," *Computational Intelligence*, vol. 37, no. 4, pp. 1633–1650, 2021.

[30] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2020.

[31] A. K. Manrai, B. H. Funke, H. L. Rehm, M. S. Olesen, B. A. Maron, P. Szolovits, D. M. Margulies, J. Loscalzo, and I. S. Kohane, "Genetic misdiagnoses and the potential for health disparities," *New England Journal of Medicine*, vol. 375, no. 7, pp. 655–665, 2016.

[32] N. C. Arpey, A. H. Gaglioti, and M. E. Rosenbaum, "How socioeconomic status affects patient perceptions of health care: a qualitative study," *Journal of primary care & community health*, vol. 8, no. 3, pp. 169–175, 2017.

[33] S. Barocas and A. D. Selbst, "Big data's disparate impact," *Calif. L. Rev.*, vol. 104, p. 671, 2016.

[34] H.-W. Liu, C.-F. Lin, and Y.-J. Chen, "Beyond state v loomis: artificial intelligence, government algorithmization and accountability," *International journal of law and information technology*, vol. 27, no. 2, pp. 122–141, 2019.

[35] M. DeCamp and C. Lindvall, "Latent bias and the implementation of artificial intelligence in medicine," *Journal of the American Medical Informatics Association*, vol. 27, no. 12, pp. 2020–2023, 2020.

[36] M. K. Cho, "Rising to the challenge of bias in health care ai," *Nature Medicine*, vol. 27, no. 12, pp. 2079–2081, 2021.

[37] S. Alelyani, "Detection and evaluation of machine learning bias," *Applied Sciences*, vol. 11, no. 14, p. 6271, 2021.

[38] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[39] M. A. Ahmad, A. Patel, C. Eckert, V. Kumar, and A. Teredesai, "Fairness in machine learning for healthcare," in *Proceedings of the 26th acm sigkdd international conference on knowledge discovery & data mining*, 2020, pp. 3529–3530.

[40] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," *Advances in neural information processing systems*, vol. 30, 2017.

[41] Y. Ahn and Y.-R. Lin, "Fairsight: Visual analytics for fairness in decision making," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 1, pp. 1086–1095, 2019.

[42] N. Bantilan, "Themis-ml: A fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation," *Journal of Technology in Human Services*, vol. 36, no. 1, pp. 15–30, 2018.

[43] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (xai): A survey," *arXiv preprint arXiv:2006.11371*, 2020.

[44] A. Preece, D. Harborne, D. Braines, R. Tomsett, and S. Chakraborty, "Stakeholders in explainable ai," *arXiv preprint arXiv:1810.00184*, 2018.

[45] C. A. of Radiologists (CAR) Artificial Intelligence Working Group, "Canadian association of radiologists white paper on ethical and legal issues related to artificial intelligence in radiology," *Canadian Association of Radiologists' Journal*, vol. 70, no. 2, pp. 107–118, 2019.

[46] H. Chen, C. Gomez, C.-M. Huang, and M. Unberath, "Explainable medical imaging ai needs human-centered design: guidelines and evidence from a systematic review," *npj Digital Medicine*, vol. 5, no. 1, pp. 1–15, 2022.

[47] D. S. Char, M. D. Abràmoff, and C. Feudtner, "Identifying ethical considerations for machine learning healthcare applications," *The American Journal of Bioethics*, vol. 20, no. 11, pp. 7–17, 2020.

[48] C. Feudtner, T. Schall, P. Nathanson, and J. Berry, "Ethical framework for risk stratification and mitigation programs for children with medical complexity," *Pediatrics*, vol. 141, no. Supplement_3, pp. S250–S258, 2018.

[49] X. Wang and M. Yin, "Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making," in *26th International Conference on Intelligent User Interfaces*, 2021, pp. 318–328.

[50] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. Weld, "Does the whole exceed its parts? the effect of ai explanations on complementary team performance," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–16.

[51] A. Wölker and T. E. Powell, "Algorithms in the newsroom? news readers' perceived credibility and selection of automated journalism," *Journalism*, vol. 22, no. 1, pp. 86–103, 2021.

[52] D. Shin, B. Zhong, and F. A. Biocca, "Beyond user experience: What constitutes algorithmic experiences?" *International Journal of Information Management*, vol. 52, p. 102061, 2020.

[53] M. T. Ribeiro, S. Singh, and C. Guestrin, """ why should i trust you?""" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[54] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[55] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *The Lancet Digital Health*, vol. 3, no. 11, pp. e745–e750, 2021.

[56] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[57] M. van Oijen, "Linear modelling: Lm, glm, gam and mixed models," in *Bayesian Compendium*. Springer, 2020, pp. 137–140.

[58] P. Ju, X. Lin, and J. Liu, "Overfitting can be harmless for basis pursuit, but only to a degree," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7956–7967, 2020.

[59] B. Ustun and C. Rudin, "Supersparse linear integer models for optimized medical scoring systems," *Machine Learning*, vol. 102, no. 3, pp. 349–391, 2016.

[60] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.

[61] N. McCauley and M. Ala, "The use of expert systems in the healthcare industry," *Information & Management*, vol. 22, no. 4, pp. 227–235, 1992.

[62] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan, "Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model," *The Annals of Applied Statistics*, vol. 9, no. 3, pp. 1350–1371, 2015.

[63] M. R. Zafar and N. Khan, "Deterministic local interpretable model-agnostic explanations for stable explainability," *Machine Learning and Knowledge Extraction*, vol. 3, no. 3, pp. 525–541, 2021.

[64] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[65] C. Z. Janikow, "Fuzzy decision trees: issues and methods," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 28, no. 1, pp. 1–14, 1998.

[66] C. Marsala, "Data mining with ensembles of fuzzy decision trees," in *2009 IEEE Symposium on Computational Intelligence and Data Mining*. IEEE, 2009, pp. 348–354.

[67] V. Levashenko, E. Zaitseva, and S. Puuronen, "Fuzzy classifier based on fuzzy decision tree," in *EUROCON 2007-The International Conference on" Computer as a Tool"*. IEEE, 2007, pp. 823–827.

[68] C.-F. Lin and S.-D. Wang, "Fuzzy support vector machines," *IEEE transactions on neural networks*, vol. 13, no. 2, pp. 464–471, 2002.

[69] Z. Bian, C. M. Vong, P. K. Wong, and S. Wang, "Fuzzy knn method with adaptive nearest neighbors," *IEEE Transactions on Cybernetics*, 2020.

[70] R. Chimatapu, H. Hagras, A. Starkey, and G. Owusu, "Explainable ai and fuzzy logic systems," in *International Conference on Theory and Practice of Natural Computing*. Springer, 2018, pp. 3–20.

[71] Y. Deng, Z. Ren, Y. Kong, F. Bao, and Q. Dai, "A hierarchical fused fuzzy deep neural network for data classification," *IEEE Transactions on Fuzzy Systems*, vol. 25, no. 4, pp. 1006–1012, 2016.

[72] S. Park, S. J. Lee, E. Weiss, and Y. Motai, "Intra-and inter-fractional variation prediction of lung tumors using fuzzy deep learning," *IEEE journal of translational engineering in health and medicine*, vol. 4, pp. 1–12, 2016.

