From Black Box to Transparency: Consistency and Cost within XAI

Julio Corona*, Pedro Rodrigues*, Leonardo Almeida*, Rafael Teixeira*, Mário Antunes*† and Rui L. Aguiar*†
*Instituto de Telecomunicações, Universidade de Aveiro, Aveiro, Portugal
E-mails: {jcamejo,pedrofrodrigues,leonardoalmeida,rafaelgteixeira,mario.antunes,ruilaa}@av.it.pt
†DETI, Universidade de Aveiro, Aveiro, Portugal
E-mails: {mario.antunes,ruilaa}@ua.pt

2024 IEEE Globecom Workshops (GC Wkshps) | 979-8-3315-0567-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/GCWkshp64532.2024.11100910

***Abstract*—As the 5G era progresses and the research community shifts its focus to 6G, an unprecedented surge in the adoption of Artificial Intelligence (AI) techniques for network development and operation is expected. AI is envisioned to play a crucial role in 6G networks, enabling intelligent network management, enhanced user experience, and unprecedented levels of connectivity. However, the opaque nature of Machine Learning (ML) models has prompted a shift towards Explainable AI (XAI) techniques to enhance decision-making transparency and auditability. This paper delves into the agreement levels of three state-of-the-art XAI techniques—SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and Permutation Importance (PI)—regarding feature importance. These techniques are applied to three groups of ML models in five distinct 5G network scenarios. Additionally, this paper sheds light on the temporal and energy costs inherent in employing these techniques. Our findings indicate that although PI stands out as the most cost-efficient XAI method, there may be disagreement concerning the relevance of features obtained through XAI methods, even when the models obtain high accuracy, highlighting the importance of considering more than one XAI method in model explanation decisions.**

***Index Terms*—Explainable AI, Machine Learning, 5G, 6G**

I. Introduction

As we stand at the forefront of the 5G era, marked by its promises of enhanced Mobile Broadband (eMBB), Ultra-Reliable Low Latency Communications (URLLC), and Massive Machine-Type Communications (mMTC) [1], anticipation for what comes next, often interchangeably referred to as beyond 5G (B5G) and 6G, is gaining momentum. In this context, the use of Explainable AI (XAI) is rising. While some works, such as [2, 3, 4] aim to understand the decisions of Machine Learning (ML) models, others, such as [5, 6] also use XAI techniques to explain relationships between network variables, either for cognitive purposes or for selecting relevant attributes when training ML models. Additionally, XAI constitutes a fundamental tool for regulatory compliance by facilitating traceability and explanation of Artificial Intelligence (AI) decisions, something essential in critical applications such as healthcare, autonomous transportation, and cybersecurity management [7].

For example, to promote transparency and fairness, the European Union's General Data Protection Regulation (GDPR) mandates explainability for AI models that make critical decisions about individuals [8]. Furthermore, the AI Act [9] requires that AI systems, especially high-risk applications, comply with stringent regulations on safety, transparency, accountability, and fundamental rights to promote trust and innovation in the European Union.

Despite the benefits that XAI techniques offer in this context, a gap has been observed in analyzing costs derived from the computational requirements of XAI techniques. The problem with a lack of understanding regarding the cost of adopting XAI is that XAI might be required regardless of its value. Therefore, selecting an efficient XAI technique becomes crucial. While this paper does not introduce new XAI methods, it explores the consistency, temporal efficiency, and energy costs of three commonly used XAI techniques across five 5G network scenarios. The main contributions of this work are: I) Analysis of the agreement between feature relevance computed by three XAI techniques; II) Insights into the execution time and energy consumption of three XAI techniques; III) Assessment of the relationship between the inference costs of ML models and their explanation costs.

The remaining sections of this document are organized as follows: Section II provides an introduction to the ML models and XAI techniques used in this work. Section III describes the datasets used and the methodology followed during the experimentation. The results are presented and discussed in Section IV. Finally, the conclusions are presented in Section V.

II. Background

In this section, we provide the necessary background for ML and XAI in subsection II-A. After we present a detailed overview of the XAI currently being used in 5G research in subsection II-B.

