International Journal of Multidisciplinary and Current Educational Research (IJMCER)
ISSN: 2581-7027 ||Volume|| 6 ||Issue|| 3 ||Pages 226-235 ||2024||

Evaluating Methods for Assessing Interpretability of Deep Neural Networks (DNNs)

[1,]Emily Barnes, EdD, PhD, [2,]James Hutson, PhD
[1,]*Capitol Technology University*
https://orcid.org/0000-0001-9401-0186
[2,]*Lindenwood University*
https://orcid.org/0000-0002-0578-6052

ABSTRACT: The interpretability of deep neural networks (DNNs) is a critical focus in artificial intelligence (AI) and machine learning (ML), particularly as these models are increasingly deployed in high-stakes applications such as healthcare, finance, and autonomous systems. In the context of these technologies, interpretability refers to the extent to which a human can understand the cause of a decision made by a model. This article evaluates various methods for assessing the interpretability of DNNs, recognizing the significant challenges posed by their complex and opaque nature. The review encompasses both quantitative metrics and qualitative evaluations, aiming to identify effective strategies that enhance model transparency without compromising performance. The structure of the article includes an exploration of quantitative metrics, such as model complexity and computational requirements, and techniques like Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). It also considers qualitative evaluations, emphasizing the role of human judgment through domain expert reviews and user studies. Additionally, the article addresses the necessity of standardized benchmarks and the importance of context-specific evaluation frameworks. Through an investigation of these approaches, the treatment provides an overview of current methods and proposes future directions for improving the interpretability of DNNs, thus enhancing trust, accountability, and transparency in AI systems.

KEYWORDS: Deep neural networks, Interpretability, SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), Model transparency

I. INTRODUCTION

The study of Deep Neural Networks (DNNs) and their interpretability has garnered significant attention in recent years, primarily due to the increasing deployment of these models in high-stakes applications such as healthcare, finance, and autonomous systems. Interpretability, as defined by An and Joe (2022), refers to the extent to which a human can understand the cause of a decision made by a model. This concept is critical for ensuring trust, accountability, and transparency in AI systems. However, the complex and often opaque nature of DNNs poses substantial challenges to interpretability, making it a pressing issue for researchers and practitioners alike. The primary objective of research in this area is to evaluate and develop methods that enhance the transparency of DNNs without compromising their performance.

One significant aspect of this research involves evaluating quantitative metrics that can objectively measure interpretability. These metrics provide a framework for assessing the complexity, efficiency, and transparency of DNNs. Techniques such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) are instrumental in generating interpretable insights from complex models. By systematically examining these metrics and techniques, researchers aim to understand the trade-offs between model performance and interpretability, paving the way for the development of more transparent and accountable AI systems. These metrics include model complexity, computation requirements, and various explanation techniques that help in elucidating the decision-making processes of DNNs (Binder et al., 2016; Carvalho et al., 2019; Ma et al., 2015).

In addition to quantitative metrics, qualitative evaluations play a crucial role in assessing the interpretability of DNNs. These evaluations involve human judgment and insights that quantitative metrics alone cannot capture. Domain experts and end-users provide feedback on the clarity, relevance, and usefulness of the model's explanations. Methods such as user studies and expert reviews are employed to gather this feedback, ensuring that DNNs not only perform well technically but also meet practical and contextual needs. Qualitative evaluations help in understanding how well the model's explanations align with user needs and expectations,

Providing a comprehensive view of model interpretability. This approach complements quantitative metrics, offering a holistic perspective on the interpretability of DNNs (Loni et al., 2022; Mitsuhara et al., 2019).
Therefore, the primary objective of this article is to evaluate the various methods used to assess the interpretability of DNNs. At the same time, due to the subjective nature of interpretability, assessing it accurately can be challenging. This article thus aims to provide a review of both quantitative and qualitative methods proposed for this purpose. Through an evaluation of these methods, the paper seeks to identify effective strategies for making DNNs more interpretable without compromising their performance. In order to achieve this and to systematically address the evaluation of interpretability methods, the following is structured into three main sections:

1. **Quantitative Metrics:** This section explores objective measures of interpretability, such as model complexity, number of parameters, and computation requirements. It will also delve into techniques like LIME and SHAP, which provide explanations for individual predictions. Quantitative metrics provide a framework for assessing the complexity, efficiency, and transparency of DNNs. Techniques such as LIME and SHAP are instrumental in generating interpretable insights from complex models. By systematically examining these metrics and techniques, researchers aim to understand the trade-offs between model performance and interpretability, paving the way for the development of more transparent and accountable AI systems (Brocki & Chung, 2022; Couteaux et al., 2019; Montavon, Samek, & Müller, 2017).

