IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 26, Issue 6, Ser. 1 (Nov. – Dec. 2024), PP 29-36
www.iosrjournals.org

Explainable Artificial Intelligence (XAI) Techniques To Enhance Transparency In Deep Learning Models

Nasir Musa Imam[1], Abubakar Ibrahim[2], Mohit Tiwari[3]
(Dept. Computer Science And Engineering/ Vivekananda Global University, India)
(Department Computer Science / Central South University, China)
(Dept. Computer Science And Engineering/ Vivekananda Global University, India)

Abstract:
Deep learning *has revolutionized many fields, but caused the 'black-box' problem, where model prediction is not* ***interpretable*** *and* ***transparent****.* ***Explainable Artificial Intelligence (XAI)*** *attempts to overcome this problem with the help of Interpretability and Transparency in AI systems. We review important XAI methods focusing on LIME, SHAP and saliency maps that explain the elements behind model predictions. The paper discusses about the role of Explainable Artificial Intelligence (XAI) in high-stake fields such as healthcare, finance and autonomous systems, emphasizing on why trust is important for these sectors and how they help adhere to regulations while promoting ethical AI use. Despite the promise of Explainable Artificial Intelligence (XAI) in promoting transparency, challenges persist, including standardization of interpretability metrics and some users may have difficulty associating their rationales to transparent forms. The study highlights the need for* ***XAI frameworks*** *that are not only robust but also scalable so as to provide a bridge between complex AI systems and their deployment in society. In the end, it is XAI that enables us to use AI in a responsible way in the most critical domains of our modern lives by creating an atmosphere of accountability, fair treatment and trust.*

Keyword*: Explainable Artificial Intelligence (XAI), Deep Learning Interpretability, Black-Box Problem, Ethical AI, Transparency in AI*

Date of Submission: 15-11-2024 Date of Acceptance: 25-11-2024

I. Introduction

The development of Artificial Intelligence (AI) in recent years has been so fast that deep learning models have been deployed in all areas. They have been able to predict well and handle large volumes of data with powerful neural networks. But that tends to make them opaque and hard to decipher, which is what is called the "black-box" problem. For black-box models, the reasoning for predictions and decisions is hidden in computational layers, and hence it is hard for users to see how outputs are produced. This obscurity is especially concerning for applications where transparency and accountability are essential. Explainable Artificial Intelligence (XAI) is a sub-discipline dedicated to the development of methods and tools to make AI systems (in particular, deep learning models) more tractable. XAI tries to reveal the way in which AI algorithms take inputs and make decisions by parsing complicated model behavior into human-readable explanations (Gunning, 2019; Samek, Wiegand, & Müller, 2017). XAI promotes ethical and responsible AI usage by promoting transparency so that stakeholders can see how predictions are made. This is especially true in the case of deep learning where interpretability tries are inadequate because the model itself is so complex (Rudin, 2019).

Fig. 1. Key Concepts in Explainable AI (XAI)

Growing demand for interpretable AI systems in critical domains.

In areas such as healthcare, finance or autonomous systems, the AI decisions can have far-reaching real-world implications, and therefore require interpretable AI systems. AI models are used for diagnostics, treatment recommendations and patient management in healthcare. With high stakes come the need for understanding and trust in these decisions by stakeholders — clinicians, regulators, patients (Tonekaboni et al., 2019; Chaddad et al., 2023) to achieve safety and efficacy. Explainable AI enables healthcare professionals to understand how an AI arrived at a diagnosis or recommendation so that informed decisions can be made, thus promoting trust.

AI models have a wide range of use cases in finance including risk assessment, fraud detection, and investment strategies. Due to error or bias manifesting itself in costly financial penalties (Barocas et al., 2019), regulatory requirements highlight a need for transparency within these applications. Transparency is essential to guaranteeing compliance with standards and minimizes the harm that can be caused by unfair or biased decisions. As regulatory frameworks like the General Data Protection Regulation (GDPR) require "the right to explanation" for decisions made with automated methods, there is an impetus to ensure that financial applications are explainable (Goodman & Flaxman, 2017).