A. ML and XAI

ML algorithms are computational models designed to enable computers to learn and make predictions or decisions without being explicitly programmed. At the core of ML is the process of training a model on a dataset, where the algorithm analyzes patterns, relationships, and features within the data. During training, the model adjusts its parameters to minimize the difference between its predictions and the actual outcomes in the training set. Once trained, the model can generalize its learning to new, unseen data, making predictions or classifications [10].

Since ML algorithms optimize their parameters without human intervention, a set of techniques aims to explain their decisions. These techniques are part of a group called XAI techniques. XAI techniques help shed light on the “black box” nature of ML algorithms by providing insights into how they arrive at their decisions. One common approach involves generating interpretable explanations for model predictions. These explanations may take the form of feature importance scores, highlighting the key factors that influenced a particular decision. Additionally, some XAI techniques use rule-based systems or decision trees to represent the decision logic in a more human-readable format [11]. These techniques can be classified into two main categories: local and global. Local techniques explain the decisions of a model for a specific instance, while global techniques explain the model’s behavior as a whole.

B. XAI for 5G networks

XAI has been applied in the context of 5G/B5G with various objectives ranging from seeking transparency in the decisions of ML models to identifying dependencies between network variables. For example, [3] focuses on energy-efficient resource utilization in Radio Resource Management (RRM) using SHapley Additive exPlanations (SHAP), CERTIFAI, and Anchors to overcome the performance-explainability trade-off in complex ML models. In [6], the authors propose a SHAP-based framework to identify the impact of dataset features on the predictions of an Random Forest (RF) model applied to predict delays using a publicly available 5G mid-band measurements dataset. In [12], the authors propose an XAI framework, incorporating a Deep Learning (DL) model to predict Key Performance Indicators (KPIs) of network slices. Various XAI techniques, including SHAP, Local Interpretable Model-agnostic Explanations (LIME), RuleFit, and Partial Dependence Plot (PDP), are developed on top of a Federated Learning (FL) model to improve transparency and explainability in FL-based decision-making for B5G ML users such as slice managers.

The authors in [5] applied various global and local explainability methods to analyze the root cause of Service Level Agreement (SLA) violation predictions in a 5G network slicing setup. In [4], the authors propose a methodology utilizing XAI techniques, such as SHAP and other global explanation methods, to enhance transparency in the decision-making processes of ML models employed in network slicing problems like short-term resource reservation.

Some research, on the other hand, focuses on using inherently explainable ML models. For example, [2] explores the use of a Decision Tree (DT) model for Quality of Experience (QoE) classification in wireless networks. Despite being less accurate than a black-box RF classifier, fuzzy DT demonstrates competitive precision and recall values. The authors in [13] propose a novel framework for using explainable ML models such as DT, but integrated with FL for QoE predictions in B5G/6G networks.

III. RESEARCH METODOLOGY

In this work, three of the most widely used XAI techniques were selected: SHAP, LIME, and Permutation Importance (PI). Each XAI technique was employed to estimate feature relevance. To ensure the independence of the results of the XAI techniques from the models, each technique was applied with thirteen ML models that are representative of the broader landscape of models used in 5G/6G networks. This includes three groups: Classical classifiers, Ensemble classifiers, and DL classifiers. The classical classifiers used are DT, K-Nearest Neighbors (KNN), Multi-Layer Perceptron (MLP), Stochastic Gradient Descent (SGD) Classifier, and Gaussian Naive Bayes (GNB). The ensemble classifiers used are RF, Voting Classifier (VC), Bagging Classifier (BC), and AdaBoost Classifier (ABC). Finally, two Deep Neural Networks (DNNs) classifiers were selected considering network architectures used in [14] and [15] for two of the datasets used in our experiments.

In the case of VC, we set DT, KNN, and SGD as the voted classifiers. Only for SGD, we set the *modified_huber* loss to enable interpretability. Default values for hyperparameters were retained for all other models.

