IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.53

Volume 4, Issue 4, March 2024

The Role of Analyzable AI and Interpretability in Trustworthy AI Systems

Guru Arun Kumar J[1] and Dr M NNachappa[2]

PG Student, Department of MSc CS-IT [1]
Professor, School of CS & IT[2]
Jain (Deemed-to-be University), Bangalore, India
guruarun584@gmail.com[1] and mn.nachappa@jainuniversity.ac.in[2]

Abstract: *Enhancing trustworthiness and transparency in artificial intelligence systems hinges on the incorporation of Analyzable AI (XAI) and interpretability within machine learning models. Understanding the reasoning behind model predictions or decisions is paramount across various real-world scenarios. This paper provides a comprehensive overview of the current landscape of XAI and interpretability techniques, particularly focusing on deep learning models. It examines methods such as feature visualization, saliency maps, decision trees, and model distillation, while weighing their respective advantages and limitations. Emphasis is placed on selecting the most suitable approach based on specific application needs. The paper concludes by addressing remaining challenges in the field, advocating for the development of standardized metrics to evaluate model interpretability and ensure the reliability and accuracy of explanations provided. In pursuit of fostering trust in AI systems and advancing the field of AI, this paper aims to offer a thorough review of XAI and interpretability techniques in machine learning.*

Keywords: Analyzable AI, XAI, Interpretability, Machine Learning, Deep Learning, Interpretability Metrics, Feature Visualization.

REFERENCES

[1] Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608.

[2] Matthew D. Zeiler and Rob Fergus (2013). Visualizing and Understanding Convolutional Networks.

[3] Molnar, C. (2019). Interpretable machine learning: A guide for making black box models explainable. Leanpub.

[4] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin (2016). Why Should I Trust You? ": Explaining the Predictions of Any Classifier.

[5] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy (2015). Towards Deep Learning Models Resistant to Adversarial Attacks.

[6] Anh Nguyen, Jason Yosinski, and Jeff Clune (2016). Opening the Black Box of Deep Neural Networks via Information.

[7] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), 93.

[8] Lipton, Z. C. (2016). The mythos of model interpretability. arXiv preprint arXiv:1606.03490.

[9] 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).

[10] Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296.

[11] Ustun, B., & Rudin, C. (2019). Visualizing and understanding deep neural networks with heatmap-based importance scores. Journal of Machine Learning Research, 20(1), 3318-3365.

[12] Xu, Z., & Rudin, C. (2019). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 1(5), 252-263.

DOI: 10.48175/IJARSCT-15926

[13] Zhang, J., & Elhoseiny, M. (2020). Interpretable machine learning in healthcare. In Interpretable AI in Healthcare and Medicine (pp. 3-22). Springer.

[14] Deng, L., Hinton, G., & Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: An overview. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 8599-8603). IEEE.