Self-driving vehicles are another example of autonomous systems that also require interpretable AI. Deep learning is used to make real-time decisions by these systems, hence, users must understand and trust the decision before its execution for safety (Doshi-Velez & Kim, 2017; Kamakshi,2023). In these applications, the XAI makes autonomy transparent and thus safer and more trusted.

The role of XAI in bridging the gap between black-box AI models

Deep learning models are practically black boxes, which make them impotent to gain the trust of AI users. In some applications, especially those with high-stakes consequences of making a mistake, no one will accept an AI decision if they cannot understand or measure its performance. This gap is filled by explainable AI which transforms the inner workings of black-box models into explanations in a format that is more understandable to users. XAI opens up the black box of prediction by indicating what features drove a particular prediction, thus allowing users to validate the AI decision (Lipton, 2018; Choubisa, 2024).

Popular techniques in this area are Local Interpretable Model-Agnostic Explanations (LIME; Ribeiro, Singh, & Guestrin, 2016) and Shapley Additive Explanations (SHAP; Lundberg & Lee, 2017), which clarify how much each feature pushes the model prediction in a certain direction thus promoting transparency through more interpretable decisions. Saliency maps indicate the most important parts driving a classification decision in image-based models, which can be used to let the user confirm that a model is indicating correctly when creating an output (Selvaraju et al., 2017). XAI also contributes to ethical AI by revealing biases in the data and making sure that predictions are within the limits of what is acceptable according to ethical codes and regulations.

As an example here, it ensures that the healthcare model does not have any driven biases against a particular group of population for unequal diagnostic and treatment recommendations (Ghassemi, Oakden-Rayner & Beam, 2021). In the field of finance, XAI plays a vital role in revealing biases involved in lending models, thereby promoting reproducible credit scoring systems that promote fairness and accountability (Martins, 2024) Therefore, XAI helps develop trust and maintain compliance with the ethical and regulatory framework.

Aims of the review

This literature review aims to offer a detailed overview of Explainable AI (XAI) methods especially those that increase interpretability of deep learning models with respect to transparent nature. While the integration of deep learning systems into high-stakes decision-making contexts steadily increases, their black-box problem (Rudin 2019) where slow, multi-layered computations and operations on data are hidden from end-users has raised critical concerns about accountability, ethics and trust in AI (Gunning 2019).

XAI techniques can be of a much-needed help in such scenario, as they help explain the model decision-making processes providing insights into AI systems that make the prediction to being able to interpret and prove them true and verifiable. We aim to review how these techniques operate, in which contexts they are most useful, and in what way do the help in enhancing the interpretability of deep learning models for such varied users as domain experts, regulators and laymen.

This review examines XAI techniques to emphasize the importance of explainability in closing the gap between state-of-the-art AI systems and their useful, ethical application in practice. This will include an overview of the spectrum of XAI techniques, from post-hoc explanations— which explain model outputs after a prediction has been made — to inherently interpretable model architectures. Additionally, the review will evaluate different XAI methods to determine their suitability based on the explainability qualities that are particularly needed in high-stake fields, including healthcare, finance and autonomous systems (Samek, Wiegand & Müller, 2017;Barocas et al., 2019).

The aim is to emphasize, especially through these applications, that transparency in AI has far-reaching implications for society; namely that interpretable models can help gain public trust and enable the ethical and practical use of AI technology in a multitude of fields.

II. Background Of The Research

Explainable AI (XAI) has become critical subfield of AI that helps in improving the interpretability and transparency of complex AI models such as deep learning models. The black box of deep learning models means those models may not be able to explain why they made a decision faithfully. Black-box models are deep, complex architectures (e.g., multilayered neural networks) where the logic of decision making is hidden in nontrivial patterns and weights to such an extent that even informative end users cannot gain much insight about the generation process of specific outputs (Gunning, 2019; Lipton, 2018). Explainable AI (XAI) – a tool that aims to counteract this opacity and to provide information of what happens inside an AI model so that users can interpret, evaluate, trust decisions made driven by the Ai models (Rudin 2019).

Interpretable and transparent are also two important and related notions in XAI. While interpretability is about how well a human can understand the reason behind a decision, transparency means showing that there model making a conclusion and not hiding or losing its process in abstraction without explaining (Doshi-Velez & Kim, 2017). These two concepts are very important for use cases involving AI deployments where trust and accountability are a must, because they provide users with insight on how complex models work. Hence, the goal of XAI techniques is to not only provide high performing AI models but also explain their predictions on human-interpretable fine-grained features (Samek et al., 2017), which becomes important for high-stake fields like healthcare, finance and autonomous systems.