To explain the features’ relevance, of each model, a global approach was used. This approach was chosen because it allows the analysis of the model’s behavior as a whole, which is relevant in the context of this work. However, LIME is a local XAI technique, usually applied to a single sample. For this work, LIME was applied to each sample of the testing set, and the average of the results was used as a global explanation.

Spearman Correlation Coefficient (SRCC) metric was used to assess the correlation of the outputs from the XAI techniques. This metric measures the statistical association between two variables without assuming a linear relationship between them, making it suitable for assessing monotonic relationships. The correlation coefficient in this context was computed using the formula provided in Equation 1,

$$\mathrm{SRCC}(X,Y)=\frac{\mathrm{cov}(\mathrm{R}(X),\mathrm{R}(Y))}{\sigma(\mathrm{R}(X))\sigma(\mathrm{R}(Y))} \tag{1}$$

where the estimated sets of features relevance X and Y are converted to the ranks $\mathrm{R}(X)$ and $\mathrm{R}(Y)$, respectively, $\mathrm{cov}(\mathrm{R}(X),\mathrm{R}(Y))$ represents the covariance of the rank variables, and $\sigma(\mathrm{R}(X))$ and $\sigma(\mathrm{R}(Y))$ are the standard deviations of the rank variables.

To estimate the energy consumption, we used the Linear Model employed by CloudSim and iFogSim [16]. According to the previously mentioned model, the energy consumption of a task expressed by Joules (J) can be calculated using the equation presented in Equation 2

$$E(u,t)=[P_{\text{idle}}+(P_{\text{max}}-P_{\text{idle}})u]t \tag{2}$$

Where P_{idle} denotes the energy consumption of the computing device while the CPU is in the idle state (measured in watts), P_{max} refers to the node’s energy consumption while the CPU is in the busy state (also measured in watts), u indicates the CPU utilization level (expressed between 0 and 1), and

t represents the time that the task needed to be executed (measured in seconds). For this work, an HP ProLiant ML110 G4 was considered, with 10640 MIPS and 135 W consumption in the busy state and 93.7 W consumption in the idle state [17].

Due to the chosen measurement methodology, the experiments' results are independent of the hardware used, as it is possible to obtain the energy consumed by the selected device by mapping the time it took on the different devices based on its MIPS. To speed up the experiments, we used a workstation equipped with an AMD Ryzen Threadripper 2920X microprocessor and 64 GB of RAM. To estimate the MIPS available in this workstation, we used the 7z application, following the methodology outlined in [18]. The measured execution times were then mapped according to the MIPS relation of this workstation and the target device. The models were executed using Python 3.10.6 with the Scikit-learn library. In the case of the DNNs models, they were implemented in Keras, and it was used also the library SciKeras as a wrapper.

To conduct the experiments, five public datasets from the networking field were selected, as the research interest lies in this domain. These datasets have been used in state-of-the-art 5G papers and provide a reasonable degree of variability regarding sample size, feature count, and classification targets. These are:

- **KPI-KQI** [19]: This synthetic dataset provides parameter values for KPIs and Key Quality Indicators (KQIs) across nine distinct 5G services. It comprises 165 samples, each with 14 features.
- **UNAC** [14]: The User Network Activities Classification (UNAC) dataset contains network traffic traces from various users and applications. The dataset comprises 389 samples with 22 features, categorized into three user activities.
- **NSR** [20]: the Network Slicing Recognition (NSR) dataset includes 16 input features such as LTE/5G user equipment categories, packet loss rate, packet delay, and various usage scenarios. It is designed to select the most suitable network slice, even during network failures. The dataset contains 31,583 labeled samples.
- **QoS-QoE** [21]: The QoS-QoE dataset aims to predict content streaming QoE. It comprises over 69,129 samples generated in a high-fidelity and fully controllable simulation environment. Each sample includes 50 features categorized into Context information, Quality of Service (QoS) metrics, Hidden QoE, and target QoE factors.
- **5G Slicing** [15]: The 5G Slicing dataset includes KPIs for both network and device characteristics, with network slices classified into eMBB, URLLC, and mMTC types. It contains 466,739 samples, each with eight input features and one output indicating the slice type.