2. **Qualitative Evaluations:** This section examines the role of human judgment in interpretability assessments. It will discuss how domain experts use user studies and expert reviews to evaluate the clarity and usefulness of model explanations. Qualitative evaluations involve human judgment and insights that quantitative metrics alone cannot capture. Domain experts and end-users provide feedback on the clarity, relevance, and usefulness of the model's explanations. Methods such as user studies and expert reviews are employed to gather this feedback, ensuring that DNNs not only perform well technically but also meet practical and contextual needs. Qualitative evaluations help in understanding how well the model's explanations align with user needs and expectations, providing a comprehensive view of model interpretability (Dong et al., 2017; Zhang et al., 2020).

3. **Benchmarking:** This section addresses the challenges of creating standardized benchmarks for interpretability. It highlights the need for context-specific evaluation frameworks and reviews existing benchmark frameworks to identify their strengths and limitations. Standardized benchmarks play a critical role in assessing the interpretability of DNNs. Developing effective benchmarks requires overcoming significant challenges and ensuring that evaluations are context-specific and multi-dimensional. Future research and collaborative efforts are necessary to create benchmarks that provide accurate, comprehensive, and meaningful assessments of model interpretability (Casper et al., 2023; Chakraborty et al., 2017). Through these considerations the paper finds its usefulness in the evaluation of methods for assessing the interpretability of DNNs, which is a critical aspect for the practical application of these models in high-stakes domains. Through the systematic review of both quantitative and qualitative methods, the paper aims to provide a balanced perspective on the strengths and limitations of various interpretability techniques. This approach not only highlights the current state of the field but also identifies areas where further research and development are needed to enhance the transparency and accountability of DNNs. The results of this study are expected to guide future efforts in making deep learning models more understandable and trustworthy, thereby facilitating their broader acceptance and integration into mission-critical applications such as healthcare, finance, and autonomous systems (Chakraborty et al., 2017; Zhang et al., 2020.

II. QUANTITATIVE METRICS :

Evaluating the interpretability of DNNs necessitates a thorough understanding of quantitative metrics that can objectively measure various aspects of these models. Quantitative metrics provide a framework for assessing the complexity, efficiency, and transparency of DNNs, offering insights into how these models function and how their decisions can be interpreted. This section delves into key quantitative metrics, including model complexity and computation requirements, and explores explanation techniques like LIME and SHAP, which are instrumental in generating interpretable insights from complex models (Binder et al., 2016; Carvalho et al., 2019; Ma et al., 2015). By systematically examining these metrics and techniques, we can better understand the trade-offs between model performance and interpretability, paving the way for the development of more transparent and accountable AI systems.

Model complexity refers to the intricacy of the structure of neural networks, including the number of parameters, layers, and the overall depth of the network. Higher complexity, as described by Zhao and Gao (2021), often correlates with an increased capacity to learn intricate patterns from data but can also lead to decreased interpretability. Simpler models are generally easier to understand and interpret, whereas more complex models may function as "black boxes," making it challenging to discern how decisions are made. Model complexity is crucial because it directly affects both the performance and interpretability of the networks (Chakraborty et al., 2017; Zhang et al., 2020).

Several metrics can be used to quantify model complexity. One such metric is the number of parameters, which refers to the total count of adjustable weights within the network. Generally, a higher number of parameters indicates a more complex model, which can learn more detailed patterns but may also become less interpretable (Lee et al., 2020). Another metric is the depth of the network, defined by the number of layers in the network. Deeper networks have a greater capacity to learn abstract representations from data, but they are also harder to interpret due to their intricate internal structures (He & Papakonstantinou, 2022).

- **Number of Parameters**: The total count of adjustable weights within the network. More parameters typically indicate a more complex model.
- **Depth of the Network**: The number of layers in the network. Deeper networks can learn more abstract representations but are harder to interpret.

Thus, while higher model complexity can enhance the ability of neural networks to learn from data, it often comes at the cost of interpretability. Balancing this trade-off is critical for developing models that are both effective and understandable. Researchers must carefully consider the number of parameters and the depth of the network when designing DNNs to ensure that the models remain interpretable while achieving high performance (Turner et al., 2017), (Hu et al., 2020).