[73] R. Chimatapu, H. Hagras, A. Starkey, and G. Owusu, "Interval type-2 fuzzy logic based stacked autoencoder deep neural network for generating explainable ai models in workforce optimization," in *2018 IEEE international conference on fuzzy systems (FUZZ-IEEE)*. IEEE, 2018, pp. 1–8.

[74] R. Chimatapu, H. Hagras, M. Kern, and G. Owusu, "Hybrid deep learning type-2 fuzzy logic systems for explainable ai," in *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2020, pp. 1–6.

[75] H. Lakkaraju, S. H. Bach, and J. Leskovec, "Interpretable decision sets: A joint framework for description and prediction," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1675–1684.

[76] S. Dash, O. Gunluk, and D. Wei, "Boolean decision rules via column generation," *Advances in neural information processing systems*, vol. 31, 2018.

[77] T. Wang and C. Rudin, "Learning optimized or's of and's," *arXiv preprint arXiv:1511.02210*, 2015.

[78] C. Lawless and O. Gunluk, "Fair decision rules for binary classification," *arXiv preprint arXiv:2107.01325*, 2021.

[79] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[80] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *Advances in neural information processing systems*, vol. 27, 2014.

[81] P. Micaelli and A. J. Storkey, "Zero-shot knowledge transfer via adversarial belief matching," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[82] A. Dhurandhar, K. Shanmugam, R. Luss, and P. A. Olsen, "Improving simple models with confidence profiles," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[83] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Wortman Vaughan, and H. Wallach, "Manipulating and measuring model interpretability," in *Proceedings of the 2021 CHI conference on human factors in computing systems*, 2021, pp. 1–52.

[84] D. Alvarez Melis and T. Jaakkola, "Towards robust interpretability with self-explaining neural networks," *Advances in neural information processing systems*, vol. 31, 2018.

[85] R. Churchill and J. Brown, *Ebook: Complex Variables and Applications*. McGraw Hill, 2014.

[86] I. Puri, A. Dhurandhar, T. Pedapati, K. Shanmugam, D. Wei, and K. R. Varshney, "Cofrnets: interpretable neural architecture inspired by continued fractions," *Advances in Neural Information Processing Systems*, vol. 34, pp. 21 668–21 680, 2021.

[87] A. Sarkar, D. Vijaykeerthy, A. Sarkar, and V. N. Balasubramanian, "A framework for learning ante-hoc explainable models via concepts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 286–10 295.

[88] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[89] L. Ibrahim, M. Mesinovic, K.-W. Yang, and M. A. Eid, "Explainable prediction of acute myocardial infarction using machine learning and shapley values," *IEEE Access*, vol. 8, pp. 210 410–210 417, 2020.

[90] M. Ancona, C. Oztireli, and M. Gross, "Explaining deep neural networks with a polynomial time algorithm for shapley value approximation," in *International Conference on Machine Learning*. PMLR, 2019, pp. 272–281.

[91] S. M. Lauritsen, M. Kristensen, M. V. Olsen, M. S. Larsen, K. M. Lauritsen, M. J. Jørgensen, J. Lange, and B. Thiesson, "Explainable artificial intelligence model to predict acute critical illness from electronic health records," *Nature communications*, vol. 11, no. 1, pp. 1–11, 2020.

[92] M. R. Zafar and N. M. Khan, "Dlime: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems," *arXiv preprint arXiv:1906.10263*, 2019.

[93] M. Staniak and P. Biecek, "Explanations of model predictions with live and breakdown packages," *arXiv preprint arXiv:1804.01955*, 2018.

[94] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das, "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," *Advances in neural information processing systems*, vol. 31, 2018.

[95] T. Miller, "Contrastive explanation: A structural-model approach," *The Knowledge Engineering Review*, vol. 36, 2021.

[96] R. Luss, P.-Y. Chen, A. Dhurandhar, P. Sattigeri, Y. Zhang, K. Shanmugam, and C.-C. Tu, "Leveraging latent features for local explanations," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 1139–1149.

[97] G. Shao, L. Tang, and H. Zhang, "Introducing image classification efficacies," *IEEE Access*, vol. 9, pp. 134 809–134 816, 2021.

[98] K. S. Gurumoorthy, A. Dhurandhar, G. Cecchi, and C. Aggarwal, "Efficient data representation by selecting prototypes with importance weights," in *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2019, pp. 260–269.

[99] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," *Advances in neural information processing systems*, vol. 29, 2016.

[100] A.-p. Nguyen and M. R. Martínez, "On quantitative aspects of model interpretability," *arXiv preprint arXiv:2007.07584*, 2020.

[101] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[102] Y. Jones, F. Deligianni, and J. Dalton, "Improving ecg classification interpretability using saliency maps," in *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, 2020, pp. 675–682.

[103] N. Arun, N. Gaw, P. Singh, K. Chang, M. Aggarwal, B. Chen, K. Hoebel, S. Gupta, J. Patel, M. Gidwani *et al.*, "Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging," *Radiology: Artificial Intelligence*, vol. 3, no. 6, 2021.

[104] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[105] A. A. Ismail, H. Corrada Bravo, and S. Feizi, "Improving deep learning interpretability by saliency guided training," *Advances in Neural Information Processing Systems*, vol. 34, pp. 26 726–26 739, 2021.

[106] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," *Advances in neural information processing systems*, vol. 31, 2018.

[107] J. Gu and V. Tresp, "Saliency methods for explaining adversarial attacks," *arXiv preprint arXiv:1908.08413*, 2019.

[108] I. Gandin, A. Scagnetto, S. Romani, and G. Barbati, "Interpretability of time-series deep learning models: A study in cardiovascular patients admitted to intensive care unit," *Journal of Biomedical Informatics*, vol. 121, p. 103876, 2021.

[109] S. Jain and B. C. Wallace, "Attention is not explanation," *arXiv preprint arXiv:1902.10186*, 2019.

[110] B. Bai, J. Liang, G. Zhang, H. Li, K. Bai, and F. Wang, "Why attentions may not be interpretable?" in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 25–34.

[111] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[112] Y. Xu, H. Ying, S. Qian, F. Zhuang, X. Zhang, D. Wang, J. Wu, and H. Xiong, "Time-aware context-gated graph attention network for clinical risk prediction," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[113] S. A. Kamal, C. Yin, B. Qian, and P. Zhang, "An interpretable risk prediction model for healthcare with pattern attention," *BMC Medical Informatics and Decision Making*, vol. 20, no. 11, pp. 1–10, 2020.

[114] Y. Zhang and J. Li, "Application of heartbeat-attention mechanism for detection of myocardial infarction using 12-lead ecg records," *Applied Sciences*, vol. 9, no. 16, p. 3328, 2019.

[115] U. Girkar, R. Uchimido, L.-w. H. Lehman, P. Szolovits, L. Celi, and W.-H. Weng, "Predicting blood pressure response to fluid bolus therapy using neural networks with clinical interpretability," *Circulation Research*, vol. 125, no. Suppl_1, pp. A448–A448, 2019.