To train the ML models, 70% of each data set was used, and the remaining 30% was reserved for testing. All available input features were utilized for QoS-QoE, with *StallLabel* selected as the target. In the case of UNAC, the *file* feature was removed as it does not provide helpful information, and *output* was used as the target. Finally, in the case of 5G Slicing, the *Slice Type* was used as the target. After that, in all datasets, no ML preprocessing was performed.

The neural network was taken from [14], we refer to as DNN1v0, and the DNN taken from [15], we refer to as DNN2v0. To extend the exploration, we also implement a variant of each one just adding more neurons per layer, which we refer to as DNN1v1 and DNN2v1, to enhance its capacity to capture complex patterns and representations. To evaluate the models' performance, Matthews Correlation Coefficient (MCC) metric [22] was selected as it provides a reliable and balanced assessment of classification models, especially in scenarios with imbalanced datasets or when evaluating performance across multiple classes is crucial.

The source code, extensive list of statistical results and figures obtained in this work are publicly available on GitHub[1]. Due to limited space in this paper, only the most significant figures are presented in the following section.

IV. Results & Discussion

As shown in Figure 1, the predictive performance of the models varied significantly across the different datasets, as indicated by the MCC. This variation highlights the models' differing abilities to generalize to unseen data. Some models performed exceptionally well across all instances, while others showed instability or consistently poor performance.

Models such as DT, RF, BC, and ABC consistently demonstrated high performance across most datasets, often achieving perfect MCC scores ($MCC = 1$). This indicates that these models are robust and capable of generalizing well for the datasets considered. In contrast, models like SGD and GNB showed more variability, with some datasets yielding poor performance ($MCC \leq 0.2$).

The DNN-based model exhibited more instability and lower performance than traditional models. This instability may be attributed to the fact that DL models tend to require large amounts of data to learn effectively, and their performance can be highly sensitive to the dataset's characteristics. In fact, only with the 5G Slicing dataset did these models perform perfectly, likely because this dataset seems simpler and has many more samples than the others.

However, given that the DNN1v0 model was taken from the UNAC dataset's original publication [14], it was expected that the DNN1v0 and DNN1v1 models would perform better in this dataset. The determining factor for this low performance appears to be the absence of dataset preprocessing, especially considering that in [14], the authors applied various preprocessing techniques.

To ensure fairness in the analysis of the results for feature correlation, the results will be divided into two groups: models with high performance ($MCC > 0.8$) and models with low performance ($MCC \leq 0.8$) [22].

During the examination of the correlation between feature relevance obtained through XAI techniques, it was expected

[1]https://github.com/FloderCC/XAI_Cost

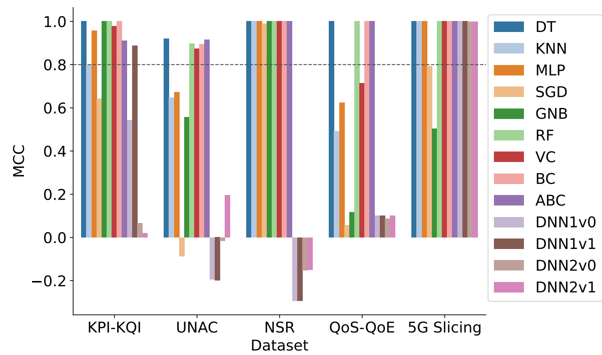

Fig. 1: Model performance measured using MCC metric. The missing bars have a MCC value of 0. The horizontal line represents the threshold between high and low performing model.

