Review Article

Received: 2023/10/09    Revised: 2023/11/06    Accepted: 2023/11/17

DOI: https://doi.org/10.15441/ceem.23.145

Explainable AI in Emergency medicine: An overview

Running title: Explainable AI in Emergency Medicine

Authors:
Yohei Okada, MD, PhD
Email: yohei_ok@duke-nus.edu.sg / okadayohei1127@yahoo.co.jp
Affiliation:
Health Services and Systems Research, Duke-NUS Medical School, Singapore, Singapore
Preventive Services, Graduate School of Medicine, Kyoto University, Kyoto, Japan

Yilin Ning, PhD
Email: yilin.ning@duke-nus.edu.sg
Affiliation: Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore, Singapore

Marcus Eng Hock Ong, MBBS, MPH
Email: marcus.ong@duke-nus.edu.sg
Affiliation: Health Services and Systems Research, Duke-NUS Medical School, Singapore, Singapore

ORCID: 0000-0001-7874-7612
Corresponding author
Yohei Okada, MD, PhD.
Health Services and Systems Research, Duke-NUS Medical School, National University of Singapore, Singapore
Preventive Services, Graduate School of Medicine, Kyoto University, Kyoto, Japan
E-mail: yohei_ok@duke-nus.edu.sg / okadayohei1127@yahoo.co.jp
Abstract

Background and Aims

Artificial Intelligence (AI) and Machine Learning (ML) have potential to revolutionize emergency medical care by enhancing triage systems, improving diagnostic accuracy, refining prognostication, and optimizing various aspects of clinical care. However, as clinicians often lack AI expertise, they might perceive AI as a "black box", leading to trust issues. To address this, "Explainable AI", which makes AI functionalities comprehensible to end-users, is important. This review introduces the definitions, importance, and role of explainable AI, and potential challenges in emergency medicine.

Results and Discussion

First, we introduce the terms: explainability, interpretability, and transparency of AI models. These terms sound similar; however, they have different roles in discussing AI. Second, we indicate why explainable AI is required in clinical settings including four reasons: justification, control, improvement, and discovery, and their examples. Third, we describe three major categories of explainability, pre-modeling explainability, interpretable models, and post-modeling explainability. Especially for post-modeling explainability, we suggest some examples such as visualization, simplification, text justification, and feature relevance. Lastly, we show the challenges of implementing AI and ML models in clinical settings and highlight the importance of collaboration between clinicians, developers, and researchers.

Conclusion

This paper summarizes the concept of "explainable AI" for emergency medicine clinicians. This review may help clinicians understand explainable AI in emergency contexts.

Key words
Artificial Intelligence, Machine learning, Resuscitation, Emergency medicine
1. **Introduction**

Artificial Intelligence (AI) and machine Learning (ML) are powerful technologies that have the potential to improve medical care. (1) AI refers to the broader concept of technology being able to carry out tasks in an autonomous and smart way, encompassing a variety of technologies, while ML is a subset of AI focused on the idea that machines can learn from data, identify patterns, and make decisions with minimal human intervention. (1-4) Particularly in emergency medicine, AI and ML are expected to play critical roles in accelerating triage, diagnosis, and prognostication to optimize individual patient care through the input of clinical information and/or image recognition. (2, 4-8) Furthermore, streamlined clinical documentation or recording using natural language processing is expected to make these tasks more efficient. (9-11) Additionally, these technologies will also contribute to drug discovery, patient monitoring, resource allocation, and epidemiological surveillance. (12-15)

Despite expectations that emergency physicians will become general users of AI and ML in the near future, critics are doubtful whether they can trust and rely on AI and ML models. (16) Physicians are usually not experts in AI and may not have an in-depth understanding of AI or ML works. When an AI model outputs a medical classification or prediction, without necessarily ‘explaining’ the underlying process or showing the variables and weights driving the prediction, physicians who are not familiar with AI algorithms may perceive the AI model as a “black box”. Such a situation may lead to doubt and mistrust in the AI's output, and it is one of the big challenges for the implementation of AI and ML tools in clinical settings. (17)

To address these concerns, the concept of “Explainable AI” has been highlighted as a possible solution for the successful implementation of AI and ML in medical practice. (18-20) Explainable AI is a concept within the field of AI that aims to make the functioning of AI systems more understandable for end-users, researchers, or developers. (18-20) As more accurate models are
developed; it may be more complicated to understand how they work. This review introduces the definitions, importance, and role of explainable AI, and potential challenges in emergency medicine.