Computational efficiency is a critical factor in the interpretability of DNNs. Efficient models are generally faster and less resource-intensive, which can facilitate more straightforward debugging and understanding of the behavior of the model. High computational requirements can obfuscate the operations of the model, making it difficult to interpret how specific inputs lead to particular outputs (Hanif et al., 2021; Messud & Chambefort, 2020). Efficient models reduce the computational overhead, which can make the models easier to analyze and interpret. Conversely, models with high computational demands may complicate the interpretability process by adding layers of complexity that obscure the inner workings of the model.

Several techniques are employed to measure computation requirements **(Table 1)**. One technique is measuring inference time, which is the time it takes for a model to process an input and produce an output. This metric is crucial as it directly impacts the usability of the model in real-time applications. Another important metric is memory usage, which indicates the amount of memory required during model training and inference. High memory usage can be a limiting factor for deploying models in resource-constrained environments. Additionally, Floating Point Operations Per Second (FLOPS) measures the number of operations required to make predictions, indicating the computational workload of the model. These metrics collectively help in understanding and managing the computational demands of DNNs (Lin et al., 2019).

Table 1. Techniques for Measuring Computation Requirements

Technique	Description
Inference Time	The time it takes for a model to process an input and produce an output.
Memory Usage	The amount of memory required during model training and inference.
Floating Point Operations Per Second (FLOPS)	A measure of the number of operations required to make predictions, indicating the computational workload.

Explanation techniques such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) are designed to enhance the interpretability of DNNs by providing insights into individual predictions **(Table 2)**. LIME approximates the original model locally with an interpretable model by perturbing the input data around the instance to be explained and observing the changes in predictions. This

approach creates a simpler model that mimics the behavior of the complex model around a particular instance (Lundberg & Lee, 2017).

Table 2. LIME Explanation Generation Steps

Step	Description
Perturbing the input data	Creating a new dataset of similar instances by making small changes to the input data.
Predicting outcomes	Using the original model to predict outcomes for these perturbed instances.
Training an interpretable model	Training an interpretable model (e.g., a linear model) on the new dataset to approximate the decision boundary of the original model locally.
Presenting the coefficients	Highlighting the contributions of different features to the prediction by presenting the coefficients of the interpretable model as explanations.

SHAP, on the other hand, uses cooperative game theory to provide a unified framework for interpreting model predictions **(Table 3)**. SHAP values represent the contribution of each feature to the prediction, ensuring a fair distribution of the prediction among the features. This technique helps in understanding the effect of each feature on the model's output and visualizes these contributions using tools like SHAP plots (Couteaux et al., 2019; Montavon et al., 2017).

Table 3. SHAP Values for Interpretability

Value	Description
Calculation	SHAP values are computed by considering all possible combinations of features and their marginal contributions.
Interpretation	These values explain the effect of each feature, showing how much they increase or decrease the prediction.
Visualization	Tools like SHAP plots can visualize these contributions, making it easier to understand the model's behavior across different instances.

Quantitative metrics, therefore, play a crucial role in assessing the interpretability of DNNs. Key metrics such as model complexity and computational efficiency provide valuable insights into how these models operate and can be interpreted. Model complexity, quantified by the number of parameters and the depth of the network, influences both the learning capacity and the interpretability of the model. Simpler models are easier to interpret but may lack the capacity to learn complex patterns, whereas more complex models can learn intricate patterns but often function as "black boxes." Computational efficiency, measured through metrics like inference time, memory usage, and FLOPS, impacts the ease with which models can be debugged and understood. Efficient models facilitate straightforward interpretation, while high computational demands can obscure the model's operations. Techniques such as LIME and SHAP further enhance interpretability by providing clear, understandable explanations for individual predictions. By leveraging these quantitative metrics and techniques, researchers and practitioners can develop more transparent and accountable AI systems, balancing the trade-offs between performance and interpretability.

III. QUALITATIVE EVALUATIONS :

Qualitative evaluations are essential for assessing the interpretability of DNNs as they incorporate human judgment and insights that quantitative metrics alone cannot capture. These evaluations involve domain experts and end-users who provide feedback on the clarity, relevance, and usefulness of the model's explanations. According to Loni et al. (2022), by focusing on real-world applicability and human understanding, qualitative evaluations ensure that DNNs not only perform well technically but also meet practical and contextual needs. This section explores the role of human judgment, the methodology of user studies, and the impact of expert reviews in assessing and enhancing the interpretability of DNNs. By involving human perspectives, these evaluations help bridge the gap between technical performance and practical utility, ensuring that AI models are not only effective but also trusted and understood by their users.