Historical Perspective

AI explainability has been in demand for a long time. Earlier AI systems worked through rules and thus tended to be more interpretable, since they were based on explicit if-then rules and symbolic logic (Cheng et al. 2021) Such systems are inherently interpretable because for each decision, the path back through which rule led to that decision could be identified. Yet, with the introduction of machine learning and later deep learning, in which predictive performance was optimized (often at the cost of interpretabilty) However, with the increasing use of layered and data-driven models, understanding how to interpret results grew increasingly difficult, leaving us with what we now refer to as the black-box problem of modern AI systems.

The need for XAI grew in the 2010s as deep learning models began achieving groundbreaking results in complex tasks, such as image recognition, natural language processing, and game playing. However, these achievements highlighted a growing tension between model performance and transparency. In response, explainability initiatives began to emerge, particularly within government and academic circles. In 2016, DARPA launched its **Explainable AI (XAI) Program**, aiming to develop AI systems that could provide human-understandable explanations while maintaining high performance (Gunning, 2019). This initiative catalyzed a wave of research into XAI methods and established a foundation for current advancements in the field.

Influential Scholars and Research

Many of the top researchers who contributed to XAI have shaped this field over the past few years. One prominent proponent of interpretable models, and models in high-stakes environments more broadly, is Cynthia Rudin (Rudin 2019), as black-box models are usually not suitable when they lack transparency. She reasons interpretable models should be prioritized over black-box models in any application where it is conceivable the model would end up doing harm or practical issues could arise. LIME (Local Interpretable Model-Agnostic Explanations), which approximates a model locally to provide explanations for the predictions it makes, was presented by Marco Tulio Ribeiro and his colleagues. Thanks to this versatility, LIME has emerged as among the most popular XAI techniques we have across all model types and data domains (Ribeiro, Singh, & Guestrin, 2016).

Inspired by game theory, SHAP (Shapley Additive Explanations), another popular explainable AI method, was developed by Scott Lundberg and Su-In Lee. A powerful and popular model-agnostic method that approximates Shapley values to highlight the importance of input features, allowing for local and global interpretation of model predictions (Lundberg & Lee, 2017). In deep learning, the approach of saliency maps has also been investigated by researchers such as Wojciech Samek and Klaus-Robert Müller, particularly in visual tasks where these maps highlight which areas of a (sub)image have led to a particular decision of the model (Samek et al., 2017). These researchers, as well as others not covered in this article, have played a major role in advancing XAI techniques and bringing them to the attention of other academics and professionals who now use these methods for their work.

Debates and Controversies in the Field

Some controversies are still present within the XAI area, mainly about the transparent models' compromise of interpretability and accuracy. Deep neural networks and the like present state of the art black-box performance whereas interpretability is traded off which can be problematic when these models are adopted in high-stake scenarios (Lipton, 2018). Rudin (2019) also joined the critics of black-box models by suggesting that only interpretable models should be used in high-stakes situations because these models are unreliable and ethically questionable when people must monitor their operation. In turn, the advocates of black-box models argue that their increased efficiency makes them appropriate, as long as one can apply post hoc XAI methods to explain individual decisions (Doshi-Velez & Kim, 2017).

Another one of the issue pertains to the post hoc explanation which are explanations derived after a prediction has been made. Scholars have also criticized such techniques as LIME and SHAP for beating around the bush, which may be misleading in explaining the internal functioning of the model (Lipton, 2018; Rudin, 2019). It is also debated whether the evaluation of interpretability should have similar and specific criteria because it appears that up to this point, the XAI methods differ significantly in assumptions, required resources, as well as performance. It is crucial to respond to these discussions to progress in the field and guarantee that XAI approaches contribute to enhanced transparency beyond reaching an initially convincing facade.