2. **What is transparency, interpretability, and explainability in AI?**

There are many types of definitions for explainable AI or related terms; however, before starting to discuss the concept of “explainability”, we would like to introduce the terms “interpretability” and “transparency”. (18-20) While they have distinct meanings, they are sometimes mistaken for explainability. AI models are often labeled as a "black box," suggesting a lack of transparency, (Figure 1). (21) In this opaque model, the process of converting input to output is invisible, making it challenging for users to understand how it is processed. In contrast, transparent/interpretable models allow users to understand how inputs are processed to produce outputs so that we can observe how the models are working internally, (Figure 1). For instance, straightforward models like linear regression with a limited number of variables or a decision tree with a few branches are relatively easier to understand (Figure 2). (21)

Explainability is different conceptually. Arrieta et al. suggested that “Given a certain audience, explainability refers to the ability of a model to show details and make its internal functioning clear or easy to understand”. (19) This definition emphasizes the audience's perspective. Different audiences, with varied backgrounds, experiences, and capacities may have different expectations of explainability in machine learning (ML). While explainability is inherent in transparent models, it is also tied to post-hoc explainability (Figure 1). This concept refers to techniques that provide the rationale or explanation to support the users’ understanding of how the model works, even if the model itself is non-interpretable (often referred to as a "black box"), by providing text or visual explanation, etc. (22)
3. Why is explainable AI needed in emergency medicine?

There are four main reasons why explainable AI is required in clinical settings: justification, control, improvement, and discovery of novel ideas. (23) These four reasons may appear to overlap, but from a clinical perspective, they capture different motivations for explainability.

Firstly, explainable AI is useful for the justification of AI model outputs, to enhance trust, and to support clinical decision-making. (23) Generally, in the process of clinical management, clinicians need to explain the medical condition, treatment plan, and expected outcomes to patients and their families. Even if the patients or their families cannot fully understand the details of their management due to medical complexity, clinicians need to make the effort to communicate to the patients and their families, to facilitate shared decision-making and trust. (24) Thus, it is obvious that clinicians require explainability in their decision-making process, and the results generated from clinical AI or ML need to build trust. Unless there is a clear and satisfactory explanation that makes sense, clinicians will likely be hesitant to trust AI. For example, a recent randomized control trial investigated the efficacy of an AI model that aims to detect cardiac arrest cases at the dispatch center, using voice data analysis during the emergency call. (25, 26) This study was unable to demonstrate the effectiveness of the AI model, with the researchers suggesting that some dispatchers might not have trusted the output from the AI model due to the absence of a reasonable explanation. Explainable AI can ensure that there is an auditable and provable way to defend algorithmic decisions as being fair and ethical. (27, 28)

Secondly, explainability helps users to maintain control of a complex technology. Indeed, a deeper understanding of how AI models work increases visibility into unknown vulnerabilities and flaws. (23) It can help to quickly identify and correct (debug) errors in critical situations. Thus, user controls can be strengthened. If the AI generates unexpected results and explanations from the AI
model are not reasonable or inconsistent with clinical experience, or if there is a potential risk of bias/discrimination, clinicians can suspend the AI and review for hidden errors or bias. (29) For example, if a patient is unexpectedly evaluated as having a low possibility of obtaining a favorable outcome and the result is mainly driven by ethnicity or socioeconomic status of the patient, clinicians may suspect an AI model is being influenced by hidden discrimination or bias in the training data. (30, 31)

Third, the other reason for developing explainable AI models is the necessity to continually improve them. If AI models can indicate the reasons for generating specific results, it provides useful information for further improvements. (32) For example, when we find that an AI model does not accurately predict an outcome, we can review how it is working using explainability features from the AI. If a certain predictor highly contributes to prediction in the model, but this variable carries a risk of measurement bias due to the absence of standardized definitions in clinical settings, we may be able to improve the model by excluding such a variable or standardizing the input. In this way, explainable AI can lay the groundwork for continuous iteration and improvement.

