International Journal of Multidisciplinary and Current Educational Research (IJMCER)
ISSN: 2581-7027 ||Volume|| 6 ||Issue|| 3 ||Pages 218-225 ||2024||

Architectural Elements Contributing to 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) has become a critical focus in artificial intelligence and machine learning, particularly as DNNs are increasingly used in high-stakes applications like healthcare, finance, and autonomous driving. Interpretability refers to the extent to which humans can understand the reasons behind a model's decisions, which is essential for trust, accountability, and transparency. However, the complexity and depth of DNN architectures often compromise interpretability as these models function as "black boxes." This article reviews key architectural elements of DNNs that affect their interpretability, aiming to guide the design of more transparent and trustworthy models. The primary objective is to examine different layer types, network depth and complexity, connectivity patterns, and attention mechanisms to provide a comprehensive understanding of how these architectural components can enhance the interpretability of DNNs. Layer types, including convolutional, recurrent, and attention layers, have unique properties affecting model interpretability. Convolutional layers are fundamental for image recognition tasks, recurrent layers handle sequential data, and attention layers improve model performance by selectively focusing on relevant parts of the input. Network depth and complexity significantly impact interpretability, with shallow networks being more interpretable but less powerful on complex tasks compared to deep networks. Connectivity patterns also play a crucial role; while fully connected layers offer high flexibility, they pose interpretability challenges due to dense connections, whereas structured connectivity patterns like residual networks provide clearer information flow. Attention mechanisms further enhance interpretability by dynamically highlighting the most relevant parts of the input data. This review helps researchers and practitioners identify strategies to design DNNs that are both performant and interpretable, contributing to the advancement of trustworthy AI applications.

KEYWORDS: Interpretability, Deep neural networks, Architectural elements, Trustworthy AI, Attention mechanisms

I. INTRODUCTION

The interpretability of Deep Neural Networks (DNNs) has emerged as a critical area of focus in the field of artificial intelligence (AI) and machine learning (ML) (Messud & Chambefort, 2020). Interpretability refers to the extent to which a human can understand the cause of a decision made by a model. As DNNs are increasingly employed in various high-stakes applications—such as healthcare, finance, and autonomous driving—the ability to interpret their outputs becomes essential for trust, accountability, and transparency (Eberle et al., 2021). These applications demand not only high performance from AI systems but also a clear understanding of the decision-making process to ensure reliability and ethical compliance. The complexity and depth of DNN architectures, characterized by multiple interconnected layers, often result in models functioning as "black boxes," where the rationale behind specific decisions is obscured. This opacity poses significant risks in critical fields, as incorrect or biased decisions can have severe consequences.

In the context of healthcare, for instance, the deployment of DNNs for diagnostic purposes necessitates an interpretable model to facilitate trust among clinicians and patients. For example, the development and validation of interpretable neural networks for predicting postoperative in-hospital mortality have shown that leveraging neural networks' ability to learn nonlinear patterns, while maintaining interpretability, can significantly improve clinical outcomes (Lee et al., 2021). Similarly, interpretable models are essential in predicting patients' choices of hospital levels to manage healthcare resources efficiently and ensure appropriate patient care (Chen et al., 2021). In finance, understanding model decisions is crucial for regulatory compliance and risk management. Interpretability in DNNs helps in identifying the most influential factors in financial models, thereby ensuring transparent and accountable decision-making processes. For instance, models that

predict stroke occurrence using large-scale electronic medical claims data have demonstrated the necessity of interpretability to identify key predictors accurately and ensure model reliability (Hung et al., 2017).Autonomous driving systems also require interpretability to ensure safety and accountability in case of accidents or system failures. DNNs used for real-time decision-making in autonomous vehicles must be interpretable to provide insights into the decision-making process and ensure that the vehicle's actions can be justified and understood by developers and regulators. Studies on cardiologist-level arrhythmia detection using DNNs have highlighted the importance of interpretability in validating and gaining trust in automated systems (Hannun et al., 2019). Hence, enhancing the interpretability of DNNs is not merely a technical challenge but a societal imperative. The ability to understand and trust these models is crucial for their adoption in critical fields, where the consequences of decisions can be life-altering. By focusing on architectural elements that improve interpretability, researchers and practitioners can develop more transparent and trustworthy AI systems, thereby fostering broader acceptance and ethical application of AI technologies.