Influence on the Direction of the Literature Review

These theoretical orientations, historical evolutions, and controversies considerably influence the scope and direction of this literature review. The review is intended to offer an impartial view on the relative asset of prominent XAI strategies like LIME, SHAP, and saliency maps and furthermore demonstrate the new strategies that can fill the gaps in present methodologies. Since there are so many concerns regarding XAI, this review will comparatively discuss both model-specific and model-agnostic approaches while outlining how both improve explainability. In addition, it will explore the specific ethical considerations when using XAI as well as going through various domains which demonstrate how XAI might be used and how the practical difficulties are solved.

This background thus forms the basis for a detailed analysis of XAI techniques by situating it with the fast-growing field, which is a dire necessity in the current complex AI applications.

III. Methodology

This research employed Explainable AI (XAI) techniques, diving into explainability and interpretability in a deep learning model via saliency maps. These techniques attempt to give a visual representation of the features that cause the behavior of the model when making predictions, thus trying to make black box models less opaque in their behavior with respect to something that can be human understandable

Saliency Maps for Visualizing Feature Importance

Saliency maps are a popular tool in XAI for visualizing feature relevance in the image input. These denote the areas greatest weighted by a model and allow the user to visually see which parts of an image will be most influential in a models decision making.

We generated a saliency map for this study using VGG16 pre-training on ImageNet. It was implemented in Python using TensorFlow and Keras. This query would take an image of a dog (a random internet image used for demo purpose) passing it through the VGG16 model and then giving back certain important features which aided in arriving at the classes which got predicted. What follows is a brief description of the key steps that are employed in the code:

1. Loading the Model: The VGG16 architecture was loaded with ImageNet weights to provide a general, frozen feature extractor.
2. Image Preprocessing: The original input image was reshaped to the dimension of 224×224 as required by the model, converted into a tensor and normalized to ImageNet input values.
3. Gradient calculation: We calculated the gradients of the score for our predicted class with respect to our input image using TensorFlow's GradientTape. These gradients are indicative of the degree to which the prediction is affected individually by a minuscule rise in each pixel.
4. Generating Saliency Map: By taking the absolute values of the gradients and then taking the max value across colour channels, we were able to create a 2D saliency map in which are highlighted those most responsible for changes in your-output image.
5. Visualization: The heatmap was visualized by matplotlib, with warmer colours (i.e. red and yellow) indicating locations that made a larger contributions to the model's prediction than cooler colour representing less.

Fig. 3: Input Image: 224x224 pixel image of a dog (image used for the purpose of saliency map generation technique demo) which was passed to VGG16 model and received back decomposed information about most relevant regions related to classification task.