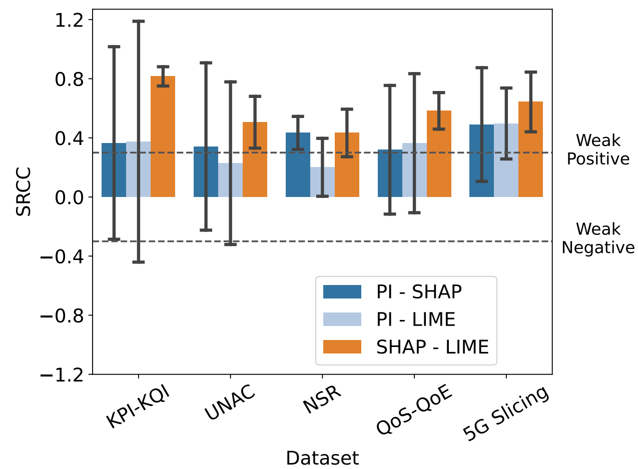

Fig. 3: Correlation between XAI results regarding datasets for low-performance models

that explanations provided for high-performing ML methods would exhibit greater consistency compared to those generated for low-performing ones. Figures 2 and 3 display averages and standard deviations of correlations grouped by dataset. As can be observed, the two groups exhibited quite different degrees of correlation. While some correlations were moderate (above 0.3), others were notably strong (above 0.8), with slightly higher agreement overall in high-performance models. The highest agreements were found with the 5G Slicing dataset, particularly among high-performance models. However, for the QoS-QoE dataset, better agreement was observed among explanations for low-performance models, especially between SHAP and LIME. This suggests that explanation consistency depends not only on the model's performance but also on its type. Moreover, with fewer classifiers, disagreement among XAI systems is less likely, especially evident in simpler datasets where achieving consensus is easier. Furthermore, standard deviations show significant variation in XAI agreement levels, depending on the type of ML model explained.

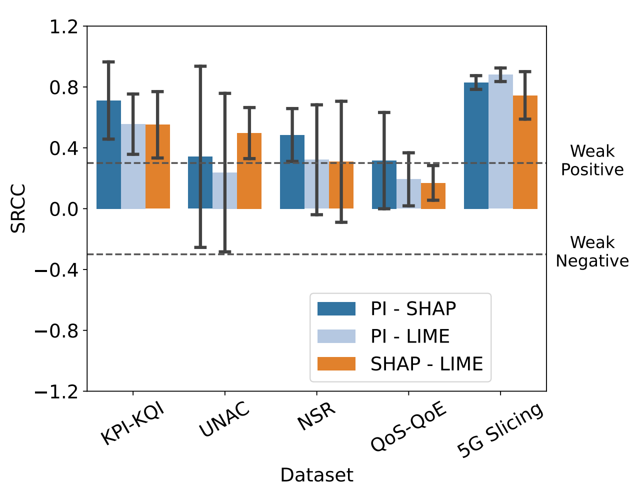

Fig. 2: Correlation between XAI results regarding datasets for high-performance models

To analyze the impact of models on XAI agreement levels, the results were grouped by models instead of datasets. Figures 4 and 5 show the average and standard deviation of correlations for each model. As expected, overall agreement was higher in the group of high-performance models. The highest levels of agreement were observed between PI and SHAP, closely followed by SHAP and LIME.

Significant standard deviations across models indicate dataset influence on explanation agreement, possibly due to varying model performances. Thus, when a model achieves better results, it becomes clearer to understand which features influence it. Other high standard deviations were obtained with the SGD, GNB, and DNN2v1 models, possibly because their performance depends heavily on the datasets. In contrast, it is observed that in the group of high-performance models, no standard deviations were obtained for SGD, DNN1v0, DNN2v0, and DNN2v1 models, as they were high-performance models with only one dataset. Similarly, in the low-performance model group, the VC model showed no standard deviation for the same reason. In this group, the feature relevances estimated for the GNB model by PI differed significantly from those estimated by SHAP and LIME, resulting in many negative correlations. This could be because with GNB models, PI struggles with feature independence assumptions, while SHAP and LIME can better account for it. Other negative correlations in this group were observed in the DNN2v1 model, likely due to its combination of low performance and high complexity.

The focus will now shift to the energy consumption associated with different XAI metrics. Since execution time and energy consumption are closely related (the longer the XAI metric takes, the more energy is consumed), the discussion will primarily concentrate on the energy consumed. For detailed insights into the relationship between execution time and energy consumption, plots illustrating these aspects are available on GitHub.