This article reviews the architectural elements of DNNs that influence their interpretability, providing insights into how these components can be designed to create more transparent and trustworthy models. It explores various types of layers, such as convolutional, recurrent, and attention layers, and their impact on model interpretability. The depth and complexity of the network, connectivity patterns, and the role of attention mechanisms are examined to understand how these factors contribute to or detract from interpretability. By identifying strategies to balance performance and transparency, this review aims to aid researchers and practitioners in developing DNNs that are not only effective but also interpretable, thus fostering broader acceptance and ethical application of AI technologies (Yu et al., 2016).

DNN architecture, known for its complexity and depth, consists of multiple layers, each contributing to the model's ability to learn from data. However, this complexity often comes at the cost of interpretability. Unlike traditional machine learning models, whose decision-making processes are more transparent, DNNs operate as black boxes, making it challenging to understand how they arrive at specific decisions (Mouton & Davel, 2022). This lack of transparency raises concerns, particularly when DNNs are used in critical applications where understanding the reasoning behind decisions is paramount. For example, in healthcare, the deployment of DNNs for diagnostic purposes necessitates interpretable models to facilitate trust among clinicians and patients (Lee et al., 2021). Similarly, in finance, regulatory compliance and risk management require clear model explanations (Hung et al., 2017). In autonomous driving, interpretability is critical for ensuring safety and accountability (Hannun et al., 2019).

The primary objective of this article is to review the architectural elements of DNNs that contribute to their interpretability. By examining different layer types, network depth and complexity, connectivity patterns, and attention mechanisms, the article aims to provide a comprehensive understanding of how these architectural components can be designed and utilized to enhance the interpretability of DNNs. This review will help researchers and practitioners identify strategies to make DNNs more transparent and trustworthy without compromising their performance (Yu et al., 2016).

The article is structured into four sections, each focusing on a key architectural element of DNNs that influences interpretability:

1. **Layer Types:** This section will explore various types of layers, including convolutional, recurrent, and attention layers, and their unique properties affecting model interpretability. Convolutional layers are fundamental for image recognition tasks, while recurrent layers handle sequential data, and attention layers improve model performance by selectively focusing on relevant parts of the input (Gandin et al., 2021).
2. **Network Depth and Complexity:** This section will discuss how the depth and complexity of networks impact interpretability, highlighting the trade-offs between shallow and deep networks. Shallow networks are more interpretable but less powerful for complex tasks compared to deep networks, which offer greater accuracy but at the expense of transparency (Miyoshi et al., 2019).
3. **Connectivity Patterns:** This section will examine how different connectivity patterns within the network, such as fully connected layers and structured connectivity like residual networks, influence interpretability. Fully connected layers provide high flexibility but pose interpretability challenges due to dense connections, whereas structured patterns like residual networks offer clearer information flow (Liu et al., 2019).
4. **Attention Mechanisms:** This section will analyze how attention mechanisms enhance interpretability by focusing on significant parts of the input data. Attention mechanisms dynamically highlight relevant input features, making the decision-making process of DNNs more transparent (Zheng et al., 2021).\

Through the systematic examination of these architectural elements, the article aims to bridge the gap between high performance and interpretability in DNNs, ensuring that these powerful models can be used effectively and responsibly in critical applications. The increasing adoption of DNNs in fields such as healthcare, finance, and autonomous driving underscores the necessity for models that not only perform well but are also interpretable and transparent (Lee et al., 2021). By focusing on architectural elements such as layer types, network depth and complexity, connectivity patterns, and attention mechanisms, this article provides a framework for designing DNNs that are both high-performing and interpretable. This balance is crucial for the broader acceptance and ethical application of AI technologies across various fields. Researchers and practitioners can utilize the insights from this review to develop AI systems that are not only effective but also transparent and trustworthy, thereby addressing the critical need for interpretability in the deployment of AI in real-world applications (Yu et al., 2016).

Layer Types : Understanding the various layer types within DNNs is crucial for grasping how these models process and interpret data (Mouton & Davel, 2022). Each layer type—convolutional, recurrent, and attention—plays a unique role in shaping the network's functionality and interpretability. Convolutional layers excel in image recognition by identifying spatial hierarchies of features, recurrent layers are adept at handling sequential data by capturing temporal dependencies, and attention layers enhance model performance by selectively focusing on the most relevant parts of the input. By examining these different layer types, one can gain insight into their respective properties and functions, as well as the interpretability challenges and techniques associated with each. This understanding is essential for developing more transparent and effective DNNs, ultimately contributing to the advancement of machine learning applications (Zhao & Gao, 2021).

