The Role of Explainability in Assuring Safety of Machine Learning in Healthcare

YAN JIA ⓘ, JOHN MCDERMID ⓘ, TOM LAWTON ⓘ, AND IBRAHIM HABLI ⓘ

Yan Jia, John McDermid, and Ibrahim Habli are with the Department of Computer Science, University of York, YO10 5DD York, U.K.
Tom Lawton is with Bradford Royal Infirmary, Bradford Institute for Health Research, BD9 6RJ Bradford, U.K.
CORRESPONDING AUTHOR: YAN JIA (yan.jia@york.ac.uk)

ABSTRACT Established approaches to assuring safety-critical systems and software are difficult to apply to systems employing ML where there is no clear, pre-defined specification against which to assess validity. This problem is exacerbated by the "opaque" nature of ML where the learnt model is not amenable to human scrutiny. Explainable AI (XAI) methods have been proposed to tackle this issue by producing human-interpretable representations of ML models which can help users to gain confidence and build trust in the ML system. However, little work explicitly investigates the role of explainability for safety assurance in the context of ML development. This paper identifies ways in which XAI methods can contribute to safety assurance of ML-based systems. It then uses a concrete ML-based clinical decision support system, concerning weaning of patients from mechanical ventilation, to demonstrate how XAI methods can be employed to produce evidence to support safety assurance. The results are also represented in a safety argument to show where, and in what way, XAI methods can contribute to a safety case. Overall, we conclude that XAI methods have a valuable role in safety assurance of ML-based systems in healthcare but that they are not sufficient in themselves to assure safety.

INDEX TERMS Explainability, machine learning, safety assurance

I. INTRODUCTION

In healthcare, machine learning (ML) is used on various problems, e.g., learning optimal treatments, or to detect abnormalities in radiology images, where it has achieved outstanding performance. However, assuring safety for such systems employing ML remains a challenge. In many domains there are well-established approaches and standards for assuring safety-critical systems and software. Assurance means establishing *justified confidence* in the system for its intended use. The assurance principles underlying these standards include validating that the system works as intended and verifying that the system meets explicit *safety* requirements. These assurance principles remain essential for systems employing ML. However, the details of these approaches and standards can be difficult to apply where systems use ML.

First, the established approaches are based, implicitly or explicitly, on the *V life-cycle model* moving from requirements, through design onto implementation then testing. In contrast, the development of ML-based systems follows a very different, much more iterative, life-cycle with four main phases: data management, ML algorithm selection, model learning, and model verification & validation, which makes it hard to apply established methods. Some emerging standards and guidance better reflect the ML life-cycle, e.g., the US Federal Drug Administration (FDA) proposed regulatory framework on AI/ML-based Software as a Medical Device (SaMD) [1] and Assurance of Machine Learning for Autonomous Systems (AMLAS) [2].

Second, because of the "black box" (opaque) nature of the ML models [3], it is hard to assess what has been learnt, which exacerbates the challenging of defining concrete requirements for the safety of SaMD in its clinical context. Instead, human performance is often used as a "gold standard" and the current practice is often to (seek to) achieve performance that is better than humans. This makes validation difficult as human performance is variable both from individual-to-individual and over time for a single individual. Also, performance will vary from patient-to-patient, e.g., with comorbidities, and clinicians might not agree on the best treatment strategy.

To overcome such problems, the ML community is actively studying "explainablity," which is intended to "peek inside the black box" and to illuminate the underlying workings of the ML models. Explainability is often equated with producing explainable artificial intelligence (XAI) methods, which seek to provide human interpretable representations of ML models [4]. Although there is considerable variation in the definition of terms such as explainability, interpretability and transparency, in this paper we adopt the view from the FDA AI/ML-enabled Medical Devices Transparency Workshop [5] that explainability is one component of transparency. Transparency is a much broader concept in their definition and we see interetability as a necessary facet of explainability, as suggested by Gilpin *et al.* [6].

In this paper we consider the role of explainability in assuring the safety of ML models in healthcare. Our focus is on development activities and deployment decisions for ML-based systems, but we briefly consider the potential role of explainability in operations.

The primary contributions of this paper are as follows. First, we developed a new conceptual model, a spider diagram, which gives a heuristic view of safety and shows how safety, as a cross-cutting concern, relates to other aspects of an ML-based system, including the role of explainability. Then we show how to assure safety by meeting relevant regulatory requirements, especially from the FDA. In particular, we show how XAI methods can be integrated into the different phases of model development and what types of XAI methods best provide safety evidence to meet the regulatory requirements. Finally, we present a concrete case study to illustrate the use of XAI methods in supporting a safety case for ML systems.

The rest of the paper is structured as follows. Section II identifies relevant related work. Section III outlines the potential role of XAI methods in the ML life-cycle. These possibilities are then investigated in Section IV using an example of weaning patients from mechanical ventilation. This is followed by a discussion and conclusions in Sections V and 6, respectively.

II. BACKGROUND & RELATED WORK

This section discusses established approaches to assurance of safety-critical systems and identifies their limitations when dealing with systems employing ML. This is followed by an introduction to the concepts of explainability and an overview of the different types of XAI methods.

A. ESTABLISHED ASSURANCE APPROACHES AND THE CHALLENGES OF ML

We use the term *assurance* to mean *confidence that the system behaviour is as intended in the environment of use*, where *as intended* includes being safe. In this context, we are interested in assurance of patient safety when ML-based systems are used in a healthcare context.

Most approaches to assurance emphasise verification and validation, although the definitions of the terms can vary. The International Medical Devices Regulator Forum (IMDRF) define the terms as follows:

- Verification – confirmation through provision of objective evidence that specified requirements have been fulfilled [7];
- Validation – confirmation through provision of objective evidence that the requirements for a specific intended use or application have been fulfilled [7].

To interpret these definitions we can say that validation is concerned with *building the right system*, including defining requirements that meet our intent and that verification is concerned with *building the system right* by demonstrating that the system meets these requirements. Verification and validation (V&V) need to encompass identified safety requirements, which are often derived to control the risk associated with *hazards*, i.e., undesirable situations that pose risk to life. Typically, risk is a combination of the likelihood and the severity of the harm arising from the hazard, although the detailed computations vary from domain to domain. Where risks are deemed too high, derived safety requirements (DSRs) are identified to reduce the likelihood of hazard occurrence, e.g., by controlling hazard causes, or to mitigate the consequences of the hazard should it arise. In healthcare, such ideas underpin some of the relevant standards, e.g., [8] produced by NHS Digital in the U.K.

We have previously investigated how to adapt traditional safety engineering processes to healthcare systems which employ ML. We have shown that, in some cases, it is possible to adapt classical safety methods to identify hazards and then to establish DSRs on the ML elements of systems [9]. Clinical judgement is also needed to produce such DSRs. Where requirements are not stated explicitly, XAI methods can help by providing explanations that enable direct validation of the ML model as a whole, e.g., showing that predictions are based on valid clinical factors and are consistent with clinical knowledge.

In many domains, including healthcare, it is accepted good practice for the safety work to culminate in the production of a safety or assurance case, see [8]. In general, a safety case is "an argument, supported by evidence, that a systems is safe to be deployed in its context of use". It is common to express the argument graphically, e.g., using the Goal Structuring Notation (GSN) [10], as a means of making the argument clear and open to review. The evidence underpinning the safety case includes the results of hazard and risk analysis, as well as the outputs from V&V activities and in this paper we will show that XAI methods can also provide such evidence.

There are a number of initiatives concerned with the assurance of ML in safety-critical systems both in healthcare and more generally. For example, AMLAS defines a process for assurance of the safety of ML-based systems to reflect the ML development life-cycle, which identifies both evidence artefacts and argument patterns (standard forms of argument that can be instantiated for a particular system) in GSN. AMLAS also considers issues of the robustness of ML-based systems, e.g., response to unexpected inputs. The FDA also

TABLE 1. Categorisation of XAI methods with examples

Type of explanation		Scope	Model Specific/ Agnostic	Examples of XAI methods
Interpretable Models		Global	Specific	A model by itself interpretable, e.g. linear/logistic regression, decision tree, GAM, etc.
Post-hoc explanation	Explanation by Approximation	Local	Agnostic	LIME
		Local	Agnostic	Anchors
		Global	Agnostic	Model extraction
	Explanation by example	Local	Agnostic	Counterfactual examples
		Local	Agnostic	Adversarial examples
		Local	Agnostic	Influential instances
		Local	Agnostic	Prototypical examples
	Feature relevance Explanation	Local	Agnostic	KernelSHAP
		Local	Specific	TreeSHAP
		Local	Specific	LRP
		Local	Specific	Integrated Gradient
		Local	Specific	DeepLIFT
	Visual Explanation	Global	Agnostic	Partial dependence plot
		Local	Agnostic	ICE

proposed a total life-cycle regulatory approach for ML-based SaMD [1]. However, these approaches are evolving in that they provide good high-level guidance and objectives, but how to meet such objectives is not sufficiently detailed. The work we present here is intended to be complementary to, and build on, these approaches and shows how XAI methods can provide evidence to meet these objectives, and thus contributes to improving their maturity.

