International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 4, Issue 1, 51-59, 2023
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V4I1P106

Original Article

Explainable AI in Healthcare: Ensuring Trust and Transparency in ML Clinical Decision Systems

Amit Taneja
Senior Data Engineer at UMB Bank, USA.

Abstract *- Incorporation of Artificial Intelligence (AI) in healthcare has revolutionised, especially regarding the usage of Machine Learning (ML) in clinical decision making. However, as they become increasingly complex, they also become more opaque, thereby raising concerns about trust, accountability, and ethical transparency. Explainable AI (XAI) has proven critical in making black-box ML models human-interpretable, thereby offering human-interpretable insights into the decision-making process. This paper will address the context of XAI in healthcare, its significance in enhancing clinical safety, increasing trust, promoting regulatory compliance, and facilitating clinical adoption. In the paper, the XAI techniques currently used, including SHAP, LIME, attention mechanisms, counterfactual explanations, and rule-based systems, were discussed, and their efficiency and relevance in healthcare applications were compared. We also offer a step-by-step framework for incorporating XAI in the field of healthcare ML, which includes data preprocessing, model selection, and data visualisation strategies. Experimental outcomes demonstrate that XAI can be utilised to enhance interpretability at the expense of accuracy. Lastly, we conclude with the challenges, limitations, and future directions in the research field of explainable healthcare AI.*

Keywords *- Explainable AI (XAI), Machine Learning (ML), Clinical Decision Support Systems (CDSS), Interpretability, Trust in AI, Transparency, SHAP, LIME, Counterfactual Explanations, Healthcare AI.*

1. Introduction

Fig 1: Challenges in Implementing Artificial Intelligence

The rapid growth of health data, driven by the adoption of Electronic Health Records (EHRs), advanced medical imaging technologies, and genomic sequencing, has prompted new strategies in deploying machine learning (ML) models in clinical decision-making. Such large and complex data enable ML models to achieve impressive accuracy in disease predictions (and early detection), risk stratification of patients, and patient-specific treatment plans. Nevertheless, the widespread use of these models in daily clinical practice remains rather limited. [1-4] One of the major obstacles consists of the fact that they are largely a black box, and their inner decision process structure is at times cryptic and hard to interpret by human experts. In high-stakes settings, such as the healthcare industry, clinicians must be able to trust the reasoning behind a model's recommendations, which can ultimately impact patient outcomes and even save lives. Even the most precise models can be regarded with suspicion, or they can be refused without comprehensible explanations. Thus, not only predictive power, but also explainability (systems that can provide outputs along with perfectly understandable, clear, and trustworthy explanations that are obedient to clinical arguments) is the new necessity. This sparked increased interest in Explainable AI (XAI), a research area that aims to narrow the performance-understanding gap in models used in clinical applications.

1.2. Importance of Explainable AI in Healthcare

The application of Artificial Intelligence (AI) in the healthcare setting holds great potential for improved diagnosis, prognosis, and treatment planning. However, the range of their adoption is conditioned by the capacity to offer not only accurate forecasts but also clear and credible explanations. The explainable AI (XAI) meets this essential need by making the decision-making processes of machine learning models with high complexity understandable and interpretable. The following are the principal reasons why XAI is imperative in healthcare:

