Explainable Artificial Intelligence for Deep Synthetic Data Generation Models

Luís Valina[1], Brígida Teixeira[2], Arsénio Reis[1], Zita Vale[2], Tiago Pinto[1]
University of Trás-os-Montes and Alto Douro, Vila Real, Portugal
LASI, GECAD, Polytechnic of Porto, Porto, Portugal
al70634@alunos.utad.pt, ars@utad.pt, tiagopinto@utad.pt

***Abstract*— Artificial intelligence encapsulates a "black box" of undiscovered knowledge, propelling the exploration of Explainable Artificial Intelligence (XAI) in generative data synthesis and deep learning. Focused on unveiling these "black box" areas, pointed into interpretability and validation in synthetic data generation, shedding light on the intricacies of generative processes. XAI techniques illuminate decision-making in complex algorithms, enhancing transparency and fostering a comprehensive understanding of non-linear relationships. Addressing the complexity of explaining deep learning models, this paper proposes an XAI solution for deep synthetic data generation explanation. The model integrates a clustering approach to identify similar training instances, reducing interpretation time for large datasets. Explanations, available in various formats, are tailored to diverse user profiles through integration with language models, generating texts with different technical detail levels. This research contributes to ethically deploying AI, bridging the gap between advanced model complexities and human interpretability in the dynamic landscape of artificial intelligence.**

***Index Terms* - Artificial intelligence, explainable artificial intelligence, synthetic data generation, deep learning, and clustering.**

I. INTRODUCTION

Nowadays, Artificial Intelligence (AI) is democratized in our everyday life. Consequently, the proliferation of AI is having a significant impact on society. Indeed, AI has already become ubiquitous, and we have become accustomed about AI making decisions for us in our daily life. However, in life-changing decisions such as disease diagnosis, it is important to know the reasons behind such a critical decision. Here, the crucial need for explaining AI, which has been gaining increasing attention recently, outcomes become fully apparent. AI algorithms suffer from opacity, that it is difficult to get insight into their internal working mechanism [1].

The union of neurosymbolic integration with AI technologies like autonomous reinforcement learning and synthetic data generation heralds a new era of AI capabilities. Synthetic data generation, empowered by generative adversarial networks, represents a revolutionary approach to address data scarcity, a perennial challenge in AI development [2]. However, as these areas are recent, not many solutions of explainable AI (XAI) related to these problems can be found. Some exceptions are a few articles directed to specific topics such as health [3] and deep neural networks [4]. Due to the scarcity of works covering this topic, this work proposes the development of explainable AI solutions for the specific problem of generative data models using deep learning.

II. OVERVIEW OF THE PROPOSED SOLUTION

Interpretability and explanations are crucial for ensuring transparency and trust in black-box models. The solution we propose takes advantage of state-of-the-art interpretability techniques, namely LIME and SHAP, based on the framework proposed in [5], to face the challenges posed by XAI in Deep Learning to generate synthetic data. Figure 1 presents the overview of the proposed approach.

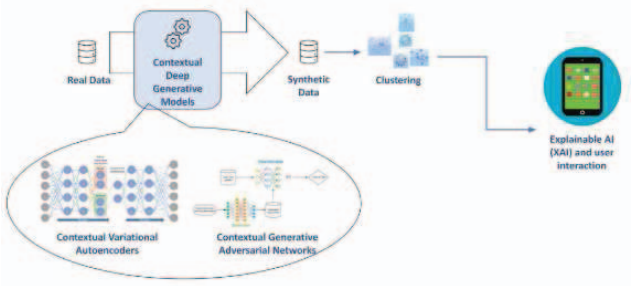

Fig. 1 – Overview of the proposed approach

Focused on improving the interpretability of complex models, especially those with large training datasets, our approach incorporates a new clustering mechanism. In response to the dimensionality of deep learning instances, we integrate clustering to group similar data points together, significantly reducing the number of instances requiring explanation [6]. This strategic adaptation not only speeds up the explanation process, but also ensures a more concise and manageable interpretability task.

Additionally, our solution includes flexibility in explanation formats. Supporting varied profiles, the model provides explanations in graphical, textual, and tabular formats. We introduce adaptability by integrating natural language processing models, namely GPT-3, to automatically generate personalized explanations with different technical backgrounds. This inclusion ensures that explanations are not only accurate, but also understandable and accessible to a diverse audience.

Our proposed solution combines state-of-the-art interpretability techniques, clustering strategies, and adaptability to make explainable AI a robust and easy-to-use tool in the context of deep learning for generating synthetic data.

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00109

III. PRELIMINARY RESULTS

This section, through Figure 2, Figure 3, Figure 4 and Figure 5, presents some preliminary results that showcasome of the expected ecpalanations supported by the proposed solution after carrying out the entire solution implementation process using SHAP and LIME, and using some examples from the framework proposed in [5].