Furthermore, through explainable AI, we may find some new ideas, hypotheses, and knowledge. (22) For example, if the explanation from the AI model shows an unexpected contribution of a certain risk factor to the prediction of outcomes, it may provide a novel hypothesis that this factor may be associated with outcomes. For example, in a study of an AI model investigating the clinical subgroups of cardiac arrest patients with high effectiveness of ECPR, one of the model features showed that creatinine values were associated with outcome. This led to the development of a novel score for considering indications for ECPR which included creatinine values. (33) As the importance of explainable AI grows, it is taking on a more critical role for future AI applications in clinical settings.
4. **How does explainable AI work?**

Explainable AI encompasses three main approaches. (18-20) The first is pre-modeling explainability such as data visualization, summarization, and transformation. (34) Before deploying AI in clinical settings, it is essential to grasp the data structure, patients’ characteristics, time trends, and proportion of the outcome for an appropriate understanding of the AI. This may include simple descriptions such as means, standard deviation, range, and missing data using data visualization or summarization. Data transformation is also crucial. It refers to the process of changing row data into a format or structure in which models can be successfully developed. (34) For example, when developing ML models using clinical data such as the date of incidence, time of emergency call, and hospital arrival, those data are generally transformed to the month or day of incidents or the duration between emergency call to time arrived at a hospital which can be easily analyzed. This data transformation is more applicable to developing models and understanding how the model works.

The second approach is developing the interpretable model, where the focus is on models that are inherently understandable or are a blend of different model types. (18-20) Models exhibit various levels of interpretability and transparency: at the level of the training algorithm (referred to as "algorithmic transparency"), at the component level (known as "decomposability"), and at the level of the entire model (or "simulatability"). (22) For example, a tree model might pose a human-understandable question, such as whether the patient is younger than 65 years old, (Figure 2). Such questions clarify the prediction process, enhancing algorithmic transparency. This model can also be broken down into individual segments, like differentiating patients with or without initial VF. (Figure 2) This allows users to grasp how each segment contributes to the overall output, showcasing decomposability. Taking the case of a 40-year-old cardiac arrest patient with initial VF as an instance, we can trace through the entire prediction pathway of the model, estimating a
survival probability of 30%, all without specialized mathematical tools. This demonstrates simulatability. In essence, models possessing these characteristics are transparent and user-friendly in their interpretation.

The hybrid interpretable model approach has also been proposed. (18-20) It includes a set of methods that attempts to combine a complex black-box model with an inherently interpretable model to build an interpretable model that achieves comparable performance to the black-box model. The AutoScore framework is an example of this hybrid interpretable model approach. (35, 36) In this framework, the process of developing the ML model is complicated; however, the final model is described as a clinical score that is familiar to clinicians. (5, 35, 37)

The last method is called post-modeling explainability. (18) It helps make complex AI models easier to understand after model development. These techniques were created based on how people try to understand things.

5. Post-modeling explainability

In this section, we introduce some examples of post-modeling explainability. There are several categories of post-modeling explainability including visualization, textual justification, making the model simpler (simplification), and showing which parts of the data are most important (feature relevance) (Figure 3). (22)

Text explanation improves the understanding of ML models by generating text-based explanations in the form of phrases or sentences using natural language generation methods. Some examples are AI models to classify pathological images, with a function to provide user-friendly explanations. (38) These models can generate sentences such as “The input image is diagnosed as tissue A type for sure because it could not be misclassified to any other tissue types”, “The input image is suspected as tissue B type, and there is a low possibility that it could be tissue C
type, D type, or E type,” or “The input image is tissue A type. However, there is a possibility that it could be tissue F type.” These explanations about the possibilities of misclassification can provide the rationale for certain predictions and help clinicians with their decision-making. (Figure 3)