In addition, it is always desirable to consider assurance "through life," as proposed by the FDA [1], not just as an activity undertaken prior to deployment. This includes getting feedback from operations to check whether or not the assumptions made in pre-deployment assurance activities are sound. This is even more important for ML-based systems than it is for "conventional" systems because of the opacity of ML models.

B. EXPLAINABLE AI METHODS

ML includes a range of different methods such as decision trees, support vector machines and neural networks (NNs). The study of XAI methods seeks to provide insight into how and why ML models make such predictions. Work on explainable AI includes formalising definitions of explainability, development of XAI methods themselves and establishing evaluation methods. In this section we provide a brief overview of XAI methods. For a more complete view of XAI, please refer to some well-cited surveys, e.g., [6], [11]. There are many different ways to categorise XAI methods, e.g., local or global based on the scope of the explanation or model agnostic or specific based on whether the XAI methods can work for any class of ML models or only work for a specific class of models. Here we adopt the taxonomy of XAI in [11] and present XAI methods in two different classes based on the explanation generating mechanism, as shown in Table 1.

Some ML models are perceived as intrinsically interpretable to the user, so we refer to these as *interpretable models*. This includes linear/logistic regression, decision trees, K-nearest neighbours, decision rules, Bayesian models, general

additive models (GAMs), etc. Note that, although often these models are viewed as intrinsically interpretable, when the number of input features are beyond human ability to grasp or when the input features are heterogeneous which is not uncommon in healthcare, it can be difficult for humans to interpret the model and care needs to be taken [12].

When it comes to explaining more complex or opaque ML models, e.g., support vector machines, tree ensembles, and NNs, which are not intrinsically interpretable, a *post-hoc* explanation can be used to provide insights without knowing the mechanisms by which the model works. We present four main *post-hoc* explanation classes as follows along with some popular techniques as illustrations.

- *Explanation by approximation* aims to use surrogate models, e.g., linear models, decision trees or decision rules to approximate the underlying complex or opaque model. These can include local or global surrogates depending on whether they are approximating a single prediction or the whole model. For example, *LIME* (Local Interpretable Model-Agnostic Explanations) [13] focuses on training a local surrogate to provide explanations for an individual prediction, which is based on the assumption that it is possible to fit a surrogate model around a single input sample that mimics the local behaviour of the complex ML model. Like LIME, *Anchors* [14] deploy a perturbation based strategy to generate local explanations for predictions in a local region resulting in if-then rules. In contrast, *model extraction* proposed in [15] trains a global surrogate to approximate a complex model. The three methods mentioned above are model-agnostic so they would also work for NNs.

- *Explanation by example* explains the ML model by selecting particular instances from the dataset or by creating new instances. It comprises counterfactual examples, adversarial examples, influential instances and prototypical examples. *Counterfactual examples* can be thought of as "what is not, but could have been" and are intended to produce a sparse human-interpretable example by changing some input features to achieve a different output, i.e., the user's desired output (what could have been). *Adversarial examples* are typically generated by adding small, intentional perturbations to the input features to cause an ML model to make an incorrect prediction. Adversarial examples are intended to deceive the ML model instead of interpreting the model. Therefore, the changes in the inputs are often imperceptible for a human observer. *Influential instances* are intended to identify which input instances have a strong effect on the trained model. They can be identified by measuring the impact of a training point on a particular prediction or on the model overall. *Prototypical examples* can summarise and represent a complex underlying data distribution, which then can be used to provide a global understanding of the model by examining prototypes along with their model predictions or a

local explanation for a specific instance by identifying the most similar training instance according to the trained model. This is different to influential instances because the training example might be influential but not representative.

- *Feature relevance explanation* techniques rank or score the input features based on their influence, relevance or importance on the model prediction where higher scores mean that the corresponding features are more important for the model. Such scores are often obtained by perturbation or gradient-based methods. For example, *SHAP* (SHapley Additive exPlanations) [16] is one of the perturbation methods based on Shapley values, which are used to explain a model prediction by treating input features as the players in a cooperative game and the model prediction as the gain resulting from the game. It includes KernelSHAP, a model agnostic weighted linear regression approximation of the exact Shapley value inspired by LIME, and TreeSHAP, a model-specific efficient estimation approach for tree-based models. The work on SHAP has wider significance as it has defined a new class of additive feature importance measures, unifying several existing XAI methods [16].
- *Visual explanation* techniques aim to facilitate model understanding by using visualisation, e.g., showing how features interact with the predicted output or other features. Such methods can involve using sensitivity analysis (SA) or partial dependence to inspect the relationship between the uncertainty in the predicted output and its input features, e.g., [17] presented several visualisations for the SA results and [18] introduced Individual Conditional Expectation (ICE) to show how the prediction of a particular instance changes along with the input features.

Due to the popularity of deep learning (DL), a different classification scheme for XAI in DL is proposed by Gilpin *et al.* [6], and it is often treated as a subfield of XAI *per se*.

- *Explanations of deep network processing*. This can be achieved by producing a "saliency map" which is a rendering of weights for the input features that highlight the salient features for the prediction. A "saliency map" can be produced by perturbation-based methods, e.g., LIME, or by calculating the gradient of the output with respect to the input, identifying which parts of the input have a significant influence on the classification [19]. Due to some limitations of directly using gradients, e.g., saturation, there are also a number of other methods proposed, for example, Integrated Gradients [20], LRP [21], and DeepLIFT [22].
- *Explanations of deep network representations*. This type of explanation aims to inspect what the model learnt. Feature visualisation [23] is helpful to understand how an NN builds up its understanding of input images throughout the network by maximising activation for the unit of interest, e.g., a specific neuron, or a
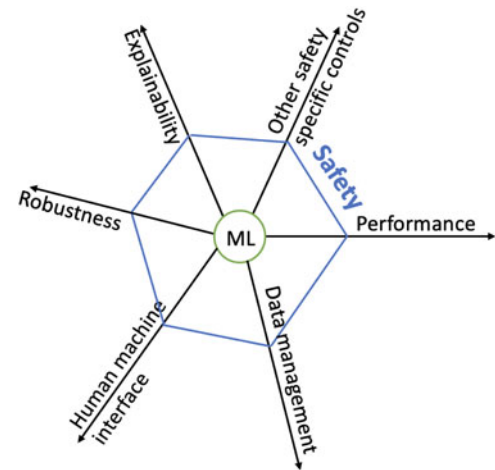

FIGURE 1. Spider diagram illustrating the role of explainability and safey.

specific layer, or a convolution channel. Further, there are concept-based methods, e.g., TCAV [24], attempting to detect concepts that are human-interpretable but embedded within the latent space learnt by the network.

- *Explanation-producing systems*. For example, attention-based networks learn a function by providing weights of the input or internal features of a NN in order to force the model to attend to the important regions with respect to the target task. Although the attention-mechanism could render an attention map to provide some insights or intuitive feeling of the model, it is important to note that the interpretation of attention as explanation is currently the subject of debate [25], [26].

Some further details on XAI methods we have used in our case study are included in Section IV.

III. THE ROLE OF EXPLAINABILITY IN SAFETY ASSURANCE

Assurance of safety is a multi-faceted, multi-dimensional concern. We start by presenting our new conceptual model as a spider diagram to show the role of explainability in safety assurance generally, and how this relates to regulatory requirements. Then, we exemplify how explainability can contribute to safety assurance in the context of the ML life-cycle.

A. EXPLAINABILITY AND SAFETY

Figure 1, which we refer to as the *spider diagram*, shows the different aspects of a ML-based system, e.g., performance and explainability, that relate to safety in its context of use. This illustrates that safety is a cross-cutting concern, rather than being a separate dimension. Figure 1 gives an impression of safety, where safety is related to six dimensions of ML-based systems and the area within the hexagon can be viewed as a heuristic evaluation of safety. Note that the dimensions in Figure 1 are not exhaustive; we have chosen them as they provide a good engineering perspective.

Performance of the ML can be assessed using accuracy or other metrics, which are well known. *Data management* for ML is crucial as the quality of data has a significant impact on the model learnt, e.g., if the data is biased, then the model might also learn discriminatory behaviour or even amplify it. *Human-machine interface* is also very important for safety as certain types of interface can be prone to human operator errors. Here we think of *Robustness* as the model prediction being stable, particularly in the presence of small variations in the input features. *Explainability* can be achieved using intrinsically interpretable models or *post-hoc* XAI methods. However, explanations of intrinsically interpretable models are more accurate than *post-hoc* XAI methods, in the sense that they have high fidelity to the task model, and this has led to some authors suggesting that only such models should be used for safety-critical tasks [27]. Finally, *Other safety specific controls* can be thought of as means to satisfy DSRs at the system level, which are not covered by the above dimensions. This is task dependent, e.g., in an online healthcare triage system, it is important to have a safety net where specific words should trigger an emergency response rather than continuing to ask questions. As shown in Figure 1, performance also matters for safety but if an intrinsically interpretable model can achieve similar performance to a "black box" model, then the interpretable one should be preferred.