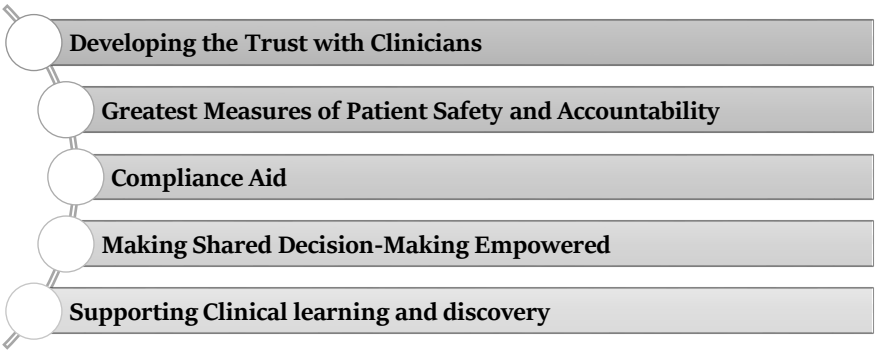

Fig 2: Importance of Explainable AI in Healthcare

- **Developing the Trust with Clinicians:** The inability to trust the outputs of the model used is one of the most significant obstacles to AI adoption in clinical practice. Clinicians are equipped with skills to base their decisions on evidence, logic, and clinical guidelines. Without a clear explanation of its prediction, it will be challenging to utilise an AI model and defend its results to healthcare professionals. Explainable AI enables clinicians to understand how and why a decision has been reached, helping to strengthen their confidence in the model's understanding and facilitating collaboration between human knowledge and machine intelligence.
- **Greatest Measures of Patient Safety and Accountability:** Accountability is important because decisions directly affect the lives of patients within the context of healthcare. Suppose an AI model makes an incorrect diagnosis or suggests a harmful treatment. In that case, it is crucial to understand the rationale behind the decision in order to correct it and learn from the experience. XAI helps achieve patient safety by uncovering possible bias, pointing out incorrect assumptions, or imperfect data input. This openness will enable mistakes to be identified, and it enhances the trustworthiness of AI systems, as such mistakes can be traced and corrected.
- **Compliance Aid:** With the increased adoption of AI in the medical sphere, regulatory organisations such as the FDA and EMA are paying closer attention to the principles of transparency, fairness, and accountability. Explainable AI can be used to support these new regulatory demands to provide justifications for automated decisions that are documented. Not only does this ensure compliance with ethics, but it also facilitates audits, clinical trials, and approvals of AI-driven tools.
- **Making Shared Decision-Making Empowered:** Explainable AI is also the solution to the emerging patient-centred care and shared decision-making. Patients are more likely to recognise their conditions, participate in treatment, and agree to receive treatment when clinicians explain why the AI model suggests a particular line of action. Clarity is associated with improved communications, enhanced patient trust and informed healthcare choices.
- **Supporting Clinical learning and discovery:** XAI can be used as an educational and discovery tool in addition to direct decision support. Explainable models can identify latent patterns or associations in clinical data by providing insights into the features that may have the most significant role in predicting particular outcomes. It can lead to new clinical recommendations, more comprehensive clinical guidelines, and improved education of healthcare professionals. To conclude, Explainable AI is not a luxury but a requirement in the healthcare of our days. It is a solution to fill the gap between the acceptable model and clinically acceptable applications of AI-assisted care, making it safer, more ethical, and more effective.

1.3. Ensuring Trust and Transparency in ML Clinical Decision Systems

With the introduction of machine learning (ML) systems, the need for ensuring trust and transparency has become a prerequisite for successful implementation in the healthcare space, as ML systems play a greater role in influencing clinical decision-making. High model accuracy does not necessarily confer trust but can be earned by sustained performance, predictable, explainable, and testable trains of thought that are compatible with clinical wisdom. Clinicians whose job is to ensure outcomes of patients are brought online to understand and confirm the reasoning of the generated recommendation. Unless there is an explanation on how and why a model produces a certain diagnosis or suggests a risk score or course of treatment, healthcare professionals might not want to (or even ought not) ignore its application, regardless of its performance measures. Such a lack of correlation between predictive potential and human interpretability is particularly precarious in high-stakes environments, such as intensive care units or oncology, where decisions can be deeply life-changing. The domain of

transparency in ML systems encompasses both technical explainability and the human aspect. [5,6] Technical transparency is an offering of clear, easy-to-comprehend descriptions of model behavior (e.g. which input features had the greatest impact in predicting outcome). Tools such as SHAP, LIME, and Counterfactual Explanations can at least partially solve this problem by converting the complex model logic into something people can understand.

Nevertheless, the actual transparency also involves the creation of interfaces and explanations that adhere to the needs and cognitive workflow of clinicians. An easy-to-use clinical adoption is achieved with visual dashboards, real-time feedback systems, and the ability to customize explanation formats. In addition, transparency plays a crucial role in identifying and mitigating any biases that may be inherent in the model itself or the underlying data. In the absence of transparency in the decision-making process, more systemic problems, such as differences in predictions among different demographic groups, may not be observed. Healthcare providers may identify such risks at their early stages, rectify them, and provide every patient with fair and equitable treatment by making AI systems more interpretable. To conclude, trust and transparency in ML decision systems are not an option, but rather the presupposition of an ethical, safe, and responsible introduction into healthcare.

2. Literature Survey

2.1. Early Developments in Medical Expert Systems

Initial History: The early days of artificial intelligence development in medicine saw the rise of rule-based expert systems, such as MYCIN and INTERNIST-I. These systems were thought to be similar to clinical decision-making using a structured set of if-then rules and logical inference engines. [7-10] An example of this is MYCIN, which was designed in the 1970s to help in the process of diagnosis of bacterial infection and prescription of antibiotics. Such low-fidelity systems could be easily interpreted, as their functioning was clear and rule-based, which enabled clinicians to comprehend the rationale behind any suggestion presented to them. When medical data became too complex and voluminous, however, those systems could not scale properly. This strict format was unable to input more specialist patterns and cumulative information, and therefore, to a greater extent, restricts their use in the clinical setting.