Visual explanations describe models by applying techniques that aim at visualizing the model’s behavior. Popular literature makes use of dimensionality reduction techniques to make simple visualizations that can be easily interpreted by humans. Visual explanation is considered to be particularly effective in conveying complex interactions between variables. (18-20) For instance, when describing black-box models to predict the probability of favorable outcomes for cardiac arrest patients, the interaction between probability and some factors (such as age and time to hospital) are difficult to recognize. To explain the interaction in this model between age and time to hospital, we can set these variables as the x-axis and the y-axis and make a scatter plot to indicate the distribution of the possibility (Figure 3, 4). (39) It enables us to grasp the distribution of the predicted probability between the factors and interaction visually.

Simplification is a technique that creates a more straightforward interpretable model from a black box model. (22) This simpler model aims to perform similarly to the original while being less complicated. One of the examples of simplification is to identify a single decision tree as the representative of a Random Forest. (40, 41) The Random Forest model consists of an ensemble of numerous decision tree models, it aggregates predictions from these individual trees to produce a final output (Figure 3). (42) Although this approach is commonly utilized in medical research, it can be challenging to interpret due to its ensemble nature. By identifying a single tree that captures the primary patterns and behaviors of the entire forest, we achieve a balance between interpretability and performance. (40, 41) This representative tree can be visualized, providing insights into the decision-making process while being rooted in the same foundational logic as
Another example is LIME (Local Interpretable Model-agnostic Explanations). This method approximates a complex black-box model with a linear regression model which is simpler and more easily interpretable. This is achieved by generating numerous samples of input data, predicting their outputs using the original model, and then training a linear model on these samples, placing more emphasis on those close to the original data point. LIME can provide the feature importance that contributes most to each prediction and helps users grasp which factors the model considers most crucial, as explained below.

Explanation by feature relevance aims to provide post-modeling explainability by assessing the internal processes of an algorithm. This type of explanation is commonly utilized in ML models in emergency medicine. It calculates relevance scores for all variables managed by the algorithm. These calculated scores quantify the importance of features critical to the model's decisions. SHAP (SHapley Additive exPlanations) is one of the methods to evaluate the contribution of each input feature to how the AI model operates. Similar to LIME, SHAP performs local linear approximations to explain the predicted risk for each individual, but by using a different approach that allows more desirable properties than LIME in points of local accuracy and consistency (for details please see the reference).

SHAP can quantify and visualize how each factor increases or decreases risk from the baseline to reach the predicted risk for each individual using a waterfall plot (Figure 5). Consider a ML model that predicts the survival rate of cardiac arrest patients using four factors: sex, age, witness status (yes/no), and time to the hospital from the emergency call (minutes). For example, consider a 45-year-old female with witnessed arrest, and the time to hospital is 37 minutes. This ML model predicts her survival rate as 21.5%. The waterfall plot in Figure 5 demonstrates how these four factors influence the prediction of ML. In this case, the baseline of the predicted value (f(X)), i.e., the average prediction across all cases is -1.495, which translates to a baseline survival rate of
18.3% via the inverse logit function, given by \([1/(1+e^{-f(X)})]\). The witnessed status (being 'yes') raises the predicted value by 0.103, which is equivalent to an increase in the survival probability to 19.9% from the baseline. This 0.103 increase in predicted value attributable to the witnessed status is the SHAP value of this factor for this individual. The sex (being female) has a SHAP value of 0.147 which further increases it to 22.4%. Age (being 45 years old) has a SHAP value of 0.22 which boosts it to 26.4%. However, the time taken to reach hospital, which has a negative SHAP value of -0.268, reduces the survival rate, bringing it down to the final predicted value of 21.5% for this particular case. Through such a case, we can understand how each variable impacts the model’s prediction using a waterfall plot and SHAP value. As in this example, SHAP can provide local explanations – that is, explanations for individual predictions. When the contributions to the predicted risk of each factor are visualized across all patients in a beeswarm plot (Figure 5), it shows the relationship between factor levels and contributions to prediction and facilitates a straightforward comparison of the extent each factor may impact the prediction.