[116] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 782–791.

[117] D. Barić, P. Fumić, D. Horvatić, and T. Lipic, "Benchmarking attention-based interpretability of deep learning in multivariate time series predictions," *Entropy*, vol. 23, no. 2, p. 143, 2021.

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

[119] K. S. Gurumoorthy, A. Dhurandhar, and G. Cecchi, "Protodash: Fast interpretable prototype selection," *arXiv preprint arXiv:1707.01212*, 2017.

[120] G. Plumb, D. Molitor, and A. S. Talwalkar, "Model agnostic supervised local explanations," *Advances in neural information processing systems*, vol. 31, 2018.

[121] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[122] T. Bäck, M. Preuss, A. Deutz, H. Wang, C. Doerr, M. Emmerich, and H. Trautmann, *Parallel Problem Solving from Nature–PPSN XVI: 16th International Conference, PPSN 2020, Leiden, The Netherlands, September 5-9, 2020, Proceedings, Part II*. Springer Nature, 2020, vol. 12270.

[123] A. Ignatiev, N. Narodytska, and J. Marques-Silva, "On relating explanations and adversarial examples," *Advances in neural information processing systems*, vol. 32, 2019.

[124] J. Chang, J. Lee, A. Ha, Y. S. Han, E. Bak, S. Choi, J. M. Yun, U. Kang, I. H. Shin, J. Y. Shin *et al.*, "Explaining the rationale of deep learning glaucoma decisions with adversarial examples," *Ophthalmology*, vol. 128, no. 1, pp. 78–88, 2021.

[125] A. Brandsæter and I. K. Glad, "Shapley values for cluster importance," *Data Mining and Knowledge Discovery*, pp. 1–32, 2022.

[126] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, "Distribution-free predictive inference for regression," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1094–1111, 2018.

[127] A. Carrillo, L. F. Cantú, and A. Noriega, "Individual explanations in machine learning models: A survey for practitioners," *arXiv preprint arXiv:2104.04144*, 2021.

[128] C. Molnar, G. König, B. Bischl, and G. Casalicchio, "Model-agnostic feature importance and effects with dependent features: a conditional subgroup approach," *Data Mining and Knowledge Discovery*, pp. 1–39, 2023.

[129] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[130] N. Agarwal and S. Das, "Interpretable machine learning tools: A survey," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2020, pp. 1528–1534.

[131] A. Yeh and A. Ngo, "Bringing a ruler into the black box: Uncovering feature impact from individual conditional expectation plots," in *Machine Learning and Principles and Practice of Knowledge Discovery in Databases: International Workshops of ECML PKDD 2021, Virtual Event, September 13-17, 2021, Proceedings, Part I*. Springer, 2022, pp. 34–48.

[132] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," *journal of Computational and Graphical Statistics*, vol. 24, no. 1, pp. 44–65, 2015.

[133] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," *arXiv preprint arXiv:1606.05386*, 2016.

[134] V. Arya, R. K. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović *et al.*, "One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques," *arXiv preprint arXiv:1909.03012*, 2019.

[135] C. Meng, L. Trinh, N. Xu, and Y. Liu, "Mimic-if: Interpretability and fairness evaluation of deep learning models on mimic-iv dataset," *arXiv preprint arXiv:2102.06761*, 2021.

[136] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," *arXiv preprint arXiv:1806.08049*, 2018.

[137] ——, "A causal framework for explaining the predictions of black-box sequence-to-sequence models," *arXiv preprint arXiv:1707.01943*, 2017.

[138] M. Ivanovs, R. Kadikis, and K. Ozols, "Perturbation-based methods for explaining deep neural networks: A survey," *Pattern Recognition Letters*, vol. 150, pp. 228–234, 2021.

[139] Q. Yang, X. Zhu, J.-K. Fwu, Y. Ye, G. You, and Y. Zhu, "Mfpp: Morphological fragmental perturbation pyramid for black-box model explanations," in *2020 25th International conference on pattern recognition (ICPR)*. IEEE, 2021, pp. 1376–1383.

[140] I. Čík, A. D. Rasamoelina, M. Mach, and P. Sinčák, "Explaining deep neural network using layer-wise relevance propagation and integrated gradients," in *2021 IEEE 19th world symposium on applied machine intelligence and informatics (SAMI)*. IEEE, 2021, pp. 000 381–000 386.

[141] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[142] Z. Qi, S. Khorram, and F. Li, "Visualizing deep networks by optimizing with integrated gradients." in *CVPR Workshops*, vol. 2, 2019, pp. 1–4.

[143] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *International conference on machine learning*. PMLR, 2017, pp. 3145–3153.

[144] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[145] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: an overview," *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 193–209, 2019.

[146] J. Grezmak, J. Zhang, P. Wang, K. A. Loparo, and R. X. Gao, "Interpretable convolutional neural network through layer-wise relevance propagation for machine fault diagnosis," *IEEE Sensors Journal*, vol. 20, no. 6, pp. 3172–3181, 2019.

[147] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[148] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 839–847.

[149] K. Vinogradova, A. Dibrov, and G. Myers, "Towards interpretable semantic segmentation via gradient-weighted class activation mapping (student abstract)," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 10, 2020, pp. 13 943–13 944.

[150] M. B. Muhammad and M. Yeasin, "Eigen-cam: Class activation map using principal components," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–7.

[151] N. Jethani, M. Sudarshan, Y. Aphinyanaphongs, and R. Ranganath, "Have we learned to explain?: How interpretability methods can learn to encode predictions in their interpretations." in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 1459–1467.

[152] J. Yoon, J. Jordon, and M. van der Schaar, "Invase: Instance-wise variable selection using neural networks," in *International Conference on Learning Representations*, 2018.

[153] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "Pcanet: A simple deep learning baseline for image classification?" *IEEE transactions on image processing*, vol. 24, no. 12, pp. 5017–5032, 2015.

[154] J. Chorowski, R. J. Weiss, S. Bengio, and A. Van Den Oord, "Unsupervised speech representation learning using wavenet autoencoders," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 12, pp. 2041–2053, 2019.

[155] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This looks like that: deep learning for interpretable image recognition," *Advances in neural information processing systems*, vol. 32, 2019.

[156] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5338–5348.

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

[158] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *International conference on machine learning*. PMLR, 2018, pp. 2668–2677.

[159] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim, "Towards automatic concept-based explanations," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[160] T. Shi, X. Zhang, P. Wang, and C. K. Reddy, "Corpus-level and concept-based explanations for interpretable document classification," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 16, no. 3, pp. 1–17, 2021.

[161] V. Kamakshi, U. Gupta, and N. C. Krishnan, "Pace: Posthoc architecture-agnostic concept extractor for explaining cnns," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.

[162] C.-K. Yeh, B. Kim, S. Arik, C.-L. Li, T. Pfister, and P. Ravikumar, "On completeness-aware concept-based explanations in deep neural networks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 20 554–20 565, 2020.

[163] S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, and Y. Goldberg, "Null it out: Guarding protected attributes by iterative nullspace projection," *arXiv preprint arXiv:2004.07667*, 2020.