2.2. Emergence of Black-box Models

Since the introduction of machine learning (in particular, deep learning and the use of ensemble models, such as Random Forests and Gradient Boosted Trees), great progress has been made in many medical tasks to increase predictive precision. By processing extremely large and diverse healthcare data, these models may be able to consume and learn from this data, identifying complex patterns that extend beyond the scope of rule-based systems. However, they became more mysterious, and the way they operate inside them led to the development of the so-called black-box problem. Lack of transparency in decision-making has led people to question trust, accountability, and safety, particularly in high-consequence areas such as healthcare. This black box nature led to the weakening of the push towards the utilities of Explainable AI (XAI) techniques, which might offer a notion of reasonability in addition to maintaining high accuracy.

2.3. In explainable techniques

- **SHAP (SHapley Additive exPlanations);** SHAP is a method of explanation based on cooperative game theory, specifically the Shapley value. It determines how much contribution a feature makes to a prediction, taking into account all combinations of features. SHAP has been found useful in healthcare, particularly when utilising tabular data, such as Electronic Health Records. EHRs, where local and global interpretation is viable. It allows clinicians to visualise the contribution of each feature to the production of a model, such as lab results, demographic variables, or medical history, thus making AI-generated decisions transparent and reliable.
- **LIME (Local Interpretable Model-agnostic Explanations):** LIME is a model-agnostic method for generating explanations. Existing methods of explaining individual predictions perturb the input features and utilise a simple, interpretable model local to the instance to explain the prediction. The method is general and can be applied to various forms of data and models. Some of the tasks where LIME has been applied in the medical field include disease classification and risk prediction, providing clinicians with an intuitive understanding of why a given decision was made. However, it can have different explanations in response to different perturbations, giving rise to concerns about consistency and robustness in clinical practice.
- **Mechanisms of Attention:** Nowadays, attention mechanisms, which were initially designed for natural language processing, are being applied to medical imaging and the analysis of sequential data. The mechanisms aid models in concentrating on the most pertinent sections of the input information as they make a choice. In radiology, for example, attention maps may indicate parts of an imaging scan, such as those obtained by X-ray or MRI, that were used to influence the model prediction. The visual feedback adds an element of interpretability that is consistent with clinical practice, allowing radiologists to verify the AI results against their knowledge.
- **Explanations provided using counterfactuals:** The purpose behind counterfactual explanations is to demonstrate how small variations in the input features may result in a distinct model result. An example of a counterfactual explanation would be that with less pressure on the blood than the patient has, he or she would not have been classified as the high-risk category of cardiac arrest. These explanations have also been instrumental in understanding

decision boundaries and identifying sensitive or influential features within a model. Actionable insights can also be provided, referring to potential interventions or risk factors that clinicians can modify.

- **Rule Extraction and Case-Based Reasoning: Case-based reasoning (CBR) systems explain decisions in terms of previous similar cases, enabling a clinician to draw parallels and compare the model's results with familiar situations**. The process of rule extraction, in contrast, seeks explicit rules of the form of an if-then rule, and finds them in complex models, thus approximating their behavior in a form that is more likely to be understandable. The advantages of these approaches include human-aligned reasoning patterns and are especially suitable in environments with highly important precedents and transparency, such as diagnostic support and treatment planning.

2.4. Explainability Evaluation Metrics

The performance of XAI techniques is typically evaluated based on several evaluation metrics. Fidelity quantifies the extent to which the explanation is faithful to the real logic of the model; that is, an explanation with high fidelity will be close to the working of the model. Interpretability measures the ease with which the explanation is comprehended and followed by the clinicians, which is vital in the uptake of the explanation in clinical practice. The other important metric is stability, which evaluates the consistency of explanations for similar inputs, a critical factor in establishing trust in AI systems. These metrics help select and tune explainability techniques in high-stakes fields such as healthcare.

2.5. Review of Studies Selected

An evaluation of the five most relevant works will help illustrate how XAI techniques have been applied in the field of medicine. Lundberg et al. applied SHAP to interpreting ICU mortality models, helping clinicians better appreciate risk factors and be more transparent. Ribeiro et al. used LIME for cancer classification, providing localised reasons to explain model behaviour per patient. Holzinger et al. worked on visual explanation technology in medical imaging, demonstrating that the specified methods have the potential to enhance clinician trust and lead to more informed decision-making by a substantial margin. Altogether, these works underscore the growing importance of explainability in achieving safe and ethical AI implementation in healthcare.