Human judgment plays a crucial role in assessing the interpretability of deep neural networks (DNNs). Domain experts bring invaluable insights and contextual understanding that purely quantitative methods cannot capture. Their evaluations are essential for determining whether the explanations of a model are meaningful, accurate, and actionable within the specific domain of application. The approach ensures that the decisions of the model are not only technically sound but also practically relevant (Mitsuhara et al., 2019). Experts typically use several

criteria to evaluate the interpretability of model explanations, including clarity, relevance, actionability, consistency, and comprehensiveness **(Table 4)**. These criteria help to ensure that explanations are not only understandable but also useful and reliable in guiding decision-making processes. By leveraging the expertise of domain specialists, qualitative evaluations provide a nuanced assessment of model interpretability that is closely aligned with real-world applications and requirements.

Table 4. Criteria Used by Experts to Evaluate Model Explanations

Criteria	Description
Clarity	How easily can the explanation be understood?
Relevance	Does the explanation focus on the most critical features or factors?
Actionability	Can the explanation guide decision-making or suggest practical actions?
Consistency	Are the explanations consistent across similar instances?
Comprehensiveness	Does the explanation cover all relevant aspects of the decision process?

User studies involve systematic investigations where end-users interact with the explanations of models. These studies are designed to gather feedback on how well users understand the model's decisions and whether the explanations meet their needs. The methodology typically includes participant selection, task design, data collection, and analysis **(Table 5)**. Participant selection involves choosing a representative sample of users, often including both domain experts and laypersons, to ensure diverse perspectives (Doshi-Velez & Kim, 2017). Task design entails creating tasks that require users to interpret model outputs and explanations. Data collection uses surveys, interviews, and observational methods to gather qualitative data on user experiences. Analysis involves identifying common themes, usability issues, and areas for improvement from the collected data. Several user studies have been conducted to evaluate the interpretability of DNNs, providing insights into how different explanation techniques impact users' understanding and trust in the model. These studies highlight the importance of user feedback in refining and improving model explanations to better serve end-user needs.

Table 5. Methodology of Conducting User Studies for Interpretability Assessment

Methodology	Description
Participant Selection	Choosing a representative sample of users, often including both domain experts and laypersons.
Task Design	Creating tasks that require users to interpret model outputs and explanations.
Data Collection	Using surveys, interviews, and observational methods to gather qualitative data on user experiences.
Analysis	Analyzing the data to identify common themes, usability issues, and areas for improvement.

Expert reviews involve soliciting feedback from domain experts who critically evaluate the explanations of models. This process usually includes initial reviews, iterative refinement, and validation **(Table 6)**. During the initial review, experts provide preliminary feedback on the explanations based on their knowledge and experience. Iterative refinement involves refining the model and its explanations based on expert feedback in multiple cycles until the explanations meet the desired standards of clarity and relevance. Validation is the final step where experts validate the refined model explanations to ensure they are accurate and reliable (Huang et al., 2019). Expert feedback is invaluable for improving the interpretability of DNNs as it helps identify flaws or gaps in the explanations the model provides and suggests ways to enhance them. This iterative process leads to the development of models that not only perform well but also provide transparent and understandable explanations. Incorporating expert reviews ensures that the model meets the high standards required for practical deployment in real-world scenarios, thus enhancing the overall trustworthiness and usability of AI systems. By combining human judgment through domain expert evaluations, user studies, and expert reviews, qualitative evaluations provide a comprehensive understanding of how well the explanations align with user needs and expectations. This approach complements quantitative metrics, offering a holistic view of model interpretability.

These qualitative insights are crucial for developing AI systems that are both effective and trustworthy, ensuring that models can be confidently used in various high-stakes applications.

Table 5. Process of Expert Reviews in Interpretability Assessment

Process	Description
Initial Review	Experts provide preliminary feedback on the model's explanations based on their knowledge and experience.
Iterative Refinement	The model and its explanations are refined iteratively based on expert feedback.
Validation	Experts validate the final model explanations to ensure they are accurate and reliable.

IV. BENCHMARKING

Benchmarking is an essential process for evaluating the interpretability of DNNs. It involves establishing standardized measures to consistently assess how understandable and transparent these models are. Effective benchmarking provides a means to compare different models and interpretability methods, ensuring that advancements in these networks do not come at the cost of losing insight into their decision-making processes. Thus, this section explores the necessity and challenges of developing standardized benchmarks, the importance of context-specific benchmarks, and reviews existing frameworks. Additionally, it proposes future directions to improve benchmarking practices, emphasizing the need for comprehensive, multi-dimensional evaluation frameworks that can adapt to various domains and applications (Casper et al., 2023; Zhang et al., 2020).