While analyzing energy consumed, it is clear that all XAI

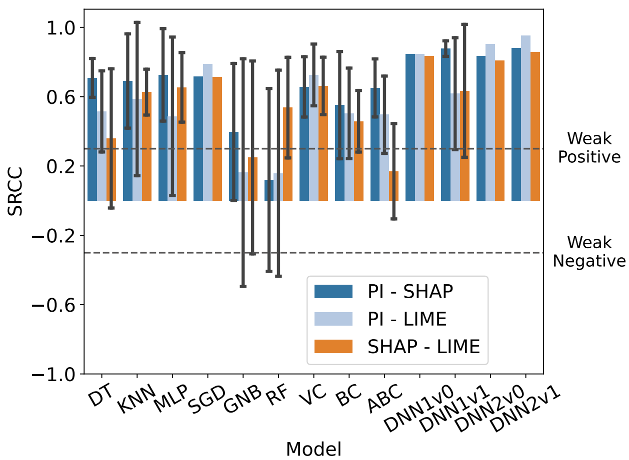

Fig. 4: Correlation between XAI results regarding models for high-performance models

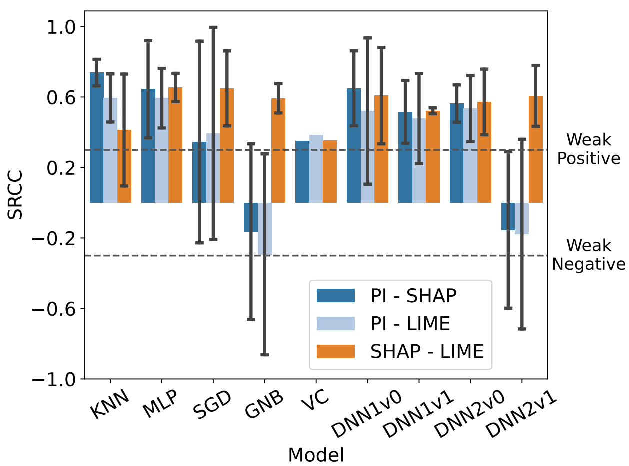

Fig. 5: Correlation between XAI results regarding models for low-performance models

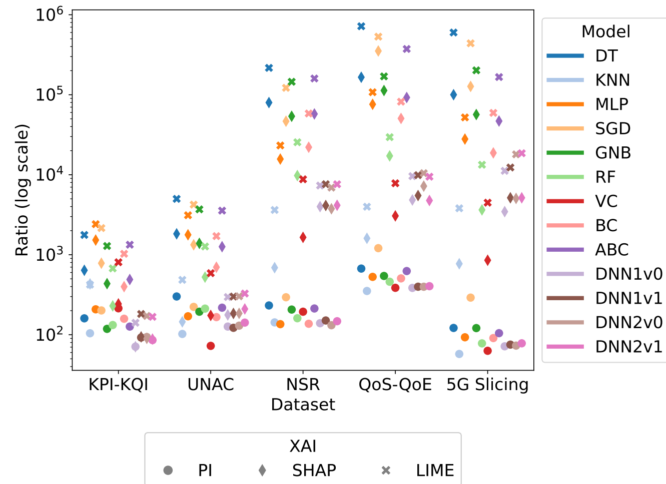

Fig. 6: Energy consumption ratios

methods exhibited a high dependence on both dataset size and the ML model used. Larger datasets, such as 5G Slicing, generally consumed more energy for the same XAI techniques. Compared to PI, SHAP and LIME tended to have significantly higher values of energy consumption in all cases. This difference was more pronounced with DNN, possibly due to the higher energy consumption of the inference. Following the DNN models, KNN and VC also consumed more energy. This was expected since KNN is a lazy learner, and KNN is one of the voted classifiers in VC. To shed light on the relationship between the energy consumption of the explainability process e compared to the inference process i, we computed the energy consumption ratio e/i. Figure 6 presents the ratios with datasets ordered according to their dimensions (number of samples $\times$ number of features).