To draw the spider diagram for a particular ML-based system requires metrics for the different dimensions. Performance is readily quantified, but it is less obvious how to measure or quantify the other dimensions; we return to this issue in Section V. In theory, if we could identify all the relevant dimensions and quantify the area in the spider diagram this would give a good basis for safety assurance. In practice, regulators are defining regulatory requirements with the intent that if these requirements are satisfied, the system can be approved as safe enough to be marketed.

In the next subsection, we focus on explanability and further investigate how it can contribute to safety in the context of the ML life-cycle, and show how the evidence generated by XAI methods can contribute to meeting the relevant FDA regulatory requirements.

B. EXPLAINABILITY IN THE ML LIFE-CYCLE
The development process for ML typically includes data management, ML algorithm selection, model learning and model V&V [2], as shown in Figure 2. Figure 2 makes explicit the need for a deployment decision prior to operation (which may be supported by a safety case). It also shows the stakeholders who might be interested in the explanations in the different phases. Our focus here is on the development activities, but we briefly consider the potential role of explainability in operation, see Section III; for a discussion of the wider role of explainability including incident and accident investigation see [28].

The rest of this section discusses the role of explainability for each stage of the process shown in Figure 2.
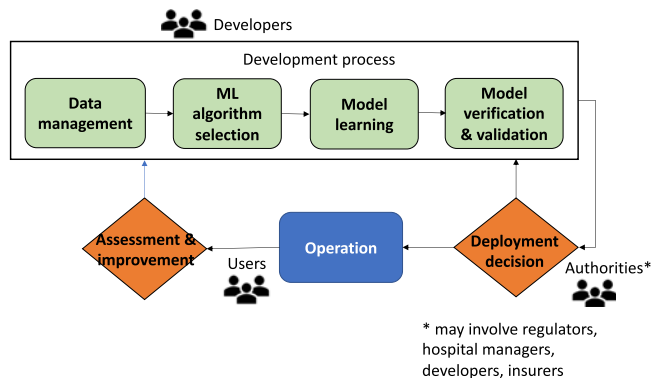

FIGURE 2. Process for development and use of an ML System.

1) DATA MANAGEMENT
The first phase of the ML development process is data management, and this aligns directly with the spider diagram. Most XAI methods are not applicable to this stage but *prototypes* are relevant; they can help to understand the datasets especially when the datasets are large and complex, although *prototypes* can also be used to approximate the learnt model [29].

The Royal Society's Policy Briefing on XAI emphasises that data quality and provenance is part of the explainability pipeline, specifically saying that *"Understanding the quality and provenance of the data used in AI systems is therefore an important part of ensuring that a system is explainable"* [30]. This includes showing that the data comes from appropriate sources for the problem addressed. A widely accepted, harmonised framework for assessment of Electronic Health Record (EHR) data quality highlights *conformance*, *completeness* and *accuracy* [31]; we prefer *accuracy* to the original term plausibility because plausibility means that the values are in the possible range but accurate means that the data is not only possible but correct. These criteria would be applicable to any ML systems developed using EHR data. Further, we also identify data *relevance* and *balance* as being particularly important to ML model development [2]. As real world data may contain biases, errors, or be incomplete, explaining how these five criteria are met can be at least as important as explaining the ML model itself.

The evidence to ensure data quality is essentially technical, for example data *conformance* would include showing that data observes defined formats, e.g., correct units for weight [31]. However, demonstrating *data relevance* and *data balance* would include a judgement that the training data contained clinically relevant factors and are balanced for the problem being addressed. We acknowledge that often it is not possible to choose data that gives both feature balance and class balance. Instead, it might be useful to explain that some important features are reasonably balanced, e.g., gender, if the model is intended to be used for both male and female patients. Class balance has long been an active research area in the ML community. In the case of skewed dataset, *prototypes* can be generated to understand the data distribution and can be used to train a model. There is evidence showing that

using *prototypes* can help with class balance. For example, Gurumoorthy *et al.* [32] have demonstrated that using good *prototypes* to train a model can give better performance than using the whole dataset or randomly sampled subsets to balance the classes.

It should be noted that data management is both crucial and labour intensive. Indeed, it may consume more effort than the rest of the ML life-cycle. Judgement of the extent to which the data meets these five criteria would be used to assess an AI/ML system in the *data management* dimension in the spider diagram.

2) ML ALGORITHM SELECTION

The second phase in the development process is ML algorithm selection (also referred to as model selection, here we use the term ML algorithm selection to avoid the confusion with model selection in the training phase where the ML algorithm is the same but hyperparameters of the model are tuned to be different). It is important to understand what kind of problem is being addressed and what kind of ML methods are suitable for the problem at hand, e.g., classification, regression, or finding an optimal policy. Another important aspect to consider at this stage is the *explainability* and *performance* dimensions of the ML model as shown in the spider diagram. In Section II we identified that some ML models are intrinsically interpretable whereas others need to be supplemented with *post-hoc* XAI methods. Guidelines on ML algorithm selection, balancing model performance against explainability, have been proposed [33].

When it comes to ML algorithm selection, safety requirements are often implicitly transformed into explainability and performance requirements. Note that sometimes people make statements such as "use of deep NNs is not safe". When they make this kind of statement, they are implicitly making the judgement that deep NNs are opaque, i.e., not interpretable. The spider diagram helps to show that this is over-simplistic. There is no binary choice "opaque/interpretable," rather *explainability* and *performance* are two of the dimensions; both matter for safety in an AI/ML application. This is why we argue that safety requirements are partially, but not wholly, transformed into explainability requirements. It would be ideal to have an interpretable model which can achieve performance as high as black box models. When this is not the case, a trade-off between explainability and performance would be necessary [33] and *post-hoc* explanations should be considered either in later phases of development or in operation to produce effective explanation. The rationale for the ML algorithm choice, including the performance-explainability trade-offs, needs to be documented in the safety case.

3) MODEL LEARNING

The third phase in the development process is model learning. The essential aim of this stage is to train a "good" model, and *performance* and *robustness* are especially important at this stage. For model learning, hyperparameter selection, loss function definition and class balance need to be considered

in order to meet safety requirements. In addition, XAI methods can have a role in improving both *performance* and *robustness*, e.g., through model debugging and adversarial training.

Automated robustness improvement – adversarial examples are often added to training data to improve model robustness in object classification tasks. This is referred to as adversarial training or robustness training [34]. There are many techniques for generating adversarial examples, e.g., by minimising the distance between the adversarial example and the input instance, which is similar to counterfactual examples (see Section III for more discussion). Popular distance metrics include: L_0, L_1, L_2 and L_∞, all of which are L_p norms [35]. L_0 counts the number of features that have changed between the two instances, L_1 measures the sum of the magnitudes of the change, L_2 measures the euclidean distance between the two instances, and L_∞ measures the maximum change among all of the features. Use of *adversarial examples* to improve model robustness is becoming widespread in domains such as autonomous vehicles, for example in improving performance at reading road signs under adverse conditions, but we believe it has wider applicability, e.g., for image classification in radiology.

"Model debugging" – influential instances are useful as they can help to understand model behaviour, specifically they can help to debug domain mismatch or fix problematic training instances where the label or input features might be incorrect [36]. We can investigate influential instances for a specific prediction or for the model overall. For example, when the test instances are being misclassified by the model, we can identify the most responsible training instances for such instances. This enables exploration of the causes of the problem, e.g., errors in the training instances, or domain mismatch, i.e., poorly represented subgroups in the dataset, which can therefore be unduly influential. Domain mismatch is not uncommon in healthcare as the population from the intended use hospital can be very different from the population used for developing the model. Further, we can also investigate the most influential training instances for the model overall, e.g., we can assess the influence of removing a certain training instance on the model by measuring the change in loss of this training instance before and after it is removed; the bigger the change in the loss the more influential the instance. Intuitively if the model has to "try hard" to accommodate the instance, then that instance is highly influential. In reality, labels in the training dataset can be noisy and it is unrealistic to expect human experts to manually review all of the data. In this case, it is useful to investigate the most influential instances for the model overall to direct the human experts' attention to the instances that actually matter.

Two approaches for identifying *influential instances* are often used – deletion diagnostics and influence functions. Deletion diagnostics is not practical for big training datasets as it needs to remove a single training instance every time to observe the effect of this instance until the effect of all of the training data has been observed. However, influence functions

can be used to approximate the effect without deleting the training instance (see Section IV for details).

In addition to the uses of the two XAI methods described above, current research is also exploring how to guide the learning process to enable the models to produce the desired form of explanations, for example by including explainable regularisers in the loss function. In [37], the authors penalise the gradient of a NN to force it to focus on regions which contain important information for the task. Others are exploring how to penalise the neighbourhood fidelity [38] in order to improve the quality of the explanation. This is an important research direction, but it relies on understanding what constitutes a good explanation and what metrics enable the "goodness" of the explanation to be evaluated, e.g., neighbourhood fidelity or stability [38]. We expect such methods to be significant for safety assurance, once the understanding of what is a good explainable regulariser is more mature, which depends on progress in the domain of explainability itself.