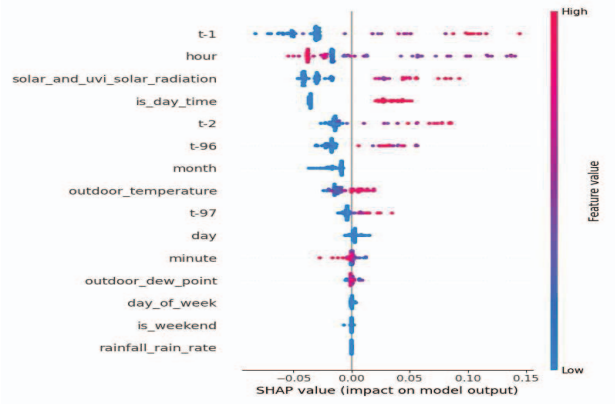

Figure 1. SHAP global interpretation

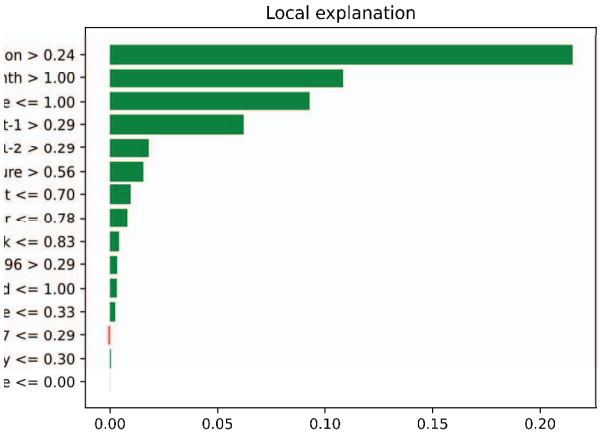

Figure 2. LIME local explanation

From Figure 2 and Figure 3, it is possible to check which feature values contributed most to the prediction of a specific value. In the specific case of generating explanations for deep generative models, such graphs are useful so that one may evaluate which features contributed most to reaching such a result and/or decision of the explanatory models, and hence the decision to use these two technologies aimed at deep learning and synthetic data generation, as they are quite effective and easy to understand.

Figure 4 enables analyzing the performance of the models, comparing the real values with the values predicted by the model. It is possible to verify that, in fact, the results predicted by the model are very close to the true values, which can help to make the models to be developed in the future more reliable.

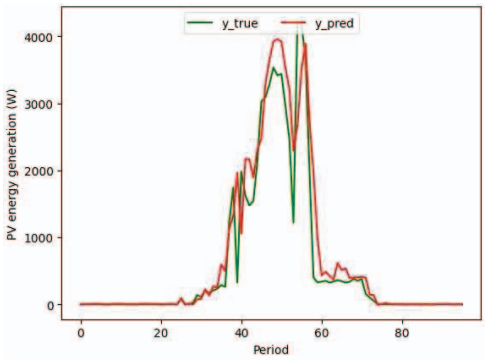

Figure 3. PV generation performance from a limited range

IV. FINAL REMARKS

In summary, our proposed solution integrates advanced interpretability techniques such as LIME and SHAP to enhance XAI in Deep Learning for synthetic data generation. The introduction of clustering streamlines the interpretability process, particularly beneficial for models with extensive training datasets. This initiative is crucial for promoting transparency in complex AI models, ensuring responsible deployment. As we move forward, our focus is on refining and expanding the solution to cater to diverse user needs and contribute to a future where AI transparency is integral to trust, understanding, and responsible innovation.

ACKNOWLEDGEMENTS

The study was developed under the project A-MoVeR – "Mobilizing Agenda for the Development of Products & Systems towards an Intelligent and Green Mobility", operation n.º 02/C05-i01.01/2022.PC646908627-00000069, approved under the terms of the call n.º 02/C05-i01/2022 – Mobilizing Agendas for Business Innovation, financed by European funds provided to Portugal by the Recovery and Resilience Plan (RRP), in the scope of the European Recovery and Resilience Facility (RRF), framed in the Next Generation UE, for the period from 2021 -2026.

REFERENCES

[1] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, Sep. 2018, doi: 10.1109/ACCESS.2018.2870052.

[2] S. Rayhan and D. Gross, "Revolutionizing Intelligence: Unraveling the Frontiers of Advanced Artificial Intelligence and its Impact on Society", doi: 10.13140/RG.2.2.18500.60808.

[3] M. Lenatti, A. Paglialonga, V. Orani, M. Ferretti, and M. Mongelli, "Characterization of Synthetic Health Data Using Rule-Based Artificial Intelligence Models," IEEE J Biomed Health Inform, vol. 27, no. 8, pp. 3760–3769, Aug. 2023, doi: 10.1109/JBHI.2023.3236722.

[4] E. Tjoa and C. Guan, "Quantifying Explainability of Saliency Methods in Deep Neural Networks With a Synthetic Dataset," IEEE Transactions on Artificial Intelligence, vol. 4, no. 4, pp. 858–870, Aug. 2023, doi: 10.1109/TAI.2022.3228834.

[5] B. Teixeira, L. Carvalhais, T. Pinto, and Z. Vale, "Application of XAI-based framework for PV Energy Generation Forecasting." [Online]. Available: https://arxiv.org/abs/1504.04909v1

[6] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A Review," 2000.