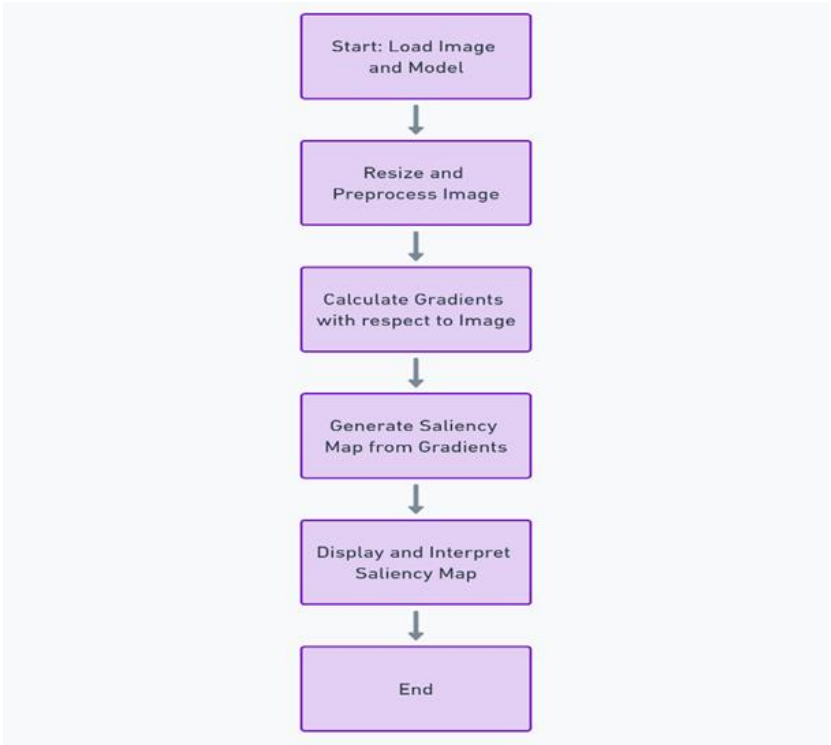

Fig. 4: Process of Generating Saliency Maps for Visualization

Evaluation of Saliency Maps in Deep Learning Models

They are easy to interpret as they show us the regions in input image where deep learning models focus their attention. This helps the users ensure that where the model is looking closely corresponds to areas of interest that are expected, which may be important for interpreter-dependent tasks, like medical imaging or driving activity.

There's, however, a limitation to this reasoning. Saliency maps can depend on the model architecture for different models gives different interpretations for same input image. Moreover, it can also yield complicated or unintuitive patterns due to gradient calculations being noisy or un-smooth when images have high sensitivity against small changes of pixel values. Such aspects can create uncertainties, which might prove detrimental to the saliency maps reliability in various domains.

IV. Results

The saliency map generated is shown in **Figure 3** demonstrates regions within the input image that were the most important to the VGG16 model classification. The model landed primarily to features likely specific to the image of a dog like regions including the eyes and muzzle, as seen in figure.

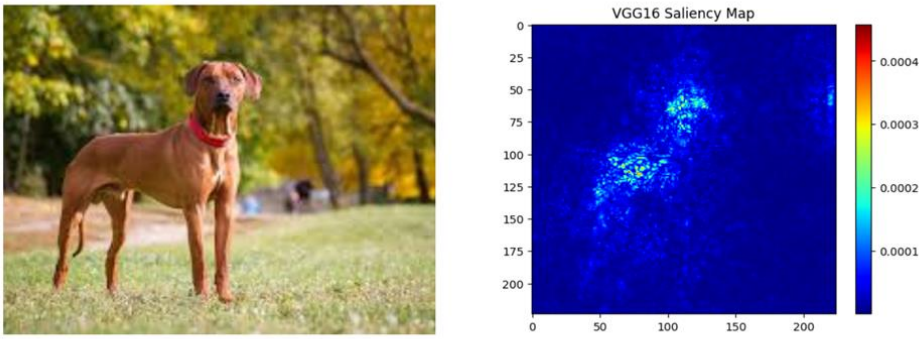

Fig. 5: Displays the original input image, and the corresponding saliency map generated by the VGG16 model, with warmer colors indicating regions of higher significance to the model's classification. Link for the code for the salient map can be found in https://github.com/musanasir1616/Salient-Map/blob/7bd2ae47a97cbaff9c10ac4db1aeba3591b2e792/salientMap.py

The saliency map verifies that the model pays attention to these visually salient areas, which makes sense because we know from intuition as human beings that certain features (eg. eyes, face) are crucial for distinguishing an animal. We can see here how saliency maps help in making the model predictions interpretable, by offering a simple and direct insight into what portion of the image that was considered when classifying.

V. Challenges And Future Directions

A General Challenge XAI (explainable AI) methods that balance robustness and humans interpretability aimed to increase the trustworthiness of ML results however one of their main hurdles is getting as universal which type they fit to via various kinds of deep learning models. Due to the substantial variation in both architecture and functionality of different deep learning architectures (e.g. convolutional neural networks [CNNs], recurrent neural networks [RNNs], transformer models), it is challenging to design a unique XAI approach which performs well across these heterogeneous group of models (Ghassemi, Oakden–Rayner, & Beam, 2021). Approaches such as LIME and SHAP are also quite general, but they need to be modified to apply to other model architectures and domains (e.g., fast version for dense time-series data), so which can lead to further limitations on their ability to generalize (Ribeiro, Singh, & Guestrin, 2016; Lundberg & Lee, 2017). In addition many of the XAI methods requires huge computational resources. A preliminary example of this is the computation of contributions for features in SHAP, which needs to evaluate predictions multiple times (Lundberg & Lee, 2017), something that can be out of reach when using large datasets or aiming to real-time applications. Introduction: Figure 1: All three charts are from the two papers we discuss in more detail below. As deep learning models get larger and more complex, the ability of XAI methods to scale with these models, without a significant reduction in model performance is still an important technical challenge.