4) MODEL VERIFICATION AND VALIDATION

The final phase in the development process is model V&V. We believe that explainability has a particular role in validation, but could also have a role in verification if there are specific explainability requirements to verify. However, such explainability requirements need to be defined in a specific situation, therefore our focus here is on validation. We derived three distinct objectives, reconciling approaches proposed by the FDA [1] and the IMDRF [7], which reflect key criteria for use of ML models in healthcare, although we note that explanations cannot guarantee that all these criteria are met [33].

First is *performance*, which can be measured using standard ML practices, e.g., evaluation of the proportion of false positives and false negatives, or the AUC-ROC. This is necessary but not sufficient to assure safety of ML.

The second objective is *analytical* or *technical validation*, showing that the software for the ML models is correctly constructed, and that it is accurate and reliable. Further, the ML model implementations produce repeatable results, giving the same predictions from the same inputs. This objective can be met by employing established safety-critical software development practices including formal specifications, traceability from specification to implementation, use of test coverage criteria and static code analysis methods [39]. We do not see a role for XAI methods for this aspect of validation.

Third is *clinical validation* which measures the ability of the system to generate a clinically meaningful output for its intended use in its operational environment. Here we define two specific sub-objectives where we believe XAI methods have a role in supporting clinical validation:

- *Clinical association* – demonstrate that the association between the system output and the targeted clinical condition in the intended population is supported by evidence;
- *Robustness* – demonstrate the ability to distinguish the different classes of intended condition or recommended

treatment without over-reliance on a specific input feature.

These explicitly relate to the *explainability* and *robustness* dimensions in the spider diagram.

Feature relevance explanations can help to demonstrate clinical association by showing that the output predictions are based on clinically meaningful and relevant factors of the input. This involves ranking input features based on their importance score or contribution score and making the rankings visible to clinicians so that they can exercise clinical judgement. One might argue that rule-based explanation can also help in this case, for example, using Anchors to generate relevant rules for clinicians to judge whether a valid clinical association has been learnt. However, for the following reasons, we believe that feature relevance explanations should be preferred in this context. First, clinicians consistently stated that knowing the feature importance driving the model outcome is crucial. This allows them to compare model decisions to their clinical judgement, especially in case of a discrepancy [40]. Second, ML is most valuable when it is used in complex clinical tasks where there are no agreed set of rules. In this case, it is easier to agree on the important and unimportant factors than on rules. For example, there is evidence [41] showing that even if an extracted rule provides 100% accuracy in the test dataset, an individual clinician still might not have confidence in it, let alone achieving consensus amongst the clinicians. Third, using feature importance can facilitate the regulatory approval process, and this avoids hampering the adoption of ML until universal rules are found and agreed by experts. However, we acknowledge that rule based methods are useful for knowledge discovery. Feature relevance explanation also has its limitations in explaining image-based datasets where the highlighted regions in a saliency map might not correspond well to high level concepts that are meaningful to humans. In this case, concept-based approaches, e.g., TCAV [24], that can quantify the degree to which a user-defined concept is important to a prediction, would be more appealing. In healthcare, such concepts would be pre-defined by clinicians.

Example-based explanation, especially *counterfactual examples*, can help to assess model robustness. Counterfactuals are generated by minimising the distance from the original input whilst producing a different prediction. As we mentioned in Section III, the distance metrics for adversarial examples are still valid for counterfactuals, although among them L_1 is the most widely explored in the literature and L_∞ is rarely used, which is unsurprising as L_1 can enforce sparsity in the generated example. As pointed out in [42], sparsity is one of the desirable properties of a good counterfactual. Intuitively, a counterfactual will be more implementable, and easier to understand, if fewer features have to be changed. Additional desirable properties for good counterfactuals are proximity (as close as possible to the original instance), plausibility (it is possible for the features to take that value) and diversity (multiple ways of achieving the desired prediction). On-going research seeks novel loss functions to incorporate

these properties to generate the counterfactuals. When using counterfactuals to assess model robustness, proximity is the most important property, in other words if the counterfactual methods don't satisfy this property, they should not be used to assess robustness. Intuitively, the greater the distance from an initial input to a counterfactual, the more robust the ML model is, i.e., the model is "harder to fool". Therefore, distance metrics can be used to define a robustness score for the ML model, which links to the robustness dimension of the spider diagram. For example, [43] used the L_1 distance metric to define the score. However, we suggest that the L_2 distance metric might be most appropriate to use based on the findings in [35] which indicates that achieving robustness against L_2 also achieves robustness against other distance metrics.

The use of XAI methods in support of ML model V&V will contribute evidence to the safety case, complementing other activities including performance assessment and safety-critical software engineering. It should be noted that explanations should be re-generated when the ML models are updated so that they reflect the state of the models.

5) OPERATION

As discussed in Section II, assurance should be considered to be a "through life" activity. This would include, for example, a clinician seeking assurance about a particular prediction, especially if acting on it can have a profound impact on patient safety. XAI methods can play a role here. Local feature relevance explanation may be helpful but counterfactuals also have a role, e.g., helping a clinician to decide whether or not a proposed change in treatment is likely to bring about the desired effect for a particular patient. Further, prototypes might also be able to help clinicians to make informed decisions for specific patients. However, current research, e.g., [29], often presents prototypes to laymen rather than domain experts to assess whether this will help them to make decisions, so more work needs to be done to see whether clinicians can benefit from prototype explanations. The role and significance of explainability in operation is examined in more detail in [28].

IV. CASE STUDY

This section presents a concrete healthcare case study to illustrate the role of XAI methods, introduced in Section III. The case study doesn't cover data management, but see our previous work for an illustration of the rationale for data inclusion for this case study [44]. The case study focuses on use of mechanical ventilation in Intensive Care Units (ICUs). Provision of mechanical ventilation is complex and consumes a significant proportion of ICU resources [45]. Invasive mechanical ventilation is used when patients cannot breathe unaided, and requires the insertion of a tube into the trachea of the patient. The term intubation is used for insertion of tube and extubation for removal of the tube. It is of critical importance to determine the right time to wean the patient from mechanical ventilation. Both early and late

TABLE 2. Performance of ML models

Model	AUC-ROC (95%CI)
Convolutional Neural Network	0.923 ± 0.010
Artificial Neural Network	0.784 ± 0.031
Logistic Regression	0.827 ± 0.000
Random Forest	0.748 ± 0.040
Decision Tree	0.808 ± 0.008
Support Vector Machine	0.826 ± 0.000

weaning are problematic. Early extubation can lead to the need for re-intubation, which may become urgent. Late extubation exposes a patient to discomfort and continued risk of complications such as pneumonia from prolonged intubation.

The case study is particularly concerned with predicting patient readiness for extubation so as to avoid the negative side effects of mis-timed extubation using the features shown in Figure 5. Put simply, the safety requirement is "prediction of readiness for extubation is timely". The case study is based on the MIMIC-III dataset [46] and used a convolutional NN (CNN) to predict readiness for extubation in the next hour.

A. ML ALGORITHM SELECTION

ML Algorithm selection is strongly influenced by performance, as previously indicated. There are a range of performance metrics, e.g., false positives, which in this case study would mean indicating that a patient is ready for extubation when it was actually premature. Here we use the AUC-ROC performance measure. The ROC curve plots the true positives against the false positives at various threshold settings. AUC-ROC represents the degree to which the model is capable of distinguishing between classes. For a "random" model the AUC-ROC would be 0.5 and for a "perfect" model it would be 1.

For the case study, the performance of a number of ML models, including CNNs, were evaluated on the same dataset to support this phase, see Table 2. CNNs have the best performance and more importantly, achieve better performance than decision trees and logistic regression which, as noted above, are often viewed as intrinsically interpretable. As mentioned in Section III there is a trade-off between performance and explainability. If performance over-rides the need for explainablility, then CNN should be chosen. Whilst if intrinsic interpretability is more important, then logistic regression should be chosen. In this case study, CNNs have been chosen, and *post-hoc* XAI methods are used to explain the model, see the rest of the section for details.

B. MODEL LEARNING

As we indicated in Section III, two XAI methods can be helpful at this stage: adversarial examples and influential instances. Because adversarial examples are difficult to generate for tabular data, here we focus on the use of influential instances for "model debugging". This shows how they provide assurance about the appropriateness of the ML model learning process, in the context of the safety requirement.
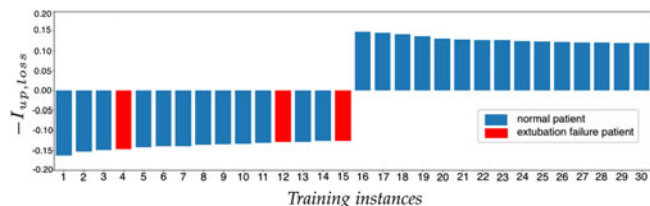
FIGURE 3. Top 30 most influential training instances.

FIGURE 4. Distribution of influential instances.