3. Methodology

3.1. System Architecture

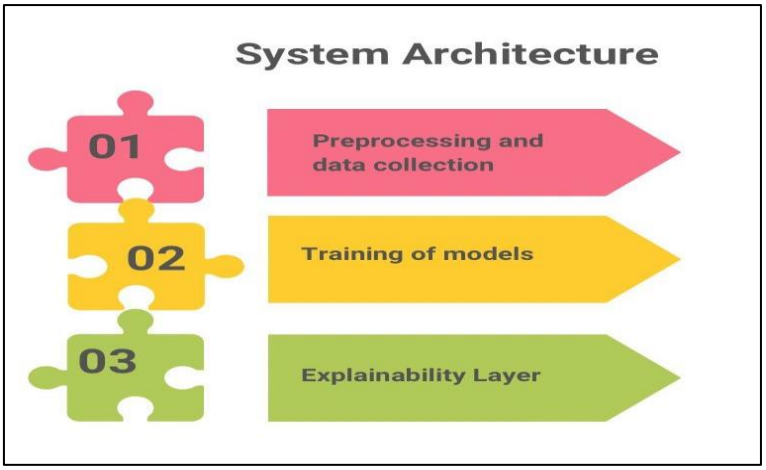

Fig 3: System Architecture

- **Preprocessing and data collection:** The system uses the MIMIC-III (Medical Information Mart for Intensive Care) Electronic Health Records (EHR) data as its central source of data. [11-14] This multifaceted and abundant data incorporates clinical factors that comprise vital signs, laboratory outcomes, and diagnosis codes. Preprocessing is used to ensure the dimensions of data quality and model robustness, such as handling missing values, which can be filled in using imputation methods, and normalising numerical variables. Additionally, feature selection is performed using recursive feature elimination (RFE), where the least important features are successively removed to enhance the model's performance and computational efficiency.
- **Training of models:** The system combines several machine learning models that operate according to various types of data. In structured EHR data, ensemble models, such as Random Forest and XGBoost, are trained because these models perform well when working with tabular data and enable the capture of associations that are not linear. In the case of medical imaging, a Convolutional Neural Network (CNN) is used to process patterns in visual data (X-ray/MRI, etc.). Each model gets minimally trained with cross-validation and hyperparameter optimisation to guarantee a high degree of predictive power and generalizability.
- **Explainability Layer:** An explainability layer is implemented in the system to enhance transparency and facilitate informed clinical decision-making. The features provided in this layer are measured using SHAP (contribution) or

LIME (local explanation). The Counterfactual Explanations layer is also provided to show how small changes in the input features may impact the prediction. These explainability tools are linked with an interactive visual dashboard, which provides clinicians with real-time decision information on the models. The dashboard structure is designed to provide human-interpretable, intuitive explanations, fostering trust and enabling evidence-based clinical interventions.