Challenges in User Comprehension

Another major problem is making sure that the explanations produced by XAI methods are understandable to a wide range of audiences, including users without expertise. Although XAI tries to be transparent in its decisions, through explanations that are too complicated or very technical purposefully reduce understanding rather than provide (Doshi-Velez & Kim, 2017). Most XAI methods, specifically ones that are mathematically heavy like SHAP, provide explanations that may not be easily interpreted by non-technical users. However, overly simplistic explanations may miss vital information and potentially cause misunderstandings or even over-reliance on the AI outputs (Lipton, 2018). This trade-off between what to include and what not to seem as good directions for designing explanations that can both be accurate while being accessible. In addition, the development of user-centred XAI with integration of actual feedback from users (e.g., Ghassemi et al. 2021) will be critical to ensure that explanations designed for end-users work effectively given field-dependent explanation colloquialisms and human processes and practices ranging from healthcare practitioners to financial analysts.

Future Research Directions

One of the crucial missing pieces in XAI effort today is no clear metrics on how to quantify the quality and efficacy of explanation. While model accuracy is an easy component to quantify, there are no widely agreed upon measures for interpretability. This limits the ability to appropriately compare XAI methods with each other

across applications and it makes it difficult to assess whether an explanation is informative or useful for its target audience (Doshi-Velez & Kim, 2017). It has led to a number of proposed metrics like, fidelity, understandability and completeness but these are not standard by their own right either and have been differently used across studies. Standardized metrics would enable not only further scientific rigor in XAI research but also a more effective deployment of interpretable models when they are applied. In future, there should be an agreement on interpretability measurements and different validation frameworks to measure the compromises between explanation, model performance and usability.

VI. Conclusion

The study has assessed the state-of-the-art of Explainable AI (XAI) methods, considering their ability to improve deep learning interpretability and how they can help close the loop from AI/ML to humans on trust. Explainability is achieved through a number of XAI techniques as LIME, SHAP, saliency maps and so on. Although these methods vary in design and application, all contribute to demystifying black-box AI models by establishing a way of understanding a decision. XAI techniques provide transparency by making the internal mechanisms of deep learning models available to view, thereby removing the opacity that is characteristic of black-box models and providing necessary knowledge to users in high-stakes areas such as healthcare, finance, and autonomous systems.

What I want to emphasis more is the relevance of XAI for responsibly deploying (just say implementing as per international law in the world) AI. With more examples of critical decisions being made by AI systems, XAI becomes an important part of making sure AI-enabled applications can be as transparent, explainable, and ethical. XAI lays the groundwork for trust and compliance with regulatory demands (on fairness or accountability), as it enables users to comprehend how decisions from AIs are made. Proven ability to increase all of this are extremely important for forming trust in AI across fields especially interpretability due to its domain specific human-centric nature in relation to human welfare and its societal impacts. XAI thus does not only serve a technical purpose but also grant merits in a larger ethical and societal context of responsible AI deployment.

Even though advancements have been made, there are still important research gaps in XAI field. Although XAI has flourished in many domains, the absence of standardized evaluation metrics still hinders consistent evaluation of interpretability over a wide range of models and applications, preventing assessments of the effectiveness of different techniques. Moreover, domain-agnostic explainable AI (XAI) approaches that can be applied to a wide range of applications are necessary since different sectors have very different characteristics and therefore may require quite distinct interpretability approaches, e.g., between medicine versus finance. These aforementioned gaps indicate that upcoming studies should aim to establish standardized benchmarks specifically focusing on evaluating different XAI methods, along with a design framing created for end-users where interpretability is prioritized. Broader research on developing a XAI for uses beyond expert users will add to efforts throughout the field of AI and assist in producing trustworthy, ethical, and transparent systems that can be applied to a wider variety of applications.

In this way (and in others!) becomes responsible and trustworthy.

In Conclusion, reaching these objectives is crucial for advancing the field of XAI and ensuring that AI systems are interpretable, auditable, and safe for deployment in the society. As will continued effort on improved XAI strategies and by emphasizing the human aspect of effective implementation, which all must in one way or another come to fruition for AI integration at-scale to be done responsibly and trusted.