When preparing the dataset for the case study, one issue that came up was whether or not to include the extubation failure patients. Here extubation failure is defined as the need for re-intubation within 48 hours. The causes of extubation failure are complex and unclear, but some of the literature suggests that premature extubation could cause extubation failure [45]. Therefore, including extubation failure patients in the training dataset might not be optimal, as it might negatively influence the prediction. To explore this issue further, we trained two CNN models to predict the readiness for extubation in the next hour in order to observe the effect of extubation failure patients. In the first model, we excluded all of the extubation failure patients in the training dataset. In the second model, we included all of the extubation failure patents in the training dataset. The accuracy of the second model is slightly changed by comparison with the first model. We randomly picked one of the test instances that was "interesting" in that the two models produced different predictions. For this instance, the first model predicted the patient should continue to be intubated, which is also the true label. However, the second model predicted that the patient was ready for extubation in the next hour. We used influence functions to identify the influential training instances for this test instance.

The key idea behind influence functions is to up-weight the loss of a training instance by an infinitestimally small step ϵ, which results in new model parameters, $\hat{\theta}_{\epsilon,z} = argmin(1 - \epsilon)\frac{1}{n}\sum_{i=1}^{n} L(z_i, \theta) + \epsilon L(z, \theta)$, where θ is the model parameter vector and $\hat{\theta}_{\epsilon,z}$ is the model parameter after upweighting z by ϵ. L is the loss function used for training the model. The influence of upweighting z on the parameters $\hat{\theta}$ given by Cook and Weisberg [47] is as follows:

$$I_{up,params}(z) = \frac{d\hat{\theta}_{\epsilon,z}}{d\epsilon}\Big|_{\epsilon=0} = -H_{\hat{\theta}}^{-1}\nabla_\theta L(z, \hat{\theta}). \quad (1)$$

Where $H_{\hat{\theta}}$ is the Hessian matrix and $\nabla_\theta L(z, \hat{\theta})$ is the loss gradient with respect to the parameters $\hat{\theta}$ for the training instance z. Next, we can apply the chain rule to calculate the influence of upweighting instance z on the loss of a test instance z_{test}

$$I_{up,loss}(z, z_{test}) = \frac{dL(z_{test}, \hat{\theta}_{\epsilon,z})}{d\epsilon}\Big|_{\epsilon=0}$$
$$= \nabla_\theta L(z_{test}, \hat{\theta})^T \frac{d\hat{\theta}_{\epsilon,z}}{d\epsilon}\Big|_{\epsilon=0}$$
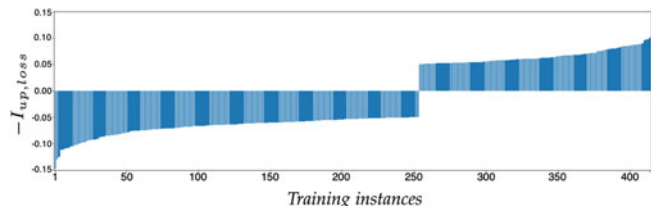$$= -\nabla_\theta L(z_{test}, \hat{\theta})^T H_{\hat{\theta}}^{-1} \nabla_\theta L(z, \hat{\theta}). \quad (2)$$

In this work, we use the influence functions algorithm developed by Koh and Liang [36] to calculate $-I_{up,loss}(z_i, z_{test})$ for each training instance z_i for this test instance. Figure 3 shows the top 15 helpful training instances (most positive $-I_{up,loss}(z_i, z_{test})$) and the top 15 harmful training instances (most negative $-I_{up,loss}(z_i, z_{test})$) for this test instance. From the figure, it shows there are three instances of patients who had extubation failure among the harmful training instances, which indicates that including the extubation failure patients made the predictions for the test instance worse. Figure 4 shows some of the most influential data points (magnitude of $-I_{up,loss}(z_i, z_{test})$ is large) from the extubation failure patients and that more of them have a negative influence than a positive influence. This suggests that the inclusion of extubation failure could make the prediction ready to extubate when it is not the case. Thus, we decided to exclude the extubation failure patients from the training dataset and the first CNN model was taken forward to the V&V stage. In a more general situation when prior knowledge is not available, i.e., we don't know what subset of the data could be problematic, we can still choose a test instance where the prediction is wrong and identify the influence of the training instances on this prediction. Then, further investigation could be done to understand what input features strongly impact the influence score, e.g., by perturbation [36] or by using decision trees.

C. MODEL VERIFICATION AND VALIDATION
In this section, we focus on clinical validation, as set out in Section III, and illustrate the use of XAI methods for demonstrating clinical association and robustness. We do not consider analytical validation here.

1) FEATURE RELEVANCE EXPLANATIONS
Here we illustrate the role of feature relevance in satisfying the clinical association safety assurance objective. This is done using DeepLIFT [22] which is a model-specific XAI method for deep NNs. It compares the activation of each neuron to its "reference activation" and attributes to each input feature an importance score based on the difference. The "reference activation" is obtained through some user-defined reference input and in this case, the reference sample is the minimum values of all of the input features obtained from the data set. We chose this method for two main reasons. First, it deals effectively with discontinuities in the gradient of the CNN model as it uses a difference from reference approach. Second, it avoids the problem of model saturation where using gradients would just assign zero to the features [22].
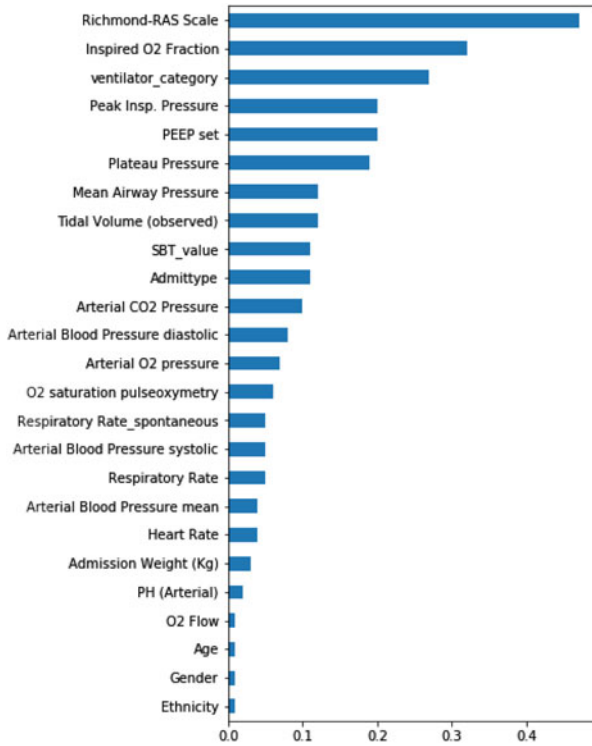
FIGURE 5. Feature importance for the CNN model.

An overview of the results of using DeepLIFT is shown in Figure 5; these values are averaged over the whole dataset, so this can be viewed as global feature importance. The feature ranking correlates well with clinical expectations, helping to give confidence in the model. Those features that score near zero in Figure 5, e.g., ethnicity, gender and age, have little influence on the weaning decision, which is as expected. The top five features also align with clinical evidence. Patients who are undergoing invasive mechanical ventilation are often sedated to maintain physiological stability and to control pain levels. Sedation is reflected in the Richardson Agitation Scale (RAS) with negative values representing sedation and 0 meaning that they are alert and calm, thus more likely to be suitable for extubation. This is consistent with the first entry in the weaning checklist used in [48] that patients are "cooperative and pain free". The second most important feature is "Inspired O2 fraction" which is the third checklist entry in [48]. The third most important feature is "ventilator category," which is the mode used for ventilation and is under direct clinician control; some modes are unsuitable for spontaneous breathing so cannot easily support weaning. The fourth and fifth most important features, peak inspiratory pressure and positive end-expiratory pressure (PEEP) set are airway pressures representing how hard the ventilator is having to work; PEEP is also the third entry in the weaning checklist in [48].

Here we have demonstrated valid clinical association through clinical evidence (relevant literature support) and expert opinion (consultation with clinicians). Overall, the

TABLE 3. Counterfactual examples for a given instance

Features	Original instance	Counterfactual Examples		
		1	2	3
Admit Type	Emergency	—	—	—
Ethnicity	White	—	—	—
Gender	Female	—	—	—
Age	78.2	—	—	—
Admission Weight	86.5	—	—	—
Heart Rate	119	—	110	—
Respiratory Rate	24	26	—	—
SpO2	98	—	—	96
Inspired O2 Fraction	100%	—	40%	—
PEEP set	10	5	5	5
Mean Airway Pressure	14	—	10	—
Tidal Volume (observed)	541	—	—	560
PH (Arterial)	7.46	—	—	—
Respiratory Rate(Spont)	0	—	24	—
Richmond-RAS Scale	-1	—	0	—
Peak Insp. Pressure	21	—	—	—
O2 Flow	5	—	—	—
Plateau Pressure	19	—	—	—
Arterial O2 pressure	124	108	118	—
Arterial CO2 Pressure	33	—	—	—
Blood Pressure (systolic)	101	—	—	—
Blood Pressure (diastolic)	65	—	—	—
Blood Pressure (mean)	76	—	—	—
Spontaneous breathing trials	No result	Successfully Completed	Successfully Completed	Successfully Completed
Ventilator Mode	CMV/ASSIST/ AutoFlow	PCV+	SIMV/PSV	SIMV/PSV
Predicted outcome	0.93	0.44	0.17	0.36

benefit of the feature importance results is that they enable clinical judgement to be applied despite the opacity of the CNN model which contributes to safety assurance.