[164] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu, "Metalearners for estimating heterogeneous treatment effects using machine learning," *Proceedings of the national academy of sciences*, vol. 116, no. 10, pp. 4156–4165, 2019.

[165] A. Feder, N. Oved, U. Shalit, and R. Reichart, "Causalm: Causal model explanation through counterfactual language models," *Computational Linguistics*, vol. 47, no. 2, pp. 333–386, 2021.

[166] E. D. Abraham, K. D'Oosterlinck, A. Feder, Y. O. Gat, A. Geiger, C. Potts, R. Reichart, and Z. Wu, "Cebab: Estimating the causal effects of real-world concepts on nlp model behavior," *arXiv preprint arXiv:2205.14140*, 2022.

[167] V. Kaul, S. Enslin, and S. A. Gross, "History of artificial intelligence in medicine," *Gastrointestinal endoscopy*, vol. 92, no. 4, pp. 807–812, 2020.

[168] I. M. Cazacu, A. Udristoiu, L. G. Gruionu, A. Iacob, G. Gruionu, and A. Saftoiu, "Artificial intelligence in pancreatic cancer: Toward precision diagnosis," *Endoscopic ultrasound*, vol. 8, no. 6, p. 357, 2019.

[169] F. Renard, S. Guedria, N. D. Palma, and N. Vuillerme, "Variability and reproducibility in deep learning for medical image segmentation," *Scientific Reports*, vol. 10, no. 1, pp. 1–16, 2020.

[170] K. Li, W. Zhou, H. Li, and M. A. Anastasio, "Assessing the impact of deep neural network-based image denoising on binary signal detection tasks," *IEEE transactions on medical imaging*, vol. 40, no. 9, pp. 2295–2305, 2021.

[171] M. M. A. Monshi, J. Poon, and V. Chung, "Deep learning in generating radiology reports: A survey," *Artificial Intelligence in Medicine*, vol. 106, p. 101878, 2020.

[172] L. Alzubaidi, M. Al-Amidie, A. Al-Asadi, A. J. Humaidi, O. Al-Shamma, M. A. Fadhel, J. Zhang, J. Santamaría, and Y. Duan, "Novel transfer learning approach for medical imaging with limited labeled data," *Cancers*, vol. 13, no. 7, p. 1590, 2021.

[173] X. Liu, L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern *et al.*, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis," *The lancet digital health*, vol. 1, no. 6, pp. e271–e297, 2019.

[174] H. Lee, E.-J. Lee, S. Ham, H.-B. Lee, J. S. Lee, S. U. Kwon, J. S. Kim, N. Kim, and D.-W. Kang, "Machine learning approach to identify stroke within 4.5 hours," *Stroke*, vol. 51, no. 3, pp. 860–866, 2020.

[175] P. Chlap, H. Min, N. Vandenberg, J. Dowling, L. Holloway, and A. Haworth, "A review of medical image data augmentation techniques for deep learning applications," *Journal of Medical Imaging and Radiation Oncology*, vol. 65, no. 5, pp. 545–563, 2021.

[176] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on mri," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, 2019.

[177] M. J. Willemink, W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R. M. Summers, D. L. Rubin, and M. P. Lungren, "Preparing medical imaging data for machine learning," *Radiology*, vol. 295, no. 1, pp. 4–15, 2020.

[178] F. Murat, O. Yildirim, M. Talo, U. B. Baloglu, Y. Demir, and U. R. Acharya, "Application of deep learning techniques for heartbeats detection using ecg signals-analysis and review," *Computers in biology and medicine*, vol. 120, p. 103726, 2020.

[179] A. Casey, E. Davidson, M. Poon, H. Dong, D. Duma, A. Grivas, C. Grover, V. Suárez-Paniagua, R. Tobin, W. Whiteley *et al.*, "A systematic review of natural language processing applied to radiology reports," *BMC medical informatics and decision making*, vol. 21, no. 1, p. 179, 2021.

[180] A. Tanwar, J. Zhang, J. Ive, V. Gupta, and Y. Guo, "Unsupervised numerical reasoning to extract phenotypes from clinical text by leveraging external knowledge," in *Multimodal AI in healthcare: A paradigm shift in health intelligence*. Springer, 2022, pp. 11–28.

[181] K. H. Goh, L. Wang, A. Y. K. Yeow, H. Poh, K. Li, J. J. L. Yeow, and G. Y. H. Tan, "Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare," *Nature communications*, vol. 12, no. 1, p. 711, 2021.

[182] J. Lin, T. Joseph, J. J. Parga-Belinkie, A. Mandel, R. Schumacher, K. Neumann, L. Scalise, J. Gaulton, L. Christ, K. Leitner *et al.*, "Development of a practical training method for a healthcare artificial intelligence (ai) chatbot," *BMJ Innovations*, vol. 7, no. 2, 2021.

[183] M. A. Casteleiro, J. Des Diz, N. Maroto, M. J. F. Prieto, S. Peters, C. Wroe, C. S. Torrado, D. M. Fernandez, R. Stevens *et al.*, "Semantic deep learning: Prior knowledge and a type of four-term embedding analogy to acquire treatments for well-known diseases," *JMIR medical informatics*, vol. 8, no. 8, p. e16948, 2020.

[184] A. Vaid, K. W. Johnson, M. A. Badgeley, S. S. Somani, M. Bicak, I. Landi, A. Russak, S. Zhao, M. A. Levin, R. S. Freeman *et al.*, "Using deep-learning algorithms to simultaneously identify right and left ventricular dysfunction from the electrocardiogram," *Cardiovascular Imaging*, vol. 15, no. 3, pp. 395–410, 2022.

[185] M. Wu, X. Du, R. Gu, and J. Wei, "Artificial intelligence for clinical decision support in sepsis," *Frontiers in Medicine*, vol. 8, p. 665464, 2021.

[186] F. Li and H. Yu, "Icd coding from clinical text using multi-filter residual convolutional neural network," in *proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 05, 2020, pp. 8180–8187.

[187] A. Sammani, A. Bagheri, P. G. van der Heijden, A. S. Te Riele, A. F. Baas, C. Oosters, D. Oberski, and F. W. Asselbergs, "Automatic multilabel detection of icd10 codes in dutch cardiology discharge letters using neural networks," *NPJ digital medicine*, vol. 4, no. 1, p. 37, 2021.

[188] K. Lybarger, M. Ostendorf, and M. Yetisgen, "Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction," *Journal of Biomedical Informatics*, vol. 113, p. 103631, 2021.

[189] X. Zhan, M. Humbert-Droz, P. Mukherjee, and O. Gevaert, "Structuring clinical text with ai: Old versus new natural language processing techniques evaluated on eight common cardiovascular diseases," *Patterns*, vol. 2, no. 7, p. 100289, 2021.

[190] M. G. Kersloot, F. J. van Putten, A. Abu-Hanna, R. Cornet, and D. L. Arts, "Natural language processing algorithms for mapping clinical text fragments onto ontology concepts: a systematic review and recommendations for future studies," *Journal of biomedical semantics*, vol. 11, pp. 1–21, 2020.