SHAP is also valuable for global explanations, which refers to understanding how the model behaves overall, considering all data points, using the average absolute SHAP value. For example, the barplot in Figure 5 indicates the average absolute SHAP value of ‘time to hospital’ is the highest, and those of ‘witnessed’ and ‘age’ are also considerably high compared to that of ‘sex’. It suggests that across the board, time to hospital, witnessed status, and age are strong predictors of survival (Figure 5). This dual capability allows users to understand both specific decisions made by the model and the broader trends and behaviors that the model exhibits across different data points.

6. **Challenges in implementation into clinical settings**

Even though the explainability of AI has advanced, there are still several challenges to the
implementation of AI models in clinical settings. One of the issues is whether the explanation is acceptable and trustworthy enough for clinicians or patients from the perspectives of the audience. (16, 18-20) Previously, an explainable AI model was developed to predict the deterioration of patients with subarachnoid hemorrhage in the intensive care unit, and to enhance the implementation of the AI tool, the mind gap between the developers and clinicians was investigated by interviewing clinicians and developers. (51) Through the interview, the study found that the developers believed that clinicians must be able to understand what the model was doing; thus, they developed the AI model with explainability providing SHAP values as mentioned above. In contrast, from the perspectives of the clinicians, the SHAP value was not thought to be helpful to understand or trust the AI model. They were more focused on the clinical plausibility based on the pathophysiological rationale or their clinical experience and holistic approach referring to the multispectral clinical information. As illustrated in this example, the kind of explainability required depends on the audience and contexts where an AI model is used. (19) In emergency settings, the contexts and patients’ conditions are always changing rapidly. Especially, during resuscitation, since it is the most time-critical situation, clinicians may not have adequate time to try to understand how AI models work. Therefore, their explainability should be understandable without taking too much time to interpret. Furthermore, it is also a considerable challenge to assess the quality/effectiveness of explainability. A previous systematic review indicated that the methods for assessment of effectiveness were varied and there are few established methods to assess explainable AI. (52) Standardized approaches to the measurement of the effectiveness of explainable AI might help to integrate AI into clinical settings and act as a tool to communicate between clinicians, researchers, and developers. (28) Last but not least, with increasing emphasis on ensuring fair and trustworthy AI-assisted decision-making in clinical settings, it would be useful to explore how explainable AI can contribute to fair AI development through a multi-disciplinary approach. (53) Considering such situations, we should highlight the importance of collaboration between AI developers, researchers, and clinicians in designing
explainable AI systems for improving the effectiveness, usability, and reliability of explainable AI in healthcare.

Conclusion
This paper summarizes the concept of “explainable AI” for clinicians working in the emergency medicine field. In an era where AI’s role is increasingly anticipated, emergency physicians and researchers will likely need knowledge of AI under pressure of necessity. Furthermore, a multidisciplinary approach is essential to develop trustworthy AI for actual use in clinical emergency medicine fields. This review will help convey the concept of explainable AI for clinicians working in emergency departments.

DECLARATIONS

Ethical approval
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials
Not applicable.

Competing interests
YO has received a research grant from the ZOLL Foundation and overseas scholarships from the Japan Society for Promotion of Science, the FUKUDA Foundation for Medical Technology, and the International Medical Research Foundation. These organizations have no role in conceptualization, data collection, or writing the manuscript. MEHO reports grants from the
Laerdal Foundation, Laerdal Medical, and Ramsey Social Justice Foundation for funding of the Pan-Asian Resuscitation Outcomes Study an advisory relationship with Global Healthcare SG, a commercial entity that manufactures cooling devices; and funding from Laerdal Medical on an observation program to their Community CPR Training Centre Research Program in Norway. MEHO is a Scientific Advisor to TIIM Healthcare SG and Global Healthcare SG.