2) COUNTERFACTUAL EXPLANATIONS

One of the concerns in model V&V is robustness and here we show how to use counterfactuals to demonstrate robustness. Diverse Counterfactual Examples (DiCE) [42] is used to generate the counterfactual examples. The reason we use DiCE is that it is one of the few methods to satisfy all of the four properties of a good counterfactual, i.e., sparsity, proximity, plausibility and diversity, introduced in Section III. Table 3 shows a set of counterfactual examples for a particular patient identifying which features need to change in order to "flip" the prediction from continued intubation to extubation using DiCE. The left hand column shows the 25 features used by the model and the prediction of the ML model is included in the bottom row. The original instance is shown first, with the three rightmost columns showing counterfactual examples. Certain features cannot be varied, e.g., age and gender, in order to satisfy the plausibility criterion; the dashes in the rightmost three columns indicate no change from the original input. The change in prediction is shown in the bottom row where a value > 0.5 indicates that mechanical ventilation should continue.

In this case, as shown in Table 3, the minimum number of features that have to change to "flip" the prediction is five, showing robustness for this instance. However, one instance is not sufficient to show ML model robustness. More of the input instances in the dataset need to be investigated in order to generate a robustness score as defined in [43].

D. OPERATIONAL USE OF THE ML MODEL

The operation of ML models is often uncertain. Thus there is merit in extending the notion of assurance to operation, providing support to a clinician to give confidence to act on the
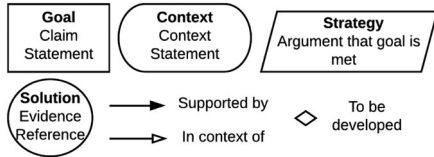
FIGURE 6. Goal structuring notation.

particular model prediction. One way of approaching this is to use local explanations. For example, we can generate the feature importance for a specific patient, similarly to Figure 5.

However, clinicians might want to find out when the patient would be ready to extubate. This brings us back to counterfactuals. The counterfactual examples shown in Table 3 could potentially help the clinician to identify actions to take so that the patient becomes ready to extubate. The model does not directly recommend a course of action; the counterfactual examples act to draw clinicians' attention to pertinent information so that they can formulate a plan from their own knowledge and experience. Note our model has not been used in operation yet, so we have just illustrated the possibilities.

E. SAFETY ARGUMENTS

As explained in Section II, it is common practice to present the arguments and evidence that provide assurance that a system is acceptably safe to deploy in a safety case. In this case study, the safety argument is presented using GSN. Before we describe the safety argument we have developed, we briefly introduce the notation.

A legend showing the key elements of GSN is presented in Figure 6; a detailed description of the notation can be found in [10]. The *goals* – claims that we wish to make and support – are shown as rectangles and they can be decomposed into sub-goals, thus forming a tree. *Goals* are understood in a *context*, e.g., the operating environment for the system or the safety requirements. Where the decomposition of goals is not obvious this is explained through a *strategy*, represented as a rhombus. The leaf-level goals are supported by *solutions*, represented as circles; the solutions provide references to evidence that supports the argument. Incomplete parts of the argument are shown with a diamond, meaning that part of the argument is to be developed.

Figure 7 presents a partial safety argument for the weaning case study, highlighting the role of explainability, e.g.,
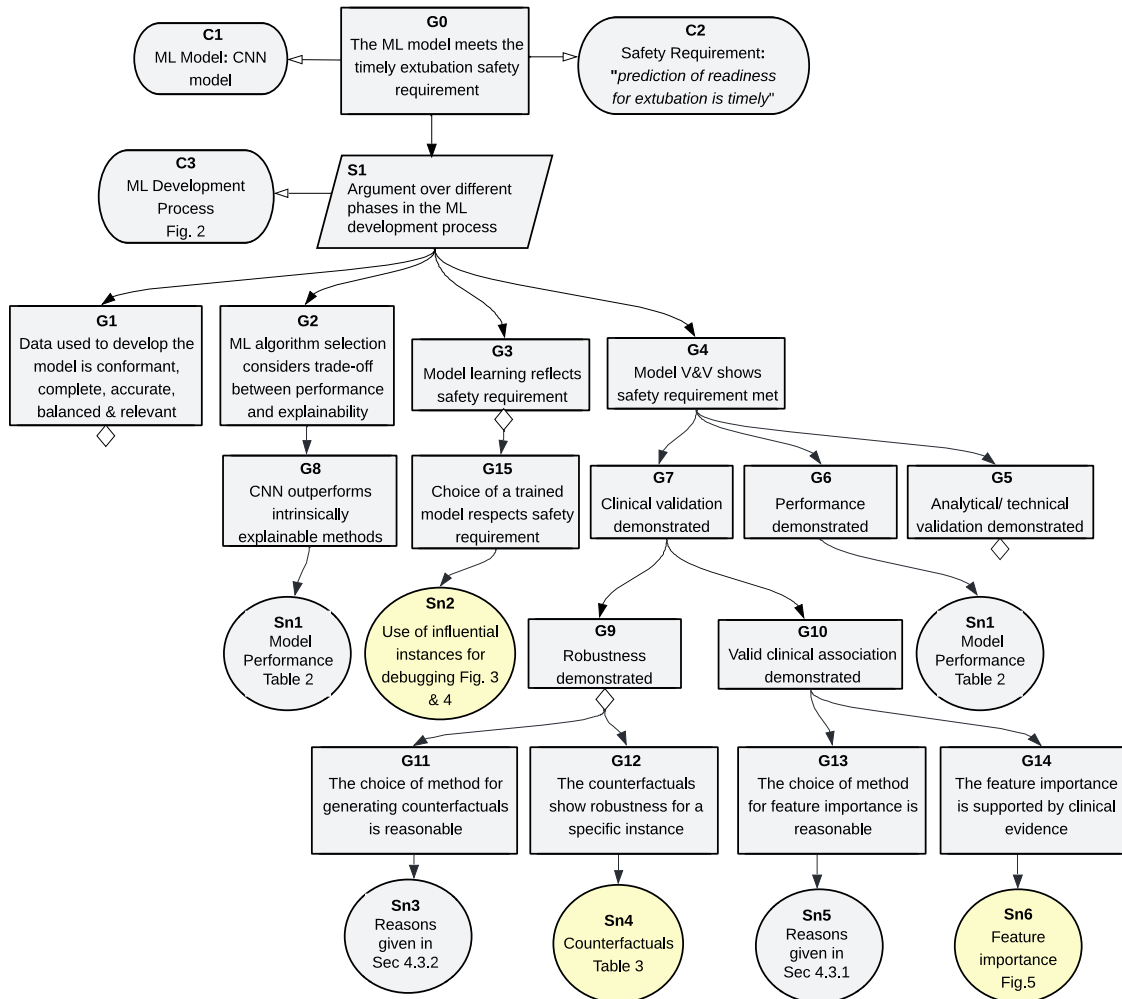

FIGURE 7. Partial safety argument for weaning ML model emphasising explainability.

solutions Sn2, Sn4 and Sn6 reflect XAI methods. The top goal (G0), which states that the ML model meets its safety requirement, is set out in the context of the definition of the ML model and the associated safety requirement – that "prediction of readiness for extubation is timely".

The top-level argument strategy is decomposition across the stages of the ML development process. As the paper does not consider data management and analytical/technical validation in detail, the corresponding goal (G1) and goal (G5) are left undeveloped.

G2: ML algorithm selection considers trade-off between performance and explainability – this is supported by the analysis in Section IV (Sn1) which shows that the CNN outperforms other available ML methods, and suitable *post-hoc* XAI methods are available.

G3: Model learning reflects safety requirement – this is *partially* supported by G15 which in turn is supported by the use of influential instances (Sn2) which show the rationale for excluding extubation failure patients in the trained model. Note that other evidence is needed (so G3 is shown as needing development), e.g., to show appropriateness of parameter selection for model training.

G4: Model V&V shows safety requirements met – this is broken down into G5: analytical/technical validation, G6: performance demonstrated and G7: clinical validation which is decomposed into two sub-goals covering the V&V criteria introduced in Section III.

G6: Performance demonstrated – this is directly supported by the AUC-ROC in Table 2 which shows the superiority of the CNN performance to others.

G9: Robustness demonstrated, which is decomposed into two subgoals, G11 concerning the reason for the choice of the specific XAI method and G12 concerning the results of using the method. It is important to document the rationale for choosing the method, which will benefit from a more systematic evaluation metric for XAI methods, see Section V for details. G9 needs further development as G12 only shows robustness for a single instance. More input instances need to be investigated to understand the model robustness.

G10: Valid clinical association demonstrated, which is decomposed into two subgoals, G13 concerning the reason for the choice of the specific XAI method and G14 concerning the results of using the method. G14 is further supported by clinical evidence in the literature and expert review by clinicians. Meeting G10, demonstrating valid clinical association, does not entail trust. Trust is a more complicated topic [33] and is outside the scope of this paper.