[191] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning–based text classification: a comprehensive review," *ACM computing surveys (CSUR)*, vol. 54, no. 3, pp. 1–40, 2021.

[192] D. M. Korngiebel and S. D. Mooney, "Considering the possibilities and pitfalls of generative pre-trained transformer 3 (gpt-3) in healthcare delivery," *NPJ Digital Medicine*, vol. 4, no. 1, p. 93, 2021.

[193] K. Elkins and J. Chun, "Can gpt-3 pass a writer's turing test?" *Journal of Cultural Analytics*, vol. 5, no. 2, 2020.

[194] W. S. Alharbi and M. Rashid, "A review of deep learning applications in human genomics using next-generation sequencing data," *Human Genomics*, vol. 16, no. 1, pp. 1–20, 2022.

[195] S. K. Wang, S. Nair, R. Li, K. Kraft, A. Pampari, A. Patel, J. B. Kang, C. Luong, A. Kundaje, and H. Y. Chang, "Single-cell multiome of the human retina and deep learning nominate causal variants in complex eye diseases," *Cell Genomics*, vol. 2, no. 8, p. 100164, 2022.

[196] T. Jo, K. Nho, P. Bice, A. J. Saykin, A. D. N. Initiative *et al.*, "Deep learning-based identification of genetic variants: application to alzheimer's disease classification," *Briefings in Bioinformatics*, vol. 23, no. 2, 2022.

[197] J. Venugopalan, L. Tong, H. R. Hassanzadeh, and M. D. Wang, "Multimodal deep learning models for early detection of alzheimer's disease stage," *Scientific reports*, vol. 11, no. 1, pp. 1–13, 2021.

[198] G. Eraslan, Ž. Avsec, J. Gagneur, and F. J. Theis, "Deep learning: new computational modelling techniques for genomics," *Nature Reviews Genetics*, vol. 20, no. 7, pp. 389–403, 2019.

[199] M. Kumaran, U. Subramanian, and B. Devarajan, "Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data," *BMC bioinformatics*, vol. 20, no. 1, pp. 1–11, 2019.

[200] G. Eraslan, Avsec, J. Gagneur, and F. Theis, "Deep learning: new computational modelling techniques for genomics," *Nature Reviews Genetics*, pp. 1–13, 2019.

[201] N. S. Abul-Husn and E. E. Kenny, "Personalized medicine and the power of electronic health records," *Cell*, vol. 177, no. 1, pp. 58–69, 2019.

[202] Y.-K. Lin, M. Lin, and H. Chen, "Do electronic health records affect quality of care? evidence from the hitech act," *Information Systems Research*, vol. 30, no. 1, pp. 306–318, 2019.

[203] O. Collaborative, E. Williamson, A. J. Walker, K. Bhaskaran, S. Bacon, C. Bates, C. E. Morton, H. J. Curtis, A. Mehrkar, D. Evans *et al.*, "Opensafely: factors associated with covid-19-related hospital death in the linked electronic health records of 17 million adult nhs patients," *MedRxiv*, pp. 2020–05, 2020.

[204] S. A. Pendergrass and D. C. Crawford, "Using electronic health records to generate phenotypes for research," *Current protocols in human genetics*, vol. 100, no. 1, p. e80, 2019.

[205] E. Mahmoudi, N. Kamdar, N. Kim, G. Gonzales, K. Singh, and A. K. Waljee, "Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review," *bmj*, vol. 369, 2020.

[206] L. Xie, Z. Li, Y. Zhou, Y. He, and J. Zhu, "Computational diagnostic techniques for electrocardiogram signal analysis," *Sensors*, vol. 20, no. 21, p. 6318, 2020.

[207] F.-T.-Z. Khanam, A. Al-Naji, and J. Chahl, "Remote monitoring of vital signs in diverse non-clinical and clinical scenarios using computer vision systems: A review," *Applied Sciences*, vol. 9, no. 20, p. 4474, 2019.

[208] Z. Ebrahimi, M. Loni, M. Daneshtalab, and A. Gharehbaghi, "A review on deep learning methods for ecg arrhythmia classification," *Expert Systems with Applications: X*, vol. 7, p. 100033, 2020.

[209] F. E. Shamout, T. Zhu, P. Sharma, P. J. Watkinson, and D. A. Clifton, "Deep interpretable early warning system for the detection of clinical deterioration," *IEEE journal of biomedical and health informatics*, vol. 24, no. 2, pp. 437–446, 2019.

[210] L. Morvan, C. Nanni, A.-V. Michaud, B. Jamet, C. Bailly, C. Bodet-Milin, S. Chauvie, C. Touzeau, P. Moreau, E. Zamagni *et al.*, "Learned deep radiomics for survival analysis with attention," in *International Workshop on PRedictive Intelligence In MEdicine*. Springer, 2020, pp. 35–45.

[211] E. Puyol-Antón, C. Chen, J. R. Clough, B. Ruijsink, B. S. Sidhu, J. Gould, B. Porter, M. Elliott, V. Mehta, D. Rueckert *et al.*, "Interpretable deep models for cardiac resynchronisation therapy response prediction," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 284–293.

[212] P. Sabol, P. Sinčák, P. Hartono, P. Kočan, Z. Benetinová, A. Blichárová, L. Verbóová, E. Štammová, A. Sabolová-Fabianová, and A. Jašková, "Explainable classifier for improving the accountability in decision-making for colorectal cancer diagnosis from histopathological images," *Journal of biomedical informatics*, vol. 109, p. 103523, 2020.

[213] P. R. Magesh, R. D. Myloth, and R. J. Tom, "An explainable machine learning model for early detection of parkinson's disease using lime on datscan imagery," *Computers in Biology and Medicine*, vol. 126, p. 104041, 2020.

[214] A. Binder, M. Bockmayr, M. Hägele, S. Wienert, D. Heim, K. Hellweg, M. Ishii, A. Stenzinger, A. Hocke, C. Denkert *et al.*, "Morphological and molecular breast cancer profiling through explainable machine learning," *Nature Machine Intelligence*, vol. 3, no. 4, pp. 355–366, 2021.

[215] Q. Zhang, Q. Du, and G. Liu, "A whole-process interpretable and multi-modal deep reinforcement learning for diagnosis and analysis of alzheimer's disease," *Journal of Neural Engineering*, vol. 18, no. 6, p. 066032, 2021.

[216] P. Afshar, F. Naderkhani, A. Oikonomou, M. J. Rafiee, A. Mohammadi, and K. N. Plataniotis, "Mixcaps: A capsule network-based mixture of experts for lung nodule malignancy prediction," *Pattern Recognition*, vol. 116, p. 107942, 2021.

[217] J. Wang, C. Liu, J. Li, C. Yuan, L. Zhang, C. Jin, J. Xu, Y. Wang, Y. Wen, H. Lu *et al.*, "icovid: interpretable deep learning framework for early recovery-time prediction of covid-19 patients," *NPJ digital medicine*, vol. 4, no. 1, p. 124, 2021.