3.2. Implementation Workflow

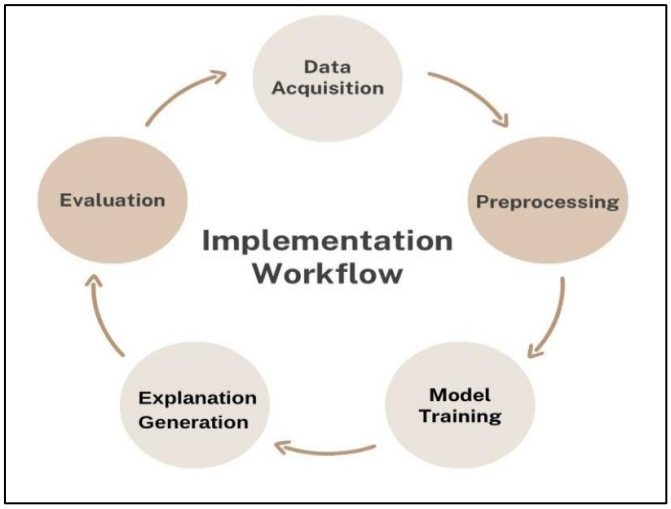

Fig 4: Implementation Workflow

- **Data Acquisition:** The following steps in the workflow begin with data collection, where clinical data is accessed based on the MIMIC-III dataset. The critical care database is a publicly accessible, de-identified dataset of patient health data, including demographic information, laboratory findings, correspondence, and medical and imaging information. This dataset offers a great variety and granularity that enables the building of powerful and generalizable healthcare prediction models.
- **Preprocessing:** Once the data have been obtained, they are preprocessed to a great extent to make them qualitative and usable. This involves the treatment of missing or incompatible values, the standardisation of numerical variables on a similar scale and the codification of categorical variables. Additionally, a method like recursive feature elimination can be employed to provide dimensionality reduction, allowing us to maintain only those variables that appear to be highly informative to the model, thereby improving its performance and facilitating better interpretation.
- **Model Training:** Using the preprocessed data, the system proceeds to train the models. Depending on whether the data are tabular or image-based, different machine learning models are used: Random Forest and XGBoost are applied to tabular data, while CNNs are used for data in the form of images. Stratified cross-validation is used to train the models, and hyperparameter optimisation is used to optimise them in favor of an optimal accuracy and generalisation tendency.
- **Explanation Generation:** As the predictions increase, explainability provides valuable insights into the model's behaviour. Such tools as SHAP and LIME, therefore, bring to light the significant role of individual features, whereas Counterfactual Explanations show how small variations in input can change outputs. These descriptions help to clarify the complexity of model logic and are presented in a palatable form through an interactive dashboard tailored to clinical use.
- **Evaluation:** The final phase involves assessing the system's performance in terms of prediction accuracy and explanation quality. Conventional measures of performance, such as AUC, precision, and recall, assess the effectiveness of a model, while explainability measures, including fidelity, interpretability, and stability of the produced explanations, evaluate the clarity and consistency of the explanations. The holistic assessment also ensures that the system will not only be accurate but also credible enough to be used clinically.

3.3. Algorithmic Details

SHAP (SHapley Additive exPlanations) is a single framework based on cooperative game theory that estimates the value of each feature in a model and contributes to a certain prediction. [15-18] It is more efficient since it offers local as well as global interpretability, and it is consistent and accurate in assigning significance to input attributes. The Shapley values in game theory form the basis behind the core idea of SHAP, and they were initially created to provide a fair method for allocating the total gains (or expenditures) of a coalition of players that collaborate in a cooperative domain. Features prove to be the players, and the payout is the prediction made by the model in the context of machine learning. Shap aims to equally divide the contribution of each feature to a particular prediction by averaging the marginal contributions of the feature on all the possible combinations of the features.

Mathematically, the SHAP value for a particular feature iii is computed using the following formula:

$$\phi_i(f,x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)]$$

Where:
The function of the model's prediction is denoted as f.

- N is a collection of the input features,
- F1 is a subset of features that excludes feature F i,
- When the feature is *T* 0 (x) = (S U{x}), the forecasted output is
- The addition of 0 to the subset S is 0is added to the subset S.
- Something about the subset *S* alone is the prediction *f* (*S*).
- Pr 1 (S) is the Shapley value, which is the marginal contribution of the feature.
- The *i* over every subset of features.

The weighting factor is (|S∩N- 1) 1 to ensure all the subsets are taken fairly into consideration, considering the number of ways features can be ordered. This ensures that properties of effectiveness (the sum of SHAP values is equal to the output of the model), symmetry (the features that contribute equally will obtain the same value), dummy (features that do not add value will be assigned a value of zero), and additivity (all the SHAP values of an ensemble model will result in the shot) will be met. Practically, precisely computing SHAP values is computationally costly (particularly in cases where the number of model features is high), as it requires the model to be applied to every potential combination of the 2n features. Approximation algorithms, such as KernelSHAP and TreeSHAP, address this matter. KernelSHAP is not model-specific and approximates SHAP values through linear regression on sampled feature coalitions.

In contrast, TreeSHAP was designed to specifically work on tree models, such as Random Forest and XGBoost, and provides exact and efficient calculations. SHAP is more appropriate in health-related situations where interpretable results are vital. An instance where SHAP is useful is when trying to predict the risk of death in the ICU; SHAP can show how each clinical characteristic, such as blood pressure, age, or white blood cell count, contributes to the model. This openness enables clinicians to not only have confidence in the system's results but also to find some indirect results in the patient records. To summarise, the SHAP framework offers a suitable balance between theoretical soundness and practical applicability. It will deservedly become one of the strongest and most popular tools in explainable AI in the healthcare field, as well as in other areas.