**Funding**

This study was supported by a scientific research grant from the JSPS KAKENHI of Japan (JP22K21143) and the Zoll Foundation. YO has received an overseas scholarship from the Japan Society for the Promotion of Science, the FUKUDA Foundation for Medical Technology, and the International Medical Research Foundation.

**Acknowledgment**

We appreciate A/Prof Liu Nan for his support and helpful advice.

**Author contribution**

YO, Conceptualization, Writing – original draft.
YN, Methodology, Writing – review & editing.
MEHO, Conceptualization, Writing – review & editing.

All authors agree to be accountable for all aspects of the work.
VF, Ventricular fibrillation, CPR, and cardiopulmonary resuscitation.
Figure 2. Decision tree model to predict the possibility of survival.

You can trace the algorithm to generate the output without any computers or devices.
Figure 3. The methods of post-explainability

Input data → Black box → Survival rate 3% → Post-hoc Explainability

Text explanation: “Because this patient is older, without witness and non-VF.”

Visualization:
- Age vs. Time to ROSC
- High risk of mortality

Simplification:
- Witness
- Age > 65

Feature relevance:
- Age
- Initial rhythm
- Witness
- Male

Survival rate 3%
This plot illustrates the relationship between the factors (age and time to hospital) and the predictions made by the random forest model, a prevalent machine-learning model. The random forest model was constructed to predict the survival probability using simulated data from cardiac arrest patients. This data includes factors such as age, gender, whether the event was witnessed, the provision of bystander CPR, and the time taken from the call to hospital arrival.

The x-axis represents age, while the y-axis denotes the time from the call to the hospital's arrival. Blue dots indicate cases with a low probability of survival (<25%), red dots highlight cases with a high likelihood of survival (>35%), and white dots signify intermediate cases (around 30%). While users might not grasp the intricacies of how the AI model predicts the survival rate, they can broadly infer that patients who are younger and have a shorter time to hospital are predicted to have higher survival rates. Conversely, older patients with a longer time to hospital are estimated to have lower survival chances. Moreover, users can also glean insights into the interactions between factors and predictions.
Figure 5. The example of SHAP values

(1) Waterfall plot

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
<th>SHAP Value</th>
<th>Contribution (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Witness</td>
<td>Yes</td>
<td>+0.103</td>
<td>19.9%</td>
</tr>
<tr>
<td>Sex</td>
<td>Women</td>
<td>+0.147</td>
<td>22.4%</td>
</tr>
<tr>
<td>age</td>
<td>45</td>
<td>-1.495</td>
<td>18.3%</td>
</tr>
<tr>
<td>TimetoHp</td>
<td>37</td>
<td>-1.293</td>
<td>21.5%</td>
</tr>
</tbody>
</table>

(2) Beeswarm plot

(3) Bar plot

TimetoHp: Time to hospital

(1) The waterfall plot indicates the predicted value (i.e., f(X)) of each factor of the case that a 45-
year-old female with witnessed, and the time to the hospital is 37 minutes. The change in the predicted value from the baseline (-1.495, corresponding to a survival rate of 18.3%) to the predicted value for this particular base (-1.293, corresponding to a survival rate of 21.5%) attributed to each factor is the SHAP value of each factor. The survival probability is calculated by the inverse logit function given by \([1/(1+e^{-(\alpha X)})]\).

(2) The beeswarm plot demonstrates the SHAP values of each factor across all cases. A central vertical line (at SHAP value=0) indicates the point of 'no influence' on the prediction. If a point is to the right of this line, it means that the factor influences the model's predictions in a positive direction (increase the survival rate); if it is to the left, it influences them in a negative direction (decrease the survival rate). The color of the dot represents the value of factors. For example, the red and blue are female and male in sex, “witnessed” and “not witnessed” in witness status. Also, the red and blue means older and younger in age.

(3) The barplot displays the absolute SHAP values, indicating that factors such as time to hospital, witnessed status, and age are relevant predictors of survival across all the cases.
Reference


toward responsible AI. Information fusion. 2020;58:82-115.


30. Okada Y, Mertens M, Liu N, Lam SSW, Ong MEH. AI and machine learning in