Convolutional Layers : Convolutional layers are fundamental components of convolutional neural networks (CNNs) and are particularly effective in image recognition tasks. These layers apply convolution operations to the input data, enabling the network to detect and learn spatial hierarchies of features through multiple layers of filters. Each filter is designed to identify specific patterns such as edges, textures, or more complex features as the data progresses through deeper layers (Jang et al., 2021). The interpretability of convolutional layers is primarily achieved through the visualization of learned filters and feature maps. By examining these visualizations, one can understand what features the network focuses on at different layers. For instance, the initial layers might capture basic edges and textures, while deeper layers identify more complex patterns and object parts. These visualizations help in demystifying the black-box nature of CNNs by providing insights into the feature extraction process and the hierarchical nature of learned representations (Gandin et al., 2021).

Recurrent Layers : Recurrent layers, such as those found in recurrent neural networks (RNNs) and long short-term memory networks (LSTMs), are designed to handle sequence data (Chaudhuri, 2019). These layers maintain a hidden state that captures information from previous time steps, making them suitable for tasks involving temporal dependencies, such as speech recognition, language modeling, and time-series prediction. Understanding the outputs of recurrent layers poses significant challenges due to their temporal dependencies and the complex interactions within the hidden states. However, techniques such as attention mechanisms and gradient-based methods can be employed to interpret recurrent layer outputs. These techniques help highlight which parts of the input sequence are most influential in the model's predictions, thus aiding in the interpretability of the model's sequential decision-making process (Bianchini & Scarselli, 2014).

Attention Layers : Attention layers enhance the network's ability to focus on the most relevant parts of the input data. By assigning different weights to different input elements, attention mechanisms enable the model to prioritize significant features, thereby improving performance in tasks such as machine translation, text summarization, and image captioning (Zhao, 2021). The interpretability of models with attention layers is significantly improved through attention visualization. These visualizations depict the weights assigned to different input elements, providing a clear indication of which parts of the input the model considers most important for a given task. This not only enhances the transparency of the model's decision-making process but also allows for a better understanding of the underlying data relationships. Attention mechanisms, therefore, serve as powerful tools for both improving model performance and facilitating interpretability (Gandin et al., 2021).

Network Depth and Complexity : The depth and complexity of neural networks are crucial factors influencing both their performance and interpretability. As neural networks become deeper, with more layers and parameters, their ability to capture intricate patterns and relationships within data increases (Mouton & Davel, 2022).

This capability is particularly beneficial for solving complex tasks that require a high degree of abstraction and generalization. However, the increased depth and complexity also introduce significant challenges in terms of understanding and interpreting the model's internal workings. Balancing the trade-off between achieving high performance and maintaining interpretability is a key concern in the design and application of deep neural networks (Yu et al., 2016); (Zhao & Gao, 2021). This section explores the properties and interpretability of both shallow and deep networks and discusses strategies for achieving an optimal balance between complexity and transparency.

Shallow Networks : Shallow networks are characterized by having fewer layers and parameters. They are typically simpler in structure and easier to train compared to deeper networks. Due to their simplicity, shallow networks are generally easier to interpret. The limited number of layers and parameters allows for a more straightforward analysis of how inputs are transformed through the network. This simplicity aids in understanding the decision-making process of the model (Mouton & Gavel, 2022). Shallow networks are particularly useful in applications where interpretability is crucial, and the complexity of the task does not demand deep architectures.

Deep Networks : Deep networks, on the other hand, consist of many layers, often resulting in increased depth and complexity. This architectural choice enables deep networks to capture and learn from intricate patterns within data, making them particularly effective for complex tasks. For instance, deep neural networks have shown remarkable performance in fields like image recognition, natural language processing, and medical diagnosis (Lee et al., 2021). The primary trade-off in deep networks lies between their high performance on complex tasks and their reduced interpretability. While deeper networks can achieve superior accuracy and generalization, understanding the internal workings of these models becomes increasingly challenging as complexity grows (Daniels et al., 2020). This opacity can hinder their adoption in critical fields where model transparency is essential for trust and accountability.