[218] L. Jia, Z. Wei, H. Zhang, J. Wang, R. Jia, M. Zhou, X. Li, H. Zhang, X. Chen, Z. Yu *et al.*, "An interpretable machine learning model based on a quick pre-screening system enables accurate deterioration risk prediction for covid-19," *Scientific Reports*, vol. 11, no. 1, p. 23127, 2021.

[219] S.-C. Liu, J. Lai, J.-Y. Huang, C.-F. Cho, P. H. Lee, M.-H. Lu, C.-C. Yeh, J. Yu, and W.-C. Lin, "Predicting microvascular invasion in hepatocellular carcinoma: a deep learning model validated across hospitals," *Cancer Imaging*, vol. 21, pp. 1–16, 2021.

[220] X. Qian, J. Pei, H. Zheng, X. Xie, L. Yan, H. Zhang, C. Han, X. Gao, H. Zhang, W. Zheng *et al.*, "Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning," *Nature biomedical engineering*, vol. 5, no. 6, pp. 522–532, 2021.

[221] P. L. Ballester, L. T. Da Silva, M. Marcon, N. B. Esper, B. N. Frey, A. Buchweitz, and F. Meneguzzi, "Predicting brain age at slice level: convolutional neural networks and consequences for interpretability," *Frontiers in Psychiatry*, vol. 12, p. 598518, 2021.

[222] E. Pierson, D. M. Cutler, J. Leskovec, S. Mullainathan, and Z. Obermeyer, "An algorithmic approach to reducing unexplained pain disparities in underserved populations," *Nature Medicine*, vol. 27, no. 1, pp. 136–140, 2021.

[223] R. Yamashita, J. Long, T. Longacre, L. Peng, G. Berry, B. Martin, J. Higgins, D. L. Rubin, and J. Shen, "Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study," *The Lancet Oncology*, vol. 22, no. 1, pp. 132–141, 2021.

[224] N. Q. K. Le, Q. H. Kha, V. H. Nguyen, Y.-C. Chen, S.-J. Cheng, and C.-Y. Chen, "Machine learning-based radiomics signatures for egfr and kras mutations prediction in non-small-cell lung cancer," *International journal of molecular sciences*, vol. 22, no. 17, p. 9254, 2021.

[225] S. Fremond, S. Andani, J. B. Wolf, J. Dijkstra, S. Melsbach, J. J. Jobsen, M. Brinkhuis, S. Roothaan, I. Jurgenliemk-Schulz, L. C. Lutgens *et al.*, "Interpretable deep learning model to predict the molecular classification of endometrial cancer from haematoxylin and eosin-stained whole-slide images: a combined analysis of the portec randomised trials and clinical cohorts," *The Lancet Digital Health*, vol. 5, no. 2, pp. e71–e82, 2023.

[226] I. Girardi, P. Ji, A.-p. Nguyen, N. Hollenstein, A. Ivankay, L. Kuhn, C. Marchiori, and C. Zhang, "Patient risk assessment and warning symptom detection using deep attention-based neural networks," *arXiv preprint arXiv:1809.10804*, 2018.

[227] W. Kongburan, M. Chignell, N. Charoenkitkarn, and J. H. Chan, "Enhancing predictive power of cluster-boosted regression with text-based indexing," *IEEE Access*, vol. 7, pp. 43 394–43 405, 2019.

[228] Z. T. Korach, K. D. Cato, S. A. Collins, M. J. Kang, C. Knaplund, P. C. Dykes, L. Wang, K. O. Schnock, J. P. Garcia, H. Jia *et al.*, "Unsupervised machine learning of topics documented by nurses about hospitalized patients prior to a rapid-response event," *Applied Clinical Informatics*, vol. 10, no. 05, pp. 952–963, 2019.

[229] S. M. Mahajan and R. Ghani, "Combining structured and unstructured data for predicting risk of readmission for heart failure patients." in *MedInfo*, 2019, pp. 238–242.

[230] M. Sun, J. Baron, A. Dighe, P. Szolovits, R. G. Wunderink, T. Isakova, and Y. Luo, "Early prediction of acute kidney injury in critical care setting using clinical notes and structured multivariate physiological measurements." *MedInfo*, vol. 264, pp. 368–72, 2019.

[231] X. Zhang, M. F. Bellolio, P. Medrano-Gracia, K. Werys, S. Yang, and P. Mahajan, "Use of natural language processing to improve predictive models for imaging utilization in children presenting to the emergency department," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, pp. 1–13, 2019.

[232] A. Danielsen, M. Fenger, S. Østergaard, K. Nielbo, and O. Mors, "Predicting mechanical restraint of psychiatric inpatients by applying machine learning on electronic health data," *Acta Psychiatrica Scandinavica*, vol. 140, no. 2, pp. 147–157, 2019.

[233] Z. B. Miled, K. Haas, C. M. Black, R. K. Khandker, V. Chandrasekaran, R. Lipton, and M. A. Boustani, "Predicting dementia with routine care emr data," *Artificial Intelligence in Medicine*, vol. 102, p. 101771, 2020.

[234] J. Gong, X. Bai, D.-a. Li, J. Zhao, and X. Li, "Prognosis analysis of heart failure based on recurrent attention model," *IRBM*, vol. 41, no. 2, pp. 71–79, 2020.

[235] S. Bacchi, S. Gluck, Y. Tan, I. Chim, J. Cheng, T. Gilbert, D. K. Menon, J. Jannes, T. Kleinig, and S. Koblar, "Prediction of general medical admission length of stay with natural language processing and deep learning: a pilot study," *Internal and emergency medicine*, vol. 15, pp. 989–995, 2020.

[236] W. Chen, Z. Lu, L. You, L. Zhou, J. Xu, K. Chen *et al.*, "Artificial intelligence–based multimodal risk assessment model for surgical site infection (amrams): development and validation study," *JMIR medical informatics*, vol. 8, no. 6, p. e18186, 2020.

[237] C.-H. Chen, J.-G. Hsieh, S.-L. Cheng, Y.-L. Lin, P.-H. Lin, and J.-H. Jeng, "Early short-term prediction of emergency department length of stay using natural language processing for low-acuity outpatients," *The American journal of emergency medicine*, vol. 38, no. 11, pp. 2368–2373, 2020.

[238] M. Fernandes, R. Mendes, S. M. Vieira, F. Leite, C. Palos, A. Johnson, S. Finkelstein, S. Horng, and L. A. Celi, "Predicting intensive care unit admission among patients presenting to the emergency department using machine learning and natural language processing," *PloS one*, vol. 15, no. 3, p. e0229331, 2020.

[239] ——, "Risk of mortality and cardiopulmonary arrest in critical patients presenting to the emergency department using machine learning and natural language processing," *PLoS One*, vol. 15, no. 4, p. e0230876, 2020.

[240] C. A. Hane, V. S. Nori, W. H. Crown, D. M. Sanghavi, and P. Bleicher, "Predicting onset of dementia using clinical notes and machine learning: case-control study," *JMIR Medical Informatics*, vol. 8, no. 6, p. e17819, 2020.

[241] J. L. Izquierdo, J. Ancochea, S. C.-. R. Group, and J. B. Soriano, "Clinical characteristics and prognostic factors for intensive care unit admission of patients with covid-19: retrospective study using machine learning and natural language processing," *Journal of medical Internet research*, vol. 22, no. 10, p. e21801, 2020.