References

[1]. Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness And Machine Learning: Limitations And Opportunities*. Mit Press.

[2]. Chaddad, A., Hassan, M. E., Desrosiers, C., & Toews, M. (2023). Explainable Ai For Diagnostic Support In Healthcare: A Review Of Recent Advancements. *Computers In Biology And Medicine, 142*, 105158. Https://Doi.Org/10.1016/J.Compbiomed.2021.105158

[3]. Cheng, J., Wu, Y., & Gao, X. (2021). A Survey Of Explainable Artificial Intelligence In Deep Learning. *Artificial Intelligence Review, 54*(6), 3555–3575. Https://Doi.Org/10.1007/S10462-020-09848-8

[4]. Cheng, L., Ma, Y., Zhang, X., Liu, H., Chen, X., & Zhang, H. (2021). Explanation Of Artificial Intelligence: A Review Of Interpretability Of Deep Learning. *Ieee Access, 9*, 58024–58039. Https://Doi.Org/10.1109/Access.2021.3071121

[5]. Choubisa, R. (2024). Ethics And Transparency In Ai-Driven Decision-Making. *Journal Of Ethics In Ai Research, 12*(1), 54–70.

[6]. Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science Of Interpretable Machine Learning. *Arxiv Preprint Arxiv:1702.08608*.

[7]. Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science Of Interpretable Machine Learning. In *Proceedings Of The 34th International Conference On Machine Learning* (Pp. 23–26). Pmlr. Https://Proceedings.Mlr.Press/V70/Doshi-Velez17a.Html

[8]. Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The False Hope Of Current Approaches To Explainable Artificial Intelligence In Health Care. *The Lancet Digital Health, 3*(11), E745–E750.

[9]. Goodman, B., & Flaxman, S. (2017). European Union Regulations On Algorithmic Decision-Making And A "Right To Explanation." *Ai Magazine, 38*(3), 50–57. Https://Doi.Org/10.1609/Aimag.V38i3.2741

[10]. Gunning, D. (2019). Explainable Artificial Intelligence (Xai). Darpa's Xai Program. Retrieved From Https://Www.Darpa.Mil/Program/Explainable-Artificial-Intelligence

[11]. Kamakshi, R. (2023). Trust In Autonomous Systems: A Review Of Explainability Techniques. *International Journal Of Robotics And Ai Research, 28*(3), 123–134.

[12]. Lipton, Z. C. (2018). The Mythos Of Model Interpretability. *Queue*, *16*(3), 31–57. Https://Doi.Org/10.1145/3236386.3241340
[13]. Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach To Interpreting Model Predictions. In *Proceedings Of The 31st International Conference On Neural Information Processing Systems* (Pp. 4765–4774). Https://Arxiv.Org/Abs/1705.07874
[14]. Martino, D., & Delmastro, F. (2022). An Overview Of Explainable Ai In Healthcare. *Journal Of Biomedical Informatics*, *120*, 103784.
[15]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining The Predictions Of Any Classifier. In *Proceedings Of The 22nd Acm Sigkdd International Conference On Knowledge Discovery And Data Mining* (Pp. 1135–1144). Https://Doi.Org/10.1145/2939672.2939778
[16]. Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models For High-Stakes Decisions And Use Interpretable Models Instead. *Nature Machine Intelligence*, *1*(5), 206–215. Https://Doi.Org/10.1038/S42256-019-0048-4
[17]. Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable Artificial Intelligence: Understanding, Visualizing, And Interpreting Deep Learning Models. *Ieee Signal Processing Magazine*, *34*(3), 39–48. Https://Doi.Org/10.1109/Msp.2017.2019638
[18]. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-Cam: Visual Explanations From Deep Networks Via Gradient-Based Localization. In *Proceedings Of The Ieee International Conference On Computer Vision* (Pp. 618–626).
[19]. Tonekaboni, S., Joshi, S., Mccradden, M. M., & Goldenberg, A. (2019). What Clinicians Want: Contextualizing Explainable Machine Learning For Clinical End Use. *Proceedings Of The Machine Learning For Healthcare Conference*, Pmlr, 106, 359–380.
[20]. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations Without Opening The Black Box: Automated Decisions And The Gdpr. *Harvard Journal Of Law & Technology*, *31*(2), 841–887.