Strategies for Balancing : To balance the depth of neural networks with the need for interpretability, several strategies can be employed. Techniques such as network pruning, which reduces the number of parameters while maintaining performance, can help simplify models. Additionally, employing visualization tools and techniques like layer-wise relevance propagation can provide insights into the decision-making process of deeper networks (Daniels et al., 2020). Another approach is using hybrid models that combine shallow interpretable layers with deeper, more complex structures, thereby leveraging the strengths of both architectures. These strategies aim to maintain the performance benefits of deep networks while enhancing their transparency and interpretability (Miyoshi et al., 2019).

Connectivity Patterns : The way neurons are connected within a neural network significantly impacts the model's interpretability. Connectivity patterns determine how information flows through the network, influencing the complexity and transparency of the learning process (Zhang et al., 2021); (Naseer et al., 2023). While fully connected layers offer high flexibility, they often pose significant challenges for interpretability due to their dense connections. In contrast, structured connectivity patterns, such as those found in residual networks, can provide a clearer flow of information, enhancing the model's transparency. This section examines the properties and interpretability challenges of fully connected layers and explores the benefits of structured connectivity patterns, particularly residual networks.

Fully Connected Layers : Fully connected layers, also known as dense layers, connect every neuron in one layer to every neuron in the next layer. This architecture offers high flexibility, allowing the network to learn complex representations by combining information from all previous layers (Zhang et al., 2021). However, the dense connections in fully connected layers make it challenging to understand the specific contributions of individual neurons and layers to the final output. This complexity complicates the interpretation of the model's decision-making process. As a result, fully connected layers often lack transparency, making it difficult for researchers and practitioners to gain insights into how the network arrives at its predictions.

Structured Connectivity Patterns : Residual networks (ResNets) introduce a more structured connectivity pattern by incorporating shortcut connections that bypass one or more layers (Chitty-Venkata & Somani, 2022). These shortcut connections allow the network to directly propagate information from earlier layers to later layers, facilitating the learning process and improving model performance. The structured connectivity in residual networks provides a clearer flow of information, making it easier to trace the contributions of different layers to the final output.

This improved clarity helps in understanding how the network processes information and makes decisions, thereby enhancing interpretability.The use of residual connections also mitigates the problem of vanishing gradients, allowing for the training of deeper networks without compromising interpretability. By maintaining a direct path for gradient flow, ResNets ensure that the learning process remains efficient even as the network depth increases. This capability makes ResNets particularly suitable for applications requiring deep architectures while still retaining a degree of interpretability (Chitty-Venkata & Somani, 2022). Through the careful design and implementation of connectivity patterns, it is possible to achieve a balance between the performance benefits of deep neural networks and the need for interpretability. Researchers and practitioners can leverage these insights to develop models that are both powerful and transparent, thereby fostering trust and accountability in AI systems used in critical applications.

Attention Mechanisms : Attention mechanisms have emerged as a pivotal architectural element in deep neural networks (DNNs), significantly enhancing their interpretability. By enabling the model to focus on the most relevant parts of the input data, attention mechanisms provide insights into which aspects of the input are most influential in the decision-making process. This section discusses the importance of attention mechanisms in DNNs, how they enhance interpretability, and their applications across various tasks.

Importance in DNNs : Attention mechanisms play a crucial role in DNNs by dynamically highlighting the most relevant parts of the input data during the processing stages. This selective focus allows the network to prioritize important features while disregarding less significant ones, thereby improving the model's performance and making its behavior more understandable (Gandin et al., 2021). By assigning different weights to different parts of the input, attention mechanisms help to clarify the internal workings of complex models. This capability is essential in tasks where understanding the model’s focus can inform improvements in both the model and the data it processes (Zou & Ding, 2021).

Interpretability Enhancement : One of the primary benefits of attention mechanisms is their ability to enhance interpretability through the visualization of attention weights. These weights reveal which parts of the input data the model is focusing on at any given time, providing a clear and intuitive understanding of the model's decision-making process (Daniels et al., 2020; Blautzik et al., 2013). By examining these visualizations, researchers and practitioners can gain valuable insights into the inner workings of the model, identifying which features are most influential and how they contribute to the final output. This transparency is particularly important in applications where the stakes are high, such as in medical diagnostics and financial forecasting.