[242] Z. T. Korach, J. Yang, S. C. Rossetti, K. D. Cato, M.-J. Kang, C. Knaplund, K. O. Schnock, J. P. Garcia, H. Jia, J. M. Schwartz *et al.*, "Mining clinical phrases from nursing notes to discover risk factors of patient deterioration," *International journal of medical informatics*, vol. 135, p. 104053, 2020.

[243] Y. Li, P. Nair, X. H. Lu, Z. Wen, Y. Wang, A. A. K. Dehaghi, Y. Miao, W. Liu, T. Ordog, J. M. Biernacka *et al.*, "Inferring multimodal latent topics from electronic health records," *Nature communications*, vol. 11, no. 1, p. 2536, 2020.

[244] B. P. Roquette, H. Nagano, E. C. Marujo, and A. C. Maiorano, "Prediction of admission in pediatric emergency department with deep neural networks and triage textual data," *Neural Networks*, vol. 126, pp. 170–177, 2020.

[245] L. Sterckx, G. Vandewiele, I. Dehaene, O. Janssens, F. Ongenae, F. De Backere, F. De Turck, K. Roelens, J. Decruyenaere, S. Van Hoecke *et al.*, "Clinical information extraction for preterm birth risk prediction," *Journal of Biomedical Informatics*, vol. 110, p. 103544, 2020.

[246] M. Topaz, K. Woo, M. Ryvicker, M. Zolnoori, and K. Cato, "Home health care clinical notes predict patient hospitalization and emergency department visits," *Nursing research*, vol. 69, no. 6, p. 448, 2020.

[247] R. Weegar and K. Sundström, "Using machine learning for predicting cervical cancer from swedish electronic health records by mining hierarchical representations," *PloS one*, vol. 15, no. 8, p. e0237911, 2020.

[248] T. Oliwa, B. Furner, J. Schmitt, J. Schneider, and J. P. Ridgway, "Development of a predictive model for retention in hiv care using natural language processing of clinical notes," *Journal of the American Medical Informatics Association*, vol. 28, no. 1, pp. 104–112, 2021.

[249] H. Dong, V. Suárez-Paniagua, W. Whiteley, and H. Wu, "Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation," *Journal of biomedical informatics*, vol. 116, p. 103728, 2021.

[250] S. Hu, F. Teng, L. Huang, J. Yan, and H. Zhang, "An explainable cnn approach for medical codes prediction from clinical text," *BMC Medical Informatics and Decision Making*, vol. 21, pp. 1–12, 2021.

[251] E. L. Barber, R. Garg, C. Persenaire, and M. Simon, "Natural language processing with machine learning to predict outcomes after ovarian cancer surgery," *Gynecologic oncology*, vol. 160, no. 1, pp. 182–186, 2021.

[252] M. Levis, C. L. Westgate, J. Gui, B. V. Watts, and B. Shiner, "Natural language processing of clinical mental health notes may add predictive value to existing suicide risk models," *Psychological medicine*, vol. 51, no. 8, pp. 1382–1391, 2021.

[253] E. Klang, B. R. Kummer, N. S. Dangayach, A. Zhong, M. A. Kia, P. Timsina, I. Cossentino, A. B. Costa, M. A. Levin, and E. K. Oermann, "Predicting adult neuroscience intensive care unit admission from emergency department triage using a retrospective, tabular-free text machine learning approach," *Scientific reports*, vol. 11, no. 1, p. 1381, 2021.

[254] H. Yang, L. Kuang, and F. Xia, "Multimodal temporal-clinical note network for mortality prediction," *Journal of Biomedical Semantics*, vol. 12, no. 1, pp. 1–14, 2021.

[255] B.-H. Kim, Z. Deng, P. S. Yu, and V. Ganapathi, "Can current explainability help provide references in clinical notes to support humans annotate medical codes?" *arXiv preprint arXiv:2210.15882*, 2022.

[256] G. López-García, J. M. Jerez, N. Ribelles, E. Alba, and F. J. Veredas, "Explainable clinical coding with in-domain adapted transformers," *Journal of Biomedical Informatics*, vol. 139, p. 104323, 2023.

[257] U. Ahmed, G. Srivastava, U. Yun, and J. C.-W. Lin, "Eandc: An explainable attention network based deep adaptive clustering model for mental health treatment," *Future Generation Computer Systems*, vol. 130, pp. 106–113, 2022.

[258] M. Z. Uddin, K. K. Dysthe, A. Følstad, and P. B. Brandtzaeg, "Deep learning for prediction of depressive symptoms in a large textual dataset," *Neural Computing and Applications*, vol. 34, no. 1, pp. 721–744, 2022.

[259] Y. Qu, X. Deng, S. Lin, F. Han, H. H. Chang, Y. Ou, Z. Nie, J. Mai, X. Wang, X. Gao *et al.*, "Using innovative machine learning methods to screen and identify predictors of congenital heart diseases," *Frontiers in Cardiovascular Medicine*, vol. 8, p. 2087, 2022.

[260] E. S. Kavvas, L. Yang, J. M. Monk, D. Heckmann, and B. O. Palsson, "A biochemically-interpretable machine learning classifier for microbial gwas," *Nature communications*, vol. 11, no. 1, p. 2580, 2020.

[261] T. Sun, Y. Wei, W. Chen, and Y. Ding, "Genome-wide association study-based deep learning for survival prediction," *Statistics in medicine*, vol. 39, no. 30, pp. 4605–4620, 2020.

[262] S. Dey, A. Bose, P. Chakraborty, M. Ghalwash, A. G. Saenz, F. Utro, K. Ng, J. Hu, L. Parida, and D. Sow, "Impact of clinical and genomic factors on sars-cov2 disease severity," *medRxiv*, 2021.

[263] A. van Hilten, S. A. Kushner, M. Kayser, M. A. Ikram, H. H. Adams, C. C. Klaver, W. J. Niessen, and G. V. Roshchupkin, "Gennet framework: interpretable deep learning for predicting phenotypes from genetic data," *Communications biology*, vol. 4, no. 1, p. 1094, 2021.

[264] Q. Yan, Y. Jiang, H. Huang, A. Swaroop, E. Y. Chew, D. E. Weeks, W. Chen, and Y. Ding, "Genome-wide association studies-based machine learning for prediction of age-related macular degeneration risk," *Translational vision science & technology*, vol. 10, no. 2, pp. 29–29, 2021.

[265] L. Zhao, Q. Dong, C. Luo, Y. Wu, D. Bu, X. Qi, Y. Luo, and Y. Zhao, "Deepomix: A scalable and interpretable multi-omics deep learning framework and application in cancer survival analysis," *Computational and structural biotechnology journal*, vol. 19, pp. 2719–2725, 2021.

[266] H. Turbé, M. Bjelogrlic, C. Lovis, and G. Mengaldo, "Evaluation of post-hoc interpretability methods in time-series classification," *Nature Machine Intelligence*, pp. 1–11, 2023.