As interpretability becomes increasingly important, understanding how humans can understand the reasoning behind models needs benchmarking (Farahani et al., 2022; Hanif et al., 2021; Naseer et al., 2023). This understanding is crucial for ensuring trust, accountability, and transparency. However, assessing interpretability poses significant challenges due to its inherently subjective nature and the diversity of applications. Different domains require different types of explanations, making it difficult to develop one-size-fits-all benchmarks. Furthermore, the lack of consensus on what constitutes "good" interpretability adds another layer of complexity.

Challenges in developing standardized benchmarks for interpretability include the subjective nature of interpretability and the diversity of applications (Moutin & Davel, 2022). Different domains require different types of explanations, making it difficult to develop one-size-fits-all benchmarks. Additionally, the lack of consensus on what constitutes "good" interpretability adds another layer of complexity (Farahani et al., 2022). These challenges highlight the need for flexible and adaptable benchmarks that can be tailored to specific contexts and use cases.
Context-specific benchmarks are crucial because they ensure that interpretability assessments are relevant and meaningful within particular domains. For example, interpretability requirements in healthcare differ significantly from those in financial services. Benchmarks must account for these variations to provide accurate and useful evaluations. Context-specific benchmarks allow for a more precise measurement of how well model explanations meet the needs and expectations of end-users in different fields (Zhang & Zhu, 2018).

Several benchmarks have been proposed to assess the interpretability of DNNs. Dong et al. (2017) discusses various criteria and frameworks used to evaluate model interpretability, such as simplicity, clarity, and completeness. These benchmarks aim to provide a standardized way to compare different models and explanation methods (Dong et al., 2017). Yet, while existing benchmarks offer valuable insights, they also have limitations. Many are too generic and do not account for the specific needs of different domains. Also, they often focus on a single aspect of interpretability, such as clarity or simplicity, without considering the broader context (Carvalho et al., 2019). This narrow focus can lead to incomplete assessments.

There is a need for more comprehensive benchmarks that consider multiple dimensions of interpretability and are adaptable to various contexts. To address the limitations of current benchmarks, future benchmarks should be designed with flexibility and multi-dimensionality in mind (Messud & Chambefort, 2020). They should incorporate criteria that reflect the specific needs of different domains and types of users. Benchmarks should also be developed collaboratively with input from a diverse range of stakeholders, including domain experts, data scientists, and end-users. This collaborative approach can help ensure that benchmarks are comprehensive and relevant.

Multi-dimensional evaluation frameworks are essential for a thorough assessment of interpretability. These frameworks should evaluate multiple aspects of model explanations, such as clarity, relevance, consistency, and actionability (Carvalho et al., 2019; Casper et al., 2023). Looking at a wide range of factors, multi-dimensional frameworks provide a more holistic view of interpretability, helping to identify strengths and weaknesses in model explanations. This comprehensive approach can guide the development of more interpretable and user-friendly DNNs. Standardized benchmarks, moreover, play a critical role in assessing the interpretability of DNNs. Developing effective benchmarks requires overcoming significant challenges and ensuring that evaluations are context-specific and multi-dimensional. Future research and collaborative efforts are necessary to create benchmarks that provide accurate, comprehensive, and meaningful assessments of model interpretability.

V. CASE STUDIES

Benchmarking is a critical component in evaluating the interpretability of DNNs and stems from the necessity to compare different models and interpretability methods consistently. However, developing such benchmarks has historically presented significant challenges due to the inherently subjective nature of interpretability and the diverse applications of these networks across various domains (Bochie et al., 2021; Rasheed et al., 2022). Examples of earlier attempts will follow, examining the complexities and importance of standardized and context-specific benchmarks. These examples will review existing frameworks and propose future directions for effective benchmarking.

For instance, Jacobs (2020) introduced "Sensie," a method for probing the sensitivity of neural networks. This approach leverages well-known techniques for visualizing and interpreting the outputs of DNNs, such as inspecting learned features (e.g., convolutional kernels and activation maps) and producing saliency maps. These visual tools allow researchers to pinpoint which parts of an input (such as an image) play the most crucial role in the model's decision-making process. Methods like occluding parts of an image (Zeiler & Fergus, 2014) and Layer-wise Relevance Propagation (Binder et al., 2016) are commonly used to enhance the interpretability of the networks.