Applications : Attention mechanisms are widely applied across various tasks to aid interpretability. For instance, in natural language processing (NLP), attention mechanisms help models focus on relevant words or phrases in a sentence, improving the understanding of language structure and context. In image recognition, attention mechanisms highlight critical regions within an image, facilitating the identification of key features that drive classification decisions (Grigg & Grady, 2010; Tutek & Šnajder, 2022). These applications demonstrate how attention mechanisms not only enhance model performance but also provide a clearer understanding of how models process and interpret data. By making the decision-making process more transparent, attention mechanisms enable the development of more trustworthy and accountable AI systems. Through the use of attention mechanisms, DNNs can achieve a balance between high performance and interpretability, ensuring that these models can be effectively and responsibly deployed in real-world applications.

II. DISCUSSION

The architectural elements reviewed thus far contribute to the interpretability of DNNs, focusing on layer types, network depth and complexity, connectivity patterns, and attention mechanisms. Convolutional, recurrent, and attention layers each have unique properties that impact interpretability. Convolutional layers allow for visualization of learned filters, aiding in understanding feature extraction. Recurrent layers handle sequential data but present challenges in understanding temporal dependencies. Attention layers enhance interpretability by highlighting important parts of the input. Network depth and complexity influence interpretability, with shallow networks being more interpretable but less powerful on complex tasks compared to deep networks. Connectivity patterns, including fully connected layers and structured patterns like residual networks, also play a significant role. Fully connected layers offer high flexibility but are difficult to interpret, while structured connectivity patterns provide clearer information flow. Attention mechanisms, by focusing on relevant input parts, significantly enhance model interpretability.

Understanding the various layer types within DNNs is crucial for grasping how these models process and interpret data (Mouton & Davel, 2022). Each layer type—convolutional, recurrent, and attention—plays a unique role in shaping the network's functionality and interpretability. Convolutional layers excel in image recognition by identifying spatial hierarchies of features, recurrent layers are adept at handling sequential data by capturing temporal dependencies, and attention layers enhance model performance by selectively focusing on the most relevant parts of the input. By examining these different layer types, one can gain insight into their respective properties and functions, as well as the interpretability challenges and techniques associated with each (Zhao & Gao, 2021).

The depth and complexity of neural networks are crucial factors influencing both their performance and interpretability. As neural networks become deeper, with more layers and parameters, their ability to capture intricate patterns and relationships within data increases (Mouton & Davel, 2022). This capability is particularly beneficial for solving complex tasks that require a high degree of abstraction and generalization. However, the increased depth and complexity also introduce significant challenges in terms of understanding and interpreting the model's internal workings. Balancing the trade-off between achieving high performance and maintaining interpretability is a key concern in the design and application of deep neural networks (Yu et al., 2016; Zhao & Gao, 2021).

The way neurons are connected within a neural network significantly impacts the model's interpretability. Connectivity patterns determine how information flows through the network, influencing the complexity and transparency of the learning process (Naseer et al., 2023; Zhang et al., 2021). While fully connected layers offer high flexibility, they often pose significant challenges for interpretability due to their dense connections. In contrast, structured connectivity patterns, such as those found in residual networks, can provide a clearer flow of information, enhancing the model's transparency. This section examines the properties and interpretability challenges of fully connected layers and explores the benefits of structured connectivity patterns, particularly residual networks.

Attention mechanisms have emerged as a pivotal architectural element in these networks, significantly enhancing their interpretability. By enabling the model to focus on the most relevant parts of the input data, attention mechanisms provide insights into which aspects of the input are most influential in the decision-making process. This selective focus allows the network to prioritize important features while disregarding less significant ones, thereby improving the model's performance and making its behavior more understandable (Gandin et al., 2021). By assigning different weights to different parts of the input, attention mechanisms help to clarify the internal workings of complex models [(Zou & Ding, 2021).

The review thus highlights the importance of architectural elements in enhancing the interpretability of deep neural networks. Convolutional layers, recurrent layers, and attention mechanisms each contribute uniquely to model transparency, while network depth and connectivity patterns significantly affect interpretability. Future research should focus on developing innovative techniques to balance the trade-off between performance and interpretability, ensuring that DNNs can be effectively and responsibly deployed in critical applications. By integrating interpretability-focused designs and leveraging advanced visualization tools, researchers can create more transparent, trustworthy, and efficient AI systems, ultimately contributing to the broader acceptance and ethical application of AI technologies.