The evidence presented above should not be taken as sufficient to justify deployment of the CNN model described here in a clinical context. For example, clinical trials will often be needed to obtain further clinical safety evidence and the necessary regulatory approval. However, the explainability argument and the supporting evidence presented in this section is a valuable part of the overall safety case.

TABLE 4. Role of XAI methods in different phases

Phases	Activity	XAI methods	Stakeholders
Development	Data Management	Prototypes	ML developers Regulators Hospital managers
	ML Algorithm Selection	Trade-off performance & explainability	ML developers
	Model Learning	Adversarial examples Influential instances	ML developers
	Model V & V	Global feature importance Counterfactual explanations	ML developers Regulators Hospital managers Insurers
Operation	Decision Support	Local feature importance Counterfactual explanations	Expert users: clinicians Decision recipients: patients

V. DISCUSSION

Safety assurance of ML models in healthcare is an active area of research. Although explainability is often said to help in safety assurance of ML, few studies so far have explored the possibilities systematically and identified precisely how explainability can help safety assurance. This paper seeks to fill this gap. We first presented the spider diagram in Figure 1 which conceptualises how explainability relates to safety. Through the analysis in Section III and the case study in Section IV we illustrated how explainability can help in safety assurance in the context of the ML life-cycle, as summarised in Table 4 along with the interested stakeholders. Although using explainability can help safety assurance, there are also associated challenges.

First is the difficulty of evaluating XAI methods. A considerable number of evaluation metrics have been suggested for assessing the quality of XAI methods. For explanation by approximation, fidelity is often proposed as the evaluation metric. For example, in [13], [38], fidelity is used to measure how accurately the approximate model matches the task model locally. For feature relevance explanation, their ideal properties, e.g., implementation invariance (the feature importance are always identical for two functionally equivalent networks) and sensitivity (if a feature changes and a prediction changes, then this feature should not have zero attribution), have been defined using axiomatic evaluation methods [20]. For explanation by examples, humans are often involved in assessing whether the explanation is useful or not. For example, in [29] prototypes were presented to users and then human accuracy, i.e.the proportion of the human predictions that correctly match the model's prediction, was measured. Among these proposed evaluation metrics, some are subjective and some are objective. One of the useful taxonomies was proposed by Doshi-Velez and Kim [49] for evaluating XAI: application-grounded, human-grounded, and functionality-grounded. However, there is still a lack of agreed formal evaluation metrics enabling a more systematic evaluation of methods. For example, [50] found Gradients & GradCAM passed their sanity checks based on their model parameter randomization test and the data randomization test, while in [51] they found GradCAM is one

of the most interpretable and reliable XAI methods but gradient didn't stand out based on their evaluation metrics. This highlights the importance of ongoing research developing systematic evaluation metrics which will allow a formal and fair comparison of available XAI methods. This should help to guide the selection of an appropriate XAI method in a specific situation as many different XAI methods can provide similar explanations, e.g., feature relevance.

Second is the limitation of explainability itself. Even if the evaluation metrics are improved, there are some intrinsic limitations of the current XAI methods. As pointed out by Rudin [27] *post-hoc* explanations "must be wrong" as if they were completely faithful to the task model, then we only need the explanation model. Further, as illustrated in [52], current XAI methods only provide descriptive accounts of the task model rather than normative evaluation to justify the model behaviour. Whilst this is valid, it is unrealistic to expect XAI methods to "close the loop" by themselves; instead this is the role of clinical judgement, as discussed above. Put more positively, explanation "serves as the unacknowledged bridge" between the task model and normative evaluation [52]. What these examples make clear is that, although the use of XAI methods can contribute to safety assurance, it is not enough to assure safety by itself, as recognised by others [33], [52].

Next, we will identify some relevant complementary methods that also contribute to safety assurance. First, it is important to adapt established methods from safety-critical software engineering for AI/ML-based SaMD. One such method is static analysis, analysing the code without executing it, to look for "bugs" (see [39] for an illustration). It is also standard practice to measure test coverage of the software, e.g., ensuring that all branches in the code have been executed at least once, when undertaking V&V. The obvious analogy for NNs is neuron coverage, although there is some debate about whether or not this is an appropriate criterion [53]. Nonetheless, coverage is significant when considering safety, as assurance is clearly undermined if there are significant parts of the ML model for which we have no test evidence. Consequently, it seems likely that understanding of what are appropriate coverage criteria will improve as experience of using AI/ML-based SaMD increases. Second, there are assurance methods that address the specific challenges of V&V for AI/ML-based software. It is possible to apply formal methods (mathematical techniques of verification) to ML models including for assessing properties such as robustness. A recent survey of such approaches can be found in [54]; it covers static analysis as well as methods such as Satisfiability Module Theories (SMT) and identifies some tools that are capable of scaling to very large models, e.g., NNs with millions of neurons.

Safety is a cross-cutting concern which interacts with technology, ethics, trust, etc. Earlier we introduced the spider diagram in Figure 1 as a heuristic representation of the engineering aspects of safety; we now briefly consider how to make it more concrete, i.e., how the area inside the hexagon *might* be estimated. There are *candidate* quantitative measures for some of the dimensions, e.g., AUC-ROC (see Table 2) for performance, and distance metrics to define robustness scores [43]. There is a conceptual basis for some of the other dimensions, e.g., fidelity and interpretability for explainability [33], and conformance with the five criteria for data management which also links to the safety argument. However, as we mentioned earlier more work is needed to develop systematic evaluation metrics for XAI methods. It is hard to measure the other safety specific controls dimension but the structured representation of controls in [9] might give a starting point. There is ongoing work on utility of explanations and what form of explanations are preferred by clinicians [55], but more empirical work is needed. At this point, the spider diagram remains a heuristic model, but we believe it is possible to make it a more "formal" tool for evaluating (comparing) alternative ways of developing SaMD, from a safety assurance perspective.

VI. CONCLUSION

To our knowledge, this is the first systematic attempt to explore the role of explainability in assuring safety of ML, with a focus on pre-deployment decision-making. We believe this will be of interest to regulators, as it illustrates how to use XAI methods to provide evidence to support relevant safety objectives, e.g., for clinical association, articulated by the FDA and IMDRF.

We first presented a spider diagram to the general relationship between explainability and safety. Then we extrapolated the safety objectives at the different phases of the ML development process and illustrated how XAI methods can help at each phase. Finally, we used a concrete healthcare case study to demonstrate how XAI methods can help to meet these safety objectives, particularly in model learning and model V&V. Specifically, we have shown the value of influential instances for model debugging during model learning, which is of particular interest to ML developers. Further, we have shown the value of feature relevance and counterfactuals in model V&V, which is of particular interest to ML developer, regulators and others involved in deployment decisions, see Table 4. The case study also shows how the use of these XAI methods feeds into a safety case, e.g., as required by healthcare standards [8].

We suggest future work should place more effort on developing and applying systematic evaluation metrics for XAI methods, which in turn will guide others in selecting appropriate XAI methods. Further, there is a need for more empirical studies to evaluate how XAI methods can best assist the end-users. This might usefully be combined with a deeper exploration of trust.

The code for applying various XAI methods is available at: https://github.com/ Yanjiayork/mechanical_ventilator.

REFERENCES

[1] US Food and Drug Administration, "Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SAMD)-discussion paper and request for feedback," 2019. Accessed: May 06, 2022. [Online]. Available: https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf

[2] R. Hawkins, C. Paterson, C. Picardi, Y. Jia, R. Calinescu, and I. Habli, "Guidance on the assurance of machine learning in autonomous systems (AMLAS)," 2021, *arXiv:2102.01564*.

[3] D. S. Watson *et al.*, "Clinical applications of machine learning algorithms: Beyond the black box," *BMJ*, vol. 364, 2019, Art. no. l886.

[4] R. Challen, J. Denny, M. Pitt, L. Gompels, T. Edwards, and K. Tsaneva-Atanasova, "Artificial intelligence, bias and clinical safety," *BMJ Qual. Saf.*, vol. 28, no. 3, pp. 231–237, 2019.

[5] M. Diamond, "Transparency artificial intelligence/ machine learning (AI/ML)-Enabled medical devices: FDA virtual public workshop," 2021. Accessed: May 06, 2022. [Online]. Available: https://www.fda.gov/media/153306/download

[6] 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 *Proc. IEEE 5th Int. Conf. Data Sci. Adv. Analytics*, 2018, pp. 80–89.

[7] IMDRF SaMD Working Group, "Software as a medical Device(SaMD): Clinical evaluation - guidance for industry and food and drug administration staff," 2017. Accessed: May 06, 2022. [Online]. Available: https://www.fda.gov/files/medical%20devices/published/Software-as-a-Medical-Device-%28SAMD%29–Clinical-Evaluation—Guidance-for-Industry-and-Food-and-Drug-Administration-Staff.pdf

[8] NHS Digital, "DCB0160: Clinical risk management: Its application in the deployment and use of health IT systems," 2018. Accessed: May 06, 2022. [Online]. Available: https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections/dcb0160-clinical-risk-management-its-application-in-the-deployment-and-use-of-health-it-systems

[9] Y. Jia, T. Lawton, J. Burden, J. McDermid, and I. Habli, "Safety-driven design of machine learning for sepsis treatment," *J. Biomed. Informat.*, vol. 117, 2021, Art. no. 103762.