On the other hand, SmoothGrad, developed by Smilkov et al. (2017), is an algorithm designed to interpret the outputs of deep neural networks. This method enhances the clarity of saliency maps by reducing visual noise, making it easier for researchers to identify important features within input data. The SmoothGrad algorithm has been particularly useful in computer vision applications, where understanding which parts of an image influence a model's prediction is crucial. Along the same lines, Rao et al. (2023) discusses the SHAP (SHapley Additive exPlanations) method in a study which was evaluated for its interpretability using Sensor Information Technology (SensIT) data. The research focused on explaining the decision-making process of a black-box model classifying two categories of military vehicles. The evaluation highlighted the effectiveness of SHAP values in providing clear, consistent, and stable explanations. Both Tree SHAP and Sampling SHAP were assessed in terms of consistency, compactness, stability, and approximation. The results demonstrated that SHAP could elucidate the underlying mechanisms of complex models, ensuring both accuracy and comprehensibility, which are essential for deploying ML models in security-sensitive sectors.

These case studies illustrate various approaches and methodologies for enhancing the interpretability of deep neural networks. From probing sensitivity with "Sensie" to generating clear visual explanations with SmoothGrad and evaluating model transparency with SHAP, these efforts contribute to the broader goal of making AI models more understandable and trustworthy across diverse applications. Each method brings unique strengths to the table, addressing different aspects of interpretability and providing valuable insights into the inner workings of DNNs.

VI. CONCLUSION

Integrating quantitative and qualitative methods offers a comprehensive approach to assessing the interpretability of deep neural networks (DNNs). Quantitative metrics provide objective measures, while qualitative evaluations incorporate human insights, ensuring that models are both technically sound and practically relevant. Strategies for integrating these methods include using quantitative metrics to identify areas needing improvement and then applying qualitative evaluations to refine and validate these enhancements (Montavon, Samek, & Müller, 2017), (Brocki & Chung, 2022). This combination of approaches helps create a more holistic and thorough understanding of model behavior and interpretability.

Assessing interpretability presents several challenges, including the subjective nature of interpretability, the diversity of applications, and the lack of standardized benchmarks. Potential solutions include developing flexible, context-specific benchmarks and fostering interdisciplinary collaboration to ensure comprehensive evaluation frameworks. By addressing these challenges, researchers can create more robust methods for evaluating interpretability that are adaptable to various contexts and use cases (Casper et al., 2023; Moutin & Davel, 2022). Ongoing research in interpretability assessment is crucial for advancing the field of AI. Future research should focus on developing multi-dimensional evaluation frameworks, exploring new explanation techniques, and addressing the challenges of interpretability in different domains. Emerging trends include the increasing use of hybrid models and the integration of interpretability assessments into the model development lifecycle, ensuring that interpretability is considered from the outset (Koo et al., 2021). Such research efforts are essential for creating AI models that are both powerful and understandable, facilitating their use in critical and high-stakes applications where transparency and trust are paramount.

Looking ahead, combining quantitative and qualitative methods provides a comprehensive framework for evaluating the interpretability of DNNs. Addressing the challenges associated with interpretability assessment and developing robust, multi-dimensional benchmarks are crucial steps toward enhancing the transparency and trustworthiness of AI models. Continued research and interdisciplinary collaboration will be vital in advancing these efforts, ultimately contributing to the development of more interpretable and reliable AI systems.