III. CONCLUSION

Deep neural networks (DNNs) have revolutionized numerous fields by providing unprecedented performance in tasks such as image recognition, natural language processing, and medical diagnosis. However, their complexity and lack of interpretability pose significant challenges, particularly in high-stakes applications where understanding the reasoning behind decisions is paramount. This article reviewed the architectural elements that contribute to the interpretability of DNNs, focusing on layer types, network depth and complexity, connectivity patterns, and attention mechanisms. The need for this study is underscored by the increasing reliance on DNNs in critical domains such as healthcare, finance, and autonomous driving. In these areas, the opacity of DNN models can hinder their adoption due to concerns over trust, accountability, and transparency. Addressing these issues requires a comprehensive understanding of how different architectural components of DNNs affect their interpretability. Our review highlighted that convolutional layers allow for visualization of learned filters, aiding in understanding feature extraction processes (Jang et al., 2021). Recurrent layers handle sequential data but present challenges in understanding temporal dependencies (Bianchini & Scarselli, 2014). Attention layers significantly enhance interpretability by highlighting important parts of the input, providing insights into the

model's focus (Gandin et al., 2021). Network depth and complexity influence interpretability, with shallow networks being more interpretable but less powerful for complex tasks compared to deep networks (Miyoshi et al., 2019). Connectivity patterns, including fully connected layers and structured patterns like residual networks, also play a significant role, with the latter providing clearer information flow (Chitty-Venkata & Somani, 2022). Future research should continue to explore ways to enhance the interpretability of DNNs without compromising their performance. Developing new visualization tools and techniques for different layer types can help demystify complex models (Lee et al., 2021). Additionally, investigating hybrid models that combine shallow and deep networks may offer a balance between performance and interpretability (Yu et al., 2016). Further research on the ethical implications of DNN interpretability, especially in high-stakes applications, is crucial. Interdisciplinary collaboration among computer scientists, ethicists, and domain experts will be essential to address the multifaceted challenges of making DNNs more transparent and accountable. Finally, the development of standardized benchmarks for evaluating interpretability across different architectures and applications will provide a clearer framework for assessing progress in this critical area. By advancing our understanding of DNN interpretability and developing innovative approaches to enhance it, we can ensure that these powerful models can be deployed responsibly and effectively, fostering broader acceptance and ethical application of AI technologies.

REFERENCES

1. An, J., & Joe, I. (2022). Attention map-guided visual explanations for deep neural networks. Applied Sciences, 12(8), 3846. https://doi.org/10.3390/app12083846
2. Bianchini, M., & Scarselli, F. (2014). On the complexity of neural network classifiers: A comparison between shallow and deep architectures. IEEE transactions on neural networks and learning systems, 25(8), 1553-1565.
3. Blautzik, J., Vetter, C., Peres, I., Gutyrchik, E., Keeser, D., Berman, A., Kirsch, V., Mueller, S., Pöppel, E., Reiser, M., Roenneberg, T., & Meindl, T. (2013). Classifying fMRI-derived resting-state connectivity patterns according to their daily rhythmicity. NeuroImage, 71, 298-306. https://doi.org/10.1016/j.neuroimage.2012.08.010
4. Chaudhuri, A. (2019). Some insights and observations on depth issues in deep learning networks. In Advances in Computational Intelligence: 15th International Work-Conference on Artificial Neural Networks, IWANN 2019, Gran Canaria, Spain, June 12-14, 2019, Proceedings, Part I 15 (pp. 583-595). Springer International Publishing. https://doi.org/10.1007/978-3-030-20521-8_48
5. Chen, L., Sheu, J., & Chuang, Y. (2021). Predicting Patient's Choices of Hospital Levels Using Deep Learning and Representation Improvements. 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 1251-1257.
6. Chitty-Venkata, K. T., & Somani, A. K. (2022). Neural architecture search survey: A hardware perspective. ACM Computing Surveys, 55(4), 1-36. https://doi.org/10.1145/3524500
7. Daniels, Z. A., Frank, L. D., Menart, C. J., Raymer, M., & Hitzler, P. (2020, April). A framework for explainable deep neural models using external knowledge graphs. In Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications II (Vol. 11413, pp. 480-499). SPIE. https://doi.org/10.1117/12.2558083
8. Eberle, H., Zhang, B., Teodorescu, C. S., Walker, G., & Carlson, T. (2021, October). An 'Ethical Black Box', Learning From Disagreement in Shared Control Systems. In 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 398-403). IEEE.
9. Gandin, I., Scagnetto, A., Romani, S., & Barbati, G. (2021). Interpretability of time-series deep learning models: A study in cardiovascular patients admitted to Intensive care unit. Journal of biomedical informatics, 103876 . https://doi.org/10.1016/j.jbi.2021.103876
10. Grigg, O., & Grady, C. (2010). Task-Related Effects on the Temporal and Spatial Dynamics of Resting-State Functional Connectivity in the Default Network. PLoS ONE, 5. https://doi.org/10.1371/journal.pone.0013311
11. 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
12. Hannun, A., Rajpurkar, P., Haghpanahi, M., Tison, G., Bourn, C., Turakhia, M., & Ng, A. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Medicine, 25, 65 - 69. https://doi.org/10.1038/s41591-018-0268-3
13. Hung, C., Chen, W., Lai, P., Lin, C., & Lee, C. (2017). Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical

claims database. 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 3110-3113. https://doi.org/10.1109/EMBC.2017.8037515

14. Jang, H., McCormack, D., & Tong, F. (2021). Noise-trained deep neural networks effectively predict human vision and its neural responses to challenging images. PLoS Biology, 19. https://doi.org/10.1371/journal.pbio.3001418
15. Lee, C., Samad, M., Hofer, I., Cannesson, M., & Baldi, P. (2021). Development and validation of an interpretable neural network for prediction of postoperative in-hospital mortality. NPJ Digital Medicine, 4. https://doi.org/10.1038/s41746-020-00377-1
16. Liu, C., Gardner, S., Wen, N., Elshaikh, M., Siddiqui, F., Movsas, B., & Chetty, I. (2019). Automatic Segmentation of the Prostate on CT Images Using Deep Neural Networks (DNN). International journal of radiation oncology, biology, physics, 104 4, 924-932 . https://doi.org/10.1016/j.ijrobp.2019.03.017
17. 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
18. 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, 263-74 . https://doi.org/10.1021/ci500747n
19. Messud, J., & Chambefort, M. (2020, November). Understanding how a deep neural network architecture choice can be related to a seismic processing task. In First EAGE Digitalization Conference and Exhibition (Vol. 2020, No. 1, pp. 1-5). European Association of Geoscientists & Engineers. https://doi.org/10.3997/2214-4609.202032076
20. 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
21. Montufar, G. F., Pascanu, R., Cho, K., & Bengio, Y. (2014). On the number of linear regions of deep neural networks. Advances in neural information processing systems, 27.
22. Mouton, C., & Davel, M. H. (2021, December). Exploring layerwise decision making in DNNs. In Southern African Conference for Artificial Intelligence Research (pp. 140-155). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-95070-5_10
23. Miyoshi, T., Higaki, A., Kawakami, H., & Yamaguchi, O. (2019). Automated interpretation of the coronary angioscopy with deep convolutional neural networks. Open Heart, 7. https://doi.org/10.1136/openhrt-2019-001177
24. 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
25. 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
26. Tutek, M., & Šnajder, J. (2022). Toward Practical Usage of the Attention Mechanism as a Tool for Interpretability. IEEE Access, PP, 1-1. https://doi.org/10.1109/access.2022.3169772
27. Yu, D., Xiong, W., Droppo, J., Stolcke, A., Ye, G., Li, J., & Zweig, G. (2016, September). Deep Convolutional Neural Networks with Layer-Wise Context Expansion and Attention. In Interspeech (pp. 17-21). https://doi.org/10.21437/Interspeech.2016-251
28. Yu, H. (2010). Network complexity analysis of multilayer feedforward artificial neural networks. Applications of Neural Networks in High Assurance Systems, 41-55. https://doi.org/10.1007/978-3-642-10690-3_3
29. Zhao, C., & Gao, X. (2021). QDNN: Deep neural networks with quantum layers. Quantum Machine Intelligence, 3(15). https://doi.org/10.1007/s42484-021-00046-w
30. Zheng, W., Amorim, E., Jing, J., Ge, W., Qiao, l., Wu, O., Ghassemi, M., Lee, J., Sivaraju, A., Pang, T., Herman, S., Gaspard, N., Ruijter, B., Sun, J., Tjepkema-Cloostermans, M., Hofmeijer, J., Putten, M., & Westover, M. (2021). Predicting Neurological Outcome in Comatose Patients after Cardiac Arrest with Multiscale Deep Neural Networks.. Resuscitation. https://doi.org/10.1016/j.resuscitation.2021.10.034
31. Zou, J., & Ding, N. (2021). Deep Neural Networks Evolve Human-like Attention Distribution during Reading Comprehension. ArXiv, abs/2107.05799.