[10] Assurance Case Working Group [ACWG], "Goal structuring notation community standard version 2," 2018. Accessed: Apr. 19, 2021. [Online]. Available: https://scsc.uk/r141B:1?t=1

[11] A. B. Arrieta *et al.*, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020.

[12] C. Mood, "Logistic regression: Why we cannot do what we think we can do, and what we can do about it," *Eur. Sociol. Rev.*, vol. 26, no. 1, pp. 67–82, 2010.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should i trust you?" Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 1135–1144.

[14] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 1527–1535.

[15] O. Bastani, C. Kim, and H. Bastani, "Interpretability via model extraction," 2017, *arXiv:1706.09773*.

[16] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4765–4774.

[17] P. Cortez and M. J. Embrechts, "Using sensitivity analysis and visualization techniques to open black box data mining models," *Inf. Sci.*, vol. 225, pp. 1–17, 2013.

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

[19] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Proc. Workshop Int. Conf. Learn. Representations*, 2014.

[20] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 3319–3328.

[21] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS One*, vol. 10, no. 7, 2015, Art. no. e0130140.

[22] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3145–3153.

[23] C. Olah, A. Mordvintsev, and L. Schubert, "Feature visualization," *Distill*, vol. 2, no. 11, 2017, Art. no. e7.

[24] B. Kim *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2668–2677.

[25] S. Jain and B. C. Wallace, "Attention is not explanation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Human Lang. Technol.*, 2019, pp. 3543–3556.

[26] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," in *Proc. Conf. Empir. Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 11–20.

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

[28] J. McDermid, Y. Jia, Z. Porter, and I. Habli, "AI explainability: The technical and ethical dimensions," *Philos. Trans. Roy. Soc. A*, vol. 379, 2021, Art. no. 20200363.

[29] S. Tan, M. Soloviev, G. Hooker, and M. T. Wells, "Tree space prototypes: Another look at making tree ensembles interpretable," in *Proc. ACM-IMS Found. Data Sci. Conf.*, 2020, pp. 23–34.

[30] The Royal Society, "Explainable AI: The basics POLICY BRIEFING," 2019. [Online]. Available: https://ec.europa.eu/futurium/en/system/files/ged/ai-and-interpretability-policy-briefing_creative_commons.pdf

[31] M. G. Kahn *et al.*, "A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data," *EGEMS*, vol. 4, no. 1, 2016, Art. no. 1244.

[32] K. S. Gurumoorthy, A. Dhurandhar, G. Cecchi, and C. Aggarwal, "Efficient data representation by selecting prototypes with importance weights," in *Proc. IEEE Int. Conf. Data Mining*, 2019, pp. 260–269.

[33] A. F. Markus, J. A. Kors, and P. R. Rijnbeek, "The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies," *J. Biomed. Informat.*, vol. 113, 2020, Art. no. 103655.

[34] E. W. Ayers, F. Eiras, M. Hawasly, and I. Whiteside, "PaRoT: A practical framework for robust deep neural network training," in *Proc. NASA Formal Methods Symp.*, 2020, pp. 63–84.

[35] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. 2017 IEEE Symp. Secur. Privacy*, 2017, pp. 39–57.

[36] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1885–1894.

[37] A. S. Ross, M. C. Hughes, and F. Doshi-Velez, "Right for the right reasons: Training differentiable models by constraining their explanations," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 2662–2670.

[38] G. Plumb, M. Al-Shedivat, A. A. Cabrera, A. Perer, E. Xing, and A. Talwalkar, "Regularizing black-box models for improved interpretability," in *Proc. Adv. Neural Inform. Process. Syst.*, 2020, pp. 10526–10536.

[39] J. McDermid and Y. Jia, "Safety of artificial intelligence: A collaborative model," in *Proc. AISafety*, 2020.

[40] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What clinicians want: Contextualizing explainable machine learning for clinical end use," in *Proc. Mach. Learn. Healthcare Conf.*, 2019, pp. 359–380.

[41] B. Kovalerchuk, E. Vityaev, and J. F. Ruiz, "Consistent and complete data and "expert" mining in medicine," *Stud. Fuzziness Soft Comput.*, vol. 60, pp. 238–281, 2001.

[42] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proc. Conf. Fairness, Accountability, Transparency*, 2020, pp. 607–617.

[43] S. Sharma, J. Henderson, and J. Ghosh, "CERTIFAI: A common framework to provide explanations and analyse the fairness and robustness of black-box models," in *Proc. AAAI/ACM Conf. AI Ethics Soc.*, 2020, pp. 166–172.

[44] Y. Jia, C. Kaul, T. Lawton, R. Murray-Smith, and I. Habli, "Prediction of weaning from mechanical ventilation using convolutional neural networks," *Artif. Intell. Med.*, vol. 117, 2021, Art. no. 102087.

[45] N. Ambrosino and L. Gabbrielli, "The difficult-to-wean patient," *Expert Rev. Respir. Med.*, vol. 4, no. 5, pp. 685–692, 2010.

[46] A. E. Johnson *et al.*, "MIMIC-III, A freely accessible critical care database," *Sci. Data*, vol. 3, no. 1, pp. 1–9, 2016.

[47] R. D. Cook and S. Weisberg, *Residuals and Influence in Regression*. New York, NY, USA: Chapman and Hall, 1982.

[48] T. Walsh, S. Dodds, and F. McArdle, "Evaluation of simple criteria to predict successful weaning from mechanical ventilation in intensive care patients," *Brit. J. Anaesth.*, vol. 92, no. 6, pp. 793–799, 2004.

[49] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, *arXiv:1702.08608*.

[50] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proc. 32nd Int. Conf. Neural Inform. Process. Syst.*, 2018, pp. 9525–9536.

[51] I. Kakogeorgiou and K. Karantzalos, "Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing," *Int. J. Appl. Earth Observation Geoinf.*, vol. 103, 2021, Art. no. 102520.

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

[53] Z. Li, X. Ma, C. Xu, and C. Cao, "Structural coverage criteria for neural networks could be misleading," in *Proc. IEEE/ACM 41st Int. Conf. Softw. Eng.: New Ideas Emerg. Results*, 2019, pp. 89–92.

[54] C. Urban and A. Miné, "A review of formal methods applied to machine learning," 2021, *arXiv:2104.02466*.

[55] O. Lahav, N. Mastronarde, and M. van der Schaar, "What is interpretable? Using machine learning to design interpretable decision-support systems," 2018, *arXiv:1811.10799*.

YAN JIA is currently a research fellow with University of York. Her work focuses on safety management in healthcare, unifying and integrating work in traditional safety engineering and in machine learning. Initial work focused on medication management for patients at risk of atrial fibrillation. More recently she has worked on the treatment of sepsis, using Reinforcement Learning (RL) in recommending treatments, showing how the use of established safety methods can inform the development of the RL elements of the system and support production of a safety case. Recent work has enhanced some of the explainability methods for ML and investigated their role in safety assurance for healthcare.

JOHN MCDERMID is currently a professor of software engineering with the University of York and has studied safety of complex systems for more than 30 years. He became director of the Lloyd's Register Foundation funded Assuring Autonomy International Programme in January 2018, focusing on safety of robotics and autonomous systems. He has acted as an advisor to many organizations, including FiveAI, Fraunhofer IESE in Kaiserslautern, the Nuclear Regulatory Commission, Siemens, and Rolls-Royce. He was appointed as a Non-Executive director of the HSE, in October 2019. He has been actively involved in standards development, including work on safety and software standards for civilian and defense applications. He currently chairs the BSI Connected Autonomous Vehicles Standards Advisory Board. He became a fellow of the Royal Academy of Engineering, in 2002 and was awarded an OBE, in 2010.

TOM LAWTON is currently a consultant in Critical Care and Anaesthesia and head of Clinical Artificial Intelligence with Bradford Teaching Hospitals NHS Foundation Trust, and director of Clinical Analytics with the Improvement Academy in the Bradford Institute for Health Research. He is a former computer programmer and current NHS-R fellow, and an affiliated member of the Yorkshire and Humber Patient Safety Translational research centre. As a visiting fellow of the Assuring Autonomy International Programme he is actively involved in the responsible introduction of AI into healthcare. He was awarded an MBE in 2020 for his work on COVID-19.

IBRAHIM HABLI is a reader in safety-critical systems with the University of York. He is the principal investigator on the UKRI-funded project AR-TAS (Assuring Responsibility for Trustworthy Autonomous Systems). He co-leads the research activities of the 12 M Lloyd's Register Foundation Assuring Autonomy International Programme (AAIP). He is also an investigator on the TAS Node in Resilience (REASON) and the Marie Curie H2020 doctoral training program Safer Autonomous Systems. In 2015, he was awarded a Royal Academy of Engineering Industrial Fellowship and collaborated with the NHS on evidence-based methods for assuring complex digital interventions. His research on the assurance of critical and autonomous systems has spanned multidisciplinary collaborations including with clinical medicine, law, economics, and ethics.