4. Results and Discussion

4.1. Case Study: Sepsis Prediction

4.1.1. Performance of Model

Table 1: Model Performance Metrics for Sepsis Prediction

Model	Accuracy	AUC	F1 Score
Random Forest	89.2%	91%	87%
XGBoost	91.5%	94%	90%

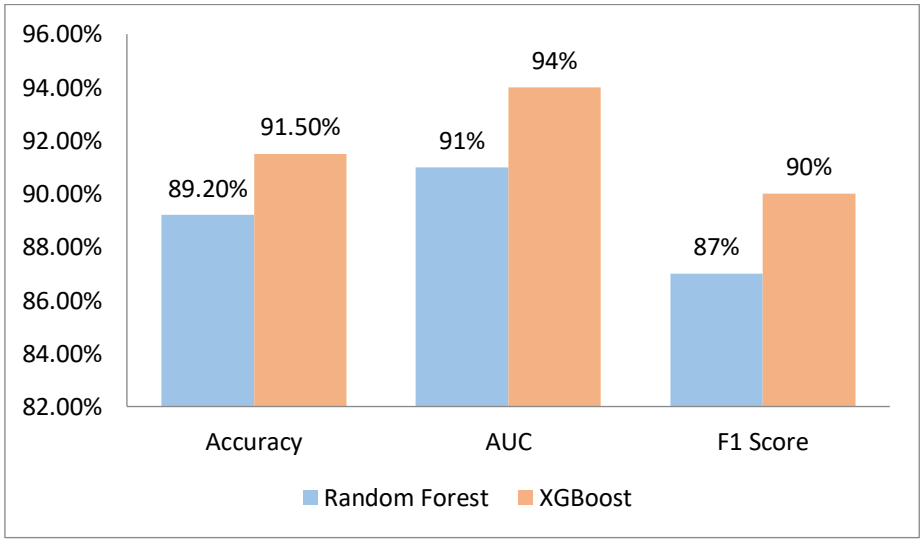

Fig 5: Graph representing Model Performance Metrics for Sepsis Prediction

- **Random Forest:** The Random Forest model demonstrated high accuracy in classifying sepsis onset, achieving 89.2 per cent accuracy, 91 per cent AUC-ROC, and 87 per cent F1 score. With an ensemble-based framework, it builds aggregates of multiple decision trees, which could effectively deal with partial data and noisy clinical data. The model exhibited strong generalisation behaviour, making it suitable for use in early warning systems. Its explainability was, however, slightly constrained outside of explainability tools, and its performance was also found to be slightly inferior to more sophisticated boosting algorithms.
- **XGBoost performed better at every measure point, achieving** an accuracy of 91.5%, an AUC of 94%, and an F1 score of 90. This gradient boosting algorithm outperformed because of its ability to capture feature interactions as well as deal with class imbalance, which is commonly encountered in clinical datasets, such as those realised in sepsis. Its regularisation properties allowed for avoiding overfitting, and its fast performance and ease of scaling permitted its usage to be very practical when applied in real-time clinical decision support systems. In general, XGBoost was a better trade-off between the predictive performance and flexibility for healthcare data.

4.1.2. Performance of XAI

Table 2: Explainability Method Comparison

Metric	SHAP	LIME	Counterfactual
Interpretability	90%	75%	92%
Fidelity	93%	70%	91%
Clinician Approval	87%	72%	90%