[267] R. Elshawi, M. H. Al-Mallah, and S. Sakr, "On the interpretability of machine learning-based model for predicting hypertension," *BMC medical informatics and decision making*, vol. 19, no. 1, pp. 1–32, 2019.

[268] X. Pang, C. B. Forrest, F. Lê-Scherban, and A. J. Masino, "Understanding early childhood obesity via interpretation of machine learning model predictions," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. IEEE, 2019, pp. 1438–1443.

[269] J. Lu, R. Jin, E. Song, M. Alrashoud, K. N. Al-Mutib, and M. S. Al-Rakhami, "An explainable system for diagnosis and prognosis of covid-19," *IEEE Internet of Things Journal*, vol. 8, no. 21, pp. 15 839–15 846, 2020.

[270] T. Dissanayake, T. Fernando, S. Denman, S. Sridharan, H. Ghaemmaghami, and C. Fookes, "A robust interpretable deep learning classifier for heart anomaly detection without segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 6, pp. 2162–2171, 2020.

[271] X. Song, A. S. Yu, J. A. Kellum, L. R. Waitman, M. E. Matheny, S. Q. Simpson, Y. Hu, and M. Liu, "Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction," *Nature communications*, vol. 11, no. 1, pp. 1–12, 2020.

[272] A. J. Barda, C. M. Horvat, and H. Hochheiser, "A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare," *BMC medical informatics and decision making*, vol. 20, no. 1, pp. 1–16, 2020.

[273] S. Penafiel, N. Baloian, H. Sanson, and J. A. Pino, "Predicting stroke risk with an interpretable classifier," *IEEE Access*, vol. 9, pp. 1154–1166, 2020.

[274] J. Hatwell, M. M. Gaber, and R. M. Atif Azad, "Ada-whips: explaining adaboost classification with applications in the health sciences," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, pp. 1–25, 2020.

[275] N. Beebe-Wang, A. Okeson, T. Althoff, and S.-I. Lee, "Efficient and explainable risk assessments for imminent dementia in an aging cohort study," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2409–2420, 2021.

[276] S.-H. Kim, E.-T. Jeon, S. Yu, K. Oh, C. K. Kim, T.-J. Song, Y.-J. Kim, S. H. Heo, K.-Y. Park, J.-M. Kim *et al.*, "Interpretable machine learning for early neurological deterioration prediction in atrial fibrillation-related stroke," *Scientific reports*, vol. 11, no. 1, pp. 1–9, 2021.

[277] M. Rashed-Al-Mahfuz, A. Haque, A. Azad, S. A. Alyami, J. M. Quinn, and M. A. Moni, "Clinically applicable machine learning approaches to identify attributes of chronic kidney disease (ckd) for use in low-cost diagnostic screening," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 9, pp. 1–11, 2021.

[278] X. Zeng, Y. Hu, L. Shu, J. Li, H. Duan, Q. Shu, and H. Li, "Explainable machine-learning predictions for complications after pediatric congenital heart surgery," *Scientific reports*, vol. 11, no. 1, pp. 1–11, 2021.

[279] Y. Zhang, D. Yang, Z. Liu, C. Chen, M. Ge, X. Li, T. Luo, Z. Wu, C. Shi, B. Wang *et al.*, "An explainable supervised machine learning predictor of acute kidney injury after adult deceased donor liver transplantation," *Journal of translational medicine*, vol. 19, no. 1, pp. 1–15, 2021.

[280] A. Pal and M. Sankarasubbu, "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*, 2021, pp. 620–628.

[281] M. A. Alves, G. Z. Castro, B. A. S. Oliveira, L. A. Ferreira, J. A. Ramírez, R. Silva, and F. G. Guimarães, "Explaining machine learning based diagnosis of covid-19 from routine blood tests with decision trees and criteria graphs," *Computers in Biology and Medicine*, vol. 132, p. 104335, 2021.

[282] A. Moncada-Torres, M. C. van Maaren, M. P. Hendriks, S. Siesling, and G. Geleijnse, "Explainable machine learning can outperform cox regression predictions and provide insights in breast cancer survival," *Scientific Reports*, vol. 11, no. 1, pp. 1–13, 2021.

[283] C. Duckworth, F. P. Chmiel, D. K. Burns, Z. D. Zlatev, N. M. White, T. W. Daniels, M. Kiuber, and M. J. Boniface, "Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during covid-19," *Scientific reports*, vol. 11, no. 1, pp. 1–10, 2021.

[284] A. M. Antoniadi, M. Galvin, M. Heverin, O. Hardiman, and C. Mooney, "Prediction of caregiver quality of life in amyotrophic lateral sclerosis using explainable machine learning," *Scientific Reports*, vol. 11, no. 1, pp. 1–13, 2021.

[285] I. R. Ward, L. Wang, J. Lu, M. Bennamoun, G. Dwivedi, and F. M. Sanfilippo, "Explainable artificial intelligence for pharmacovigilance: What features are important when predicting adverse outcomes?" *Computer Methods and Programs in Biomedicine*, vol. 212, p. 106415, 2021.

[286] O. Zhang, C. Ding, T. Pereira, R. Xiao, K. Gadhoumi, K. Meisel, R. J. Lee, Y. Chen, and X. Hu, "Explainability metrics of deep convolutional networks for photoplethysmography quality assessment," *IEEE Access*, vol. 9, pp. 29 736–29 745, 2021.

[287] L. M. Thimoteo, M. M. Vellasco, J. Amaral, K. Figueiredo, C. L. Yokoyama, and E. Marques, "Explainable artificial intelligence for covid-19 diagnosis through blood test variables," *Journal of Control, Automation and Electrical Systems*, vol. 33, no. 2, pp. 625–644, 2022.

[288] J. Hao, S. C. Kosaraju, N. Z. Tsaku, D. H. Song, and M. Kang, "Page-net: interpretable and integrative deep learning for survival analysis using histopathological images and genomic data," in *Pacific Symposium on Biocomputing 2020*. World Scientific, 2019, pp. 355–366.

[289] H. Ren, A. B. Wong, W. Lian, W. Cheng, Y. Zhang, J. He, Q. Liu, J. Yang, C. J. Zhang, K. Wu *et al.*, "Interpretable pneumonia detection by combining deep learning and explainable models with multisource data," *IEEE Access*, vol. 9, pp. 95 872–95 883, 2021.

[290] R. J. Chen, M. Y. Lu, D. F. Williamson, T. Y. Chen, J. Lipkova, Z. Noor, M. Shaban, M. Shady, M. Williams, B. Joo *et al.*, "Pan-cancer integrative histology-genomic analysis via multimodal deep learning," *Cancer Cell*, vol. 40, no. 8, pp. 865–878, 2022.

[291] E. Herrett, A. M. Gallagher, K. Bhaskaran, H. Forbes, R. Mathur, T. Van Staa, and L. Smeeth, "Data resource profile: clinical practice research datalink (cprd)," *International journal of epidemiology*, vol. 44, no. 3, pp. 827–836, 2015.

[292] P. Kirkpatrick, "New clues in the acetaminophen mystery," *Nature Reviews Drug Discovery*, vol. 4, no. 11, pp. 883–883, 2005.