REFERENCES

1. An, J., & Joe, I. (2022). Attention map-guided visual explanations for deep neural networks. *Applied Sciences*. https://doi.org/10.3390/app12083846
2. Binder, A., Montavon, G., Lapuschkin, S., Müller, K., & Samek, W. (2016). Layer-wise relevance propagation for neural networks with local renormalization layers. *ArXiv*, abs/1604.00825.
3. Bianchini, M., & Scarselli, F. (2014). On the complexity of shallow and deep neural network classifiers. *The European Symposium on Artificial Neural Networks*.
4. Bochie, K., Gilbert, M. S., Gantert, L., Barbosa, M. S., Medeiros, D. S., & Campista, M. E. M. (2021). A survey on deep learning for challenged networks: Applications and trends. *Journal of Network and Computer Applications, 194*, 103213.
5. Brocki, L., & Chung, N. (2022). Evaluation of interpretability methods and perturbation artifacts in Deep Neural Networks. *ArXiv, abs/2203.02928*. https://doi.org/10.48550/arXiv.2203.02928
6. Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*. doi:10.3390/electronics8080832
7. Casper, S., Bu, T., Li, Y., Li, J., Zhang, K., Hariharan, K., & Hadfield-Menell, D. (2023). Red teaming deep neural networks with feature synthesis tools. *Advances in Neural Information Processing Systems, 36*, 80470-80516.
8. Chakraborty, S., Tomsett, R., Raghavendra, R., Harborne, D., Alzantot, M., Cerutti, F., Srivastava, M., Preece, A., Julier, S., Rao, R., Kelley, T., Braines, D., Sensoy, M., Willis, C., & Gurram, P. (2017). Interpretability of deep learning models: A survey of results. *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation* (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), 1-6. https://doi.org/10.1109/UIC-ATC.2017.8397411
9. Chaudhuri, A. (2019). Some insights and observations on depth issues in deep learning networks. *International Work-Conference on Artificial and Natural Neural Networks*. https://doi.org/10.1007/978-3-030-20521-8_48
10. Chitty-Venkata, K. T., & Somani, A. K. (2022). Neural architecture search survey: A hardware perspective. *ACM Computing Surveys*, 55, 1-36. https://doi.org/10.1145/3524500
11. Couteaux, V., Nempont, O., Pizaine, G., & Bloch, I. (2019). Towards interpretability of segmentation networks by analyzing deepdreams. In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support: Second International Workshop, iMIMIC 2019, and 9th International Workshop, ML-CDS 2019, Held in Conjunction with MICCAI 2019*, Shenzhen, China, October 17, 2019, Proceedings 9 (pp. 56-63). Springer International Publishing.
12. Daniels, Z. A., Frank, L., Menart, C., Raymer, M., & Hitzler, P. (2020). A framework for explainable deep neural models using external knowledge graphs. *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications II*. https://doi.org/10.1117/12.2558083

13. Dong, Y., Su, H., Zhu, J., & Zhang, B. (2017). Improving interpretability of deep neural networks with semantic information. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 975-983. https://doi.org/10.1109/CVPR.2017.110
14. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608.*
15. Farahani, F. V., Fiok, K., Lahijanian, B., Karwowski, W., & Douglas, P. K. (2022). Explainable AI: A review of applications to neuroimaging data. *Frontiers in Neuroscience*, 16. https://doi.org/10.3389/fnins.2022.906290
16. Hanif, A., Zhang, X., & Wood, S. (2021). A survey on explainable artificial intelligence techniques and challenges. *2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW)*, 81-89. https://doi.org/10.1109/UPCON59197.2023.10434608
17. He, S., & Papakonstantinou, P. (2022). Deep Neural Networks: The Missing Complexity Parameter. Electron. *Colloquium Comput. Complex.*, TR22.
18. Hu, X., Liu, W., Bian, J., & Pei, J. (2020, August). Measuring model complexity of neural networks with curve activation functions. In *Proceedings of the 26th ACM SIGKDD International Conference on knowledge discovery & data mining* (pp. 1521-1531). https://doi.org/10.1145/3394486.3403203
19. Huang, S., Xu, X., Zheng, L., & Wornell, G. W. (2019). An information theoretic interpretation to deep neural networks. *Entropy*, 24. https://doi.org/10.3390/e24010135
20. Jacobs, C. (2020). "Sensie": Probing the sensitivity of neural networks. *J. Open Source Softw.*, 5, 2180. https://doi.org/10.21105/joss.02180
21. Koo, P., Majdandzic, A., Ploenzke, M., Anand, P., & Paul, S. (2021). Global importance analysis: An interpretability method to quantify importance of genomic features in deep neural networks. *PLoS Computational Biology*, 17. https://doi.org/10.1371/journal.pcbi.1008925
22. Kumar, P., & Sharma, M. (2020). Predicting academic performance of international students using machine learning techniques and human interpretable explanations using LIME—Case study of an Indian university. https://doi.org/10.1007/978-981-15-1286-5_25
23. Lee, Y., Lee, J., Hwang, S. J., Yang, E., & Choi, S. (2020). Neural complexity measures. *Advances in Neural Information Processing Systems*, *33*, 9713-9724.
24. Liang, Y., Li, S., Yan, C., Li, M., & Jiang, C. (2021). Explaining the black-box model: A survey of local interpretation methods for deep neural networks. *Neurocomputing*, 419, 168-182. https://doi.org/10.1016/j.neucom.2020.08.011
25. Lin, Z. Q., Shafiee, M. J., Bochkarev, S., Jules, M. S., Wang, X. Y., & Wong, A. (2019). Do explanations reflect decisions? A machine-centric strategy to quantify the performance of explainability algorithms. *arXiv preprint arXiv:1910.07387*.
26. Loni, M., Sinaei, S., Zoljodi, A., Daneshtalab, M., & Sjödin, M. (2020). DeepMaker: A multi-objective optimization framework for deep neural networks in embedded systems. *Microprocess. Microsystems*, 73, 102989. https://doi.org/10.1016/j.micpro.2020.102989
27. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
28. Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E., & Svetnik, V. (2015). Deep neural nets as a method for quantitative structure-activity relationships. *Journal of Chemical Information and Modeling*, 55(2), 263-274. https://doi.org/10.1021/ci500747n
29. Messud, J., & Chambefort, M. (2020). Understanding how a deep neural network architecture choice can be related to a seismic processing task. https://doi.org/10.3997/2214-4609.202032076
30. Mitsuhara, M., Fukui, H., Sakashita, Y., Ogata, T., Hirakawa, T., Yamashita, T., & Fujiyoshi, H. (2019). Embedding human knowledge into deep neural network via attention map. *VISIGRAPP*. https://doi.org/10.5220/0010335806260636
31. Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digital signal processing*, *73*, 1-15. https://doi.org/10.1016/j.dsp.2017.10.011
32. Montúfar, G., Pascanu, R., Cho, K., & Bengio, Y. (2014). On the number of linear regions of deep neural networks. *Neural Information Processing Systems*.
33. Mouton, C., & Davel, M. H. (2022). Exploring layerwise decision making in DNNs. *SACAIR*. https://doi.org/10.1007/978-3-030-95070-5_10
34. Naseer, M., Hasan, O., & Shafique, M. (2023). QuanDA: GPU accelerated quantitative deep neural network analysis. *ACM Transactions on Design Automation of Electronic Systems*, 28, 1-21. https://doi.org/10.1145/3611671
35. Rao, T., Agarwal, S., & Singh, N. (2023). An empirical evaluation of Shapley additive explanations: A military implication. *2023 10th IEEE Uttar Pradesh Section International Conference on Electrical,*