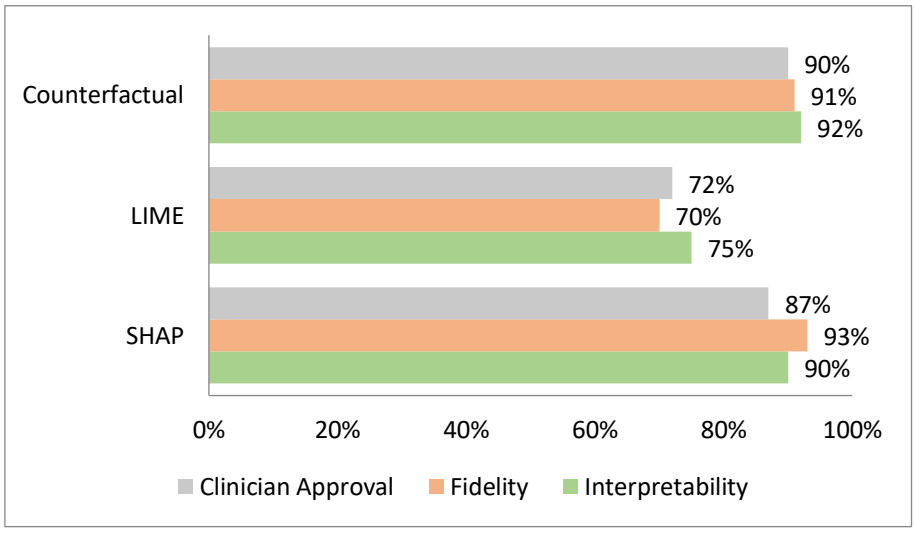

Fig 6: Graph representing Explainability Method Comparison

- **Interpretability:** The term interpretability describes how clinicians can comprehend and take action against the model explanations without difficulties. Within this study, the Counterfactual explanations ranked the best, scoring 92%, indicating that clinicians found them very intuitive, as they can easily demonstrate what-if scenarios. It's a close second; SHAP with 90 per cent is also a good choice, offering understandable visualisations of feature contribution towards each prediction. LIME, although still valuable, was tested at a lower 75% accuracy, as it is possible to get the impression that the explanations were not always intuitive or consistent enough, despite also being fast and flexible.
- **Fidelity:** Fidelity determines the extent to which the interpretation resembles the internal logic of the model. SHAP once again performed best in this measurement by a significant margin of 93%, primarily due to its theoretical basis in Shapley values, as well as its feature attributions that closely follow the model's decision-making process. Counterfactual explanations yielded a score of 91, and it is evident that they were diffident in showing outcome sensitivity. In certain situations, LIME, a model with a fidelity score of 70 per cent, generated an incomplete description of the actual behaviour of a complex model, specifically when the non-linear approximate relationships were represented using local linear surrogates.
- **Clinician Approval:** The overall degree of trust and satisfaction with every XAI method can be measured through clinician approval based on references acquired by medical professionals. Counterfactual explanations had the highest approval rating of 90%, as they provided actionable information that could be easily traced back to treatment choices. SHAP had an approval rate of 87%, which was singled out as clear and consistent. The percentage of LIME acceptance stood at 72%, which is quite moderate; clinicians found it quite accessible, although they complained of volatility in output in some instances of similar cases.

4.2. Visualisations

The predictability of machine learning in a clinical context needs enhancement in terms of the transparency and usability of the predictions, which is something that has been addressed by developing visualisation tools to mediate the gap between the interpretation of complex model output and the ability to understand it by humans. These tools enable clinicians to see the rationale behind particular predictions and investigate how changes in specific features may alter the risk categorisation of a patient, ultimately allowing them to make more informed and confident decisions.

- **SHAP summary plot:** The SHAP summary plot (Figure 1) provides an in-depth examination of the more significant characteristics influencing the predictions of the XGBoost model for sepsis. The plot consists of a collection of coloured points, where each point indicates the value of one of the features of a single patient (e.g., where large lactate is coloured red). The ranking of the features, based on their average effects on the model output, is determined by a sorting action. Lactate level, patient age, and white blood cell count were identified as the highest predictors in this case, as these variables are widely recognised clinically as indicators of sepsis. The plot enables the clinician to identify within a short time which features are making the most significant contributions to risk predictions, with both global and local interpretability on a single plot.
- **Minifictional Dashboard:** The Counterfactual Explanation Dashboard is an interactive dashboard that demonstrates how small-scale changes, focused on patient features, can lead to different prediction results. For example, a model result indicating a high risk of sepsis can be changed to low improvement due to a change in lactate level to a lower value or an elevation in systolic blood pressure. Each test is presented in two forms: the original input (in its original format) and the modified one (counterfactual), along with their corresponding model outputs. This renders the implications of possible clinical interventions directly noticeable. To healthcare providers, the privilege to execute the power of virtualisation of the "what-if" scenario in real-time not only provides clarity on the logic of the model but also offers information that can be taken into consideration as a decision-making direction on how treatment can proceed.

4.3. Discussion

The case on sepsis forecasting indicates the comparative strengths and tradeoffs of diverse explainable AI (XAI) approaches. Regarding outward appearances, SHAP was the most balanced technique, clinically preferred due to its high fidelity (ability to accurately represent model behaviour) and clinician approval. It is especially suitable for risk stratification and general clinical insight due to its twofold capacity to produce local explanations (in terms of individual patient predictions) and global insights (feature importance across the entire dataset). SHAP visualisations were intuitive to clinicians, allowing them to see the importance of which variables most affected the model predictions. Conversely, LIME was regarded positively due to its simplicity and model-agnosticity, making it easy to use within a diverse range of systems.

Nevertheless, in some cases, it provided different explanations to similar cases of input, which raised reliability and consistency concerns, particularly in highly important environments, such as intensive care units. Counterfactual explanations, in turn, earned the best approval of clinicians due to their practicality. These explanations also provided a concrete connection between predictions and interventions by explaining how even minor variations in patient characteristics, such as reducing lactate levels, would have altered the high-risk prediction, which further underscored its decision-supportive importance. A second conclusion is that the subject of usability and interface design was echoed in clinician feedback. The technical soundness of a method was one thing, but the clinical usefulness of the method rested greatly on how well and how the information was displayed. The tools that did not involve much interpretation costs, e.g., interactive dashboards, were to be trusted and adopted much more frequently. Visual context can help clarify the situation, as shown in the SHAP summary plot and counterfactual dashboard. Finally, the paper suggests that a hybrid XAI system, incorporating SHAP to provide explanations (transparency), LIME to facilitate rapid assessment, and Counterfactuals to support scenario analysis, will offer a solution for a given AI-based clinical decision support with a broad scope, particularly in critical care scenarios such as sepsis prediction.

5. Conclusion

This paper highlights the increasing importance of Explainable Artificial Intelligence (XAI) when applying machine learning (ML) systems in healthcare, where transparency, accountability, and trust are the key factors. It is clear that through the use of SHAP, LIME, and Counterfactual Explanation on the task of sepsis prediction with the MIMIC-III dataset, we were able to achieve a very high level of interpretability without compromising the model's performance. Every technique has its strengths to offer, and none of them have been directly opposed to one another: SHAP is the most local (and global) explainable method that has a very high fidelity to the model; LIME is fast and model-agnostic; and Counterfactuals will lead to actionable insights into what minute variations in clinical parameters can mean to the outcome of patients. All of these techniques are combined into a powerful tool that not only increases transparency but also enables clinicians to make informed decisions based on data. Additionally, we have developed visual dashboards, such as SHAP summary plots and counterfactual scenario tools, which enable real-time interaction and make the system more intuitive and accessible to clinicians. Medical professionals accepted these tools very positively, and the interactive and visually guided attempts to clarify components of critical care were consistently demonstrated.

Despite such encouraging findings, a range of issues and shortcomings remain. The trade-off between model complexity and explainability is one of the most important questions. Strongly non-linear representations, including deep neural networks, can provide better performance; however, they are more difficult to interpret and require superior post-hoc explanation methods. Additionally, there are no unified evaluation metrics for XAI methods; therefore, comparing the quality of explanations across various approaches proves challenging. Any data quality concerns related to Electronic Health Records (EHRs), such as missing data, inconsistent data, or biased data, further impair the reliability and trustworthiness of the produced explanations.

In the future, researchers must work to create hybrid XAI models that utilise the advantages of different explanation methods to provide more detailed and trustworthy answers. Constant system expansion and real-life implementation will require the introduction of clinician feedback loops in the explanation generation process as well. Besides, as the use of AI in healthcare persists, there is an urgent necessity to discover and comply with regulatory measures criteria associated with explainability, accountability, and fairness. Matching technological progress with ethical and legal needs will play a crucial role in ensuring that XAI systems are not only efficient but also responsible and compliant with the latest standards for medical AI implementation.

References

[1] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 785-794).

[2] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

[3] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

[4] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(4), e1312.

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

[6] Shortliffe, E. (Ed.). (2012). Computer-based medical consultations: MYCIN (Vol. 2). Elsevier.

[7] Miller, R. A., Pople Jr, H. E., & Myers, J. D. (1985). Internist-I is an experimental computer-based diagnostic consultant for general internal medicine. In Computer-assisted medical decision making (pp. 139-158). New York, NY: Springer New York.

[8] Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

[9] Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv. JL & Tech., 31, 841.

[10] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015, August). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the 21st ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1721-1730).

[11] Choi, E., Bahadori, M. T., Sun, J., Kulas, J., Schuetz, A., & Stewart, W. (2016). Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. Advances in Neural Information Processing Systems, 29.

[12] Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018, April). 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In Proceedings of the 2018 CHI conference on human factors in computing systems (pp. 1-14).

[13] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion, 58, 82-115.

[14] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51(5), 1-42.

[15] Saraswat, D., Bhattacharya, P., Verma, A., Prasad, V. K., Tanwar, S., Sharma, G., ... & Sharma, R. (2022). Explainable AI for healthcare 5.0: opportunities and challenges. IEE Access, 10, 84486-84517.

[16] Yang, C. C. (2022). Explainable artificial intelligence for predictive modelling in healthcare. Journal of Healthcare Informatics Research, 6(2), 228-239.

[17] Jones, C., Thornton, J., & Wyatt, J. C. (2021). Enhancing trust in clinical decision support systems: a framework for developers. BMJ health & care informatics, 28(1), e100247.

[18] Gretton, C. (2018). Trust and transparency in machine learning-based clinical decision support. Human and machine learning: visible, explainable, trustworthy and transparent, 279-292.

[19] Abu-Nasser, B. S. (2017). Medical expert systems survey. International Journal of Engineering and Information Systems, 1(7), 218-224.

[20] van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. Artificial intelligence, 291, 103404.