Electronics and Computer Engineering (UPCON), 10, 1390-1397. https://doi.org/10.1109/UPCON59197.2023.10434608

36. Rasheed, K., Qayyum, A., Ghaly, M., Al-Fuqaha, A., Razi, A., & Qadir, J. (2022). Explainable, trustworthy, and ethical machine learning for healthcare: A survey. *Computers in Biology and Medicine, 149*, 106043.
37. Smilkov, D., Thorat, N., Kim, B., Viégas, F. B., & Wattenberg, M. (2017). SmoothGrad: Removing noise by adding noise. *ArXiv*, abs/1706.03825. https://doi.org/10.48550/arXiv.1706.03825
38. Turner, B., Miletić, S., & Forstmann, B. (2017). Outlook on deep neural networks in computational cognitive neuroscience. *NeuroImage, 180*, 117-118. https://doi.org/10.1016/j.neuroimage.2017.12.078
39. Yu, D., Xiong, W., Droppo, J., Stolcke, A., Ye, G., Li, J., & Zweig, G. (2016). Deep convolutional neural networks with layer-wise context expansion and attention. *Interspeech*. https://doi.org/10.21437/Interspeech.2016-251
40. Yu, H. (2010). Network complexity analysis of multilayer feedforward artificial neural networks. *Applications of Neural Networks in High Assurance Systems*. https://doi.org/10.1007/978-3-642-10690-3_3
41. Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13* (pp. 818-833). Springer International Publishing.
42. Zhang, Y., Tiňo, P., Leonardis, A., & Tang, K. (2020). A Survey on Neural Network Interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence, 5*, 726-742. https://doi.org/10.1109/TETCI.2021.3100641
43. Zhang, W., Zhang, L., Zhang, Z., & Sun, M. (2021). IBD: The metrics and evaluation method for DNN processor benchmark while doing inference task. *J. Intell. Fuzzy Syst.*, 40, 9949-9961. https://doi.org/10.3233/JIFS-202552
44. Zhang, Q., & Zhu, S. (2018). Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering, 19*, 27 - 39. https://doi.org/10.1631/FITEE.1700808
45. Zhao, C., & Gao, X. (2021). QDNN: Deep neural networks with quantum layers. *Quantum Machine Intelligence*, 3(15). 10.1007/s42484-021-00046-w