The energy consumption ratio analysis revealed that the explainability process can consume a lot more energy than the inference. Lower ratios were observed for KNN, VC, and DNNs explanations, indicating their inference process is computationally more expensive. Therefore, the energy consumption of the overall explainability process is less noticeable. LIME exhibited the highest temporal and energy consumption, which suggests that it might not be suitable for global feature relevance inference. Finally, looking through the results using the five datasets, we also observed that for each model and XAI technique, not only do the energy consumptions of ML model inference and XAI depend on the dataset, but also the energy consumption ratio. One of the main reasons for this is that the dataset dimensions can affect the energy consumption of the explainability algorithm.

Considering the overall results, it is worth noting that PI stood out as the most cost-efficient. Furthermore, it is noticeable that, even when considering models that achieve high MCC values, there can be considerable disagreement in the XAI methods feature's relevance. Consequently, when choosing how to explain a model, using just one XAI method can lead to bias/false results. So, in addition to each XAI already having a considerable execution cost, the cost is increased even further by requiring more than one method to ensure that the model is correctly explained.

V. Conclusions

This research explored the agreement levels and time and energy costs of various XAI techniques, identifying PI as notably cost-efficient across diverse ML models and datasets. However, despite its cost efficiency, our findings highlight that even though it achieves better agreement levels, it can also present considerable discrepancies in some situations. Therefore, the selection of the XAI method cannot be limited by its cost alone but also by its ability to "explain" the model. Furthermore, the frequency of XAI application should be judiciously determined, considering its higher time and energy consumption than standard inference, and balanced against the necessity for explanation in the given context.

Acknowledgements

This study was funded by the PRR – Plano de Recuperação e Resiliência and by the NextGenerationEU funds at University of Aveiro, through the scope of the Agenda for Business Innovation “NEXUS: Pacto de Inovação – Transição Verde e Digital para Transportes, Logística e Mobilidade” (Project nº 53 with the application C645112083-00000059).

References

[1] Z. Zhu, X. Li, and Z. Chu, “Three major operating scenarios of 5g: embb, mmtc, urllc,” *Intell. Sens. Commun. Internet Everything*, vol. 1, pp. 15–76, 2022.

[2] A. Renda, P. Ducange, G. Gallo, and F. Marcelloni, “Xai models for quality of experience prediction in wireless networks,” in *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2021, pp. 1–6.

[3] A. Marcu, S. K. G. Peesapati, J. M. Cortes, S. Imtiaz, and J. Gross, “Explainable artificial intelligence for energy-efficient radio resource management,” in *2023 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2023, pp. 1–6.

[4] P. Barnard, I. Macaluso, N. Marchetti, and L. A. DaSilva, “Resource reservation in sliced networks: an explainable artificial intelligence (xai) approach,” in *ICC 2022-IEEE International Conference on Communications*. IEEE, 2022, pp. 1530–1535.

[5] A. Terra, R. Inam, S. Baskaran, P. Batista, I. Burdick, and E. Fersman, “Explainability methods for identifying root-cause of sla violation prediction in 5g network,” in *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 2020, pp. 1–7.

[6] E. D. Spyrou and V. Kappatos, “Xai using shap for outdoor-to-indoor 5g mid-band network,” in *2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT)*. IEEE, 2023, pp. 862–866.

[7] S. Saafi, J. Hosek, and A. Kolackova, “Enabling next-generation public safety operations with mission-critical networks and wearable applications,” *Sensors*, vol. 21, no. 17, p. 5790, 2021.

[8] E. Bayamlıoğlu, “The right to contest automated decisions under the general data protection regulation: Beyond the so-called “right to explanation”,” *Regulation & Governance*, vol. 16, no. 4, pp. 1058–1078, 2022.

[9] “Artificial intelligence act,” 2024, accessed: July 30, 2024. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[10] B. Mahesh, “Machine learning algorithms-a review,” *International Journal of Science and Research (IJSR).[Internet]*, vol. 9, no. 1, pp. 381–386, 2020.

[11] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan *et al.*, “Explainable ai (xai): Core ideas, techniques, and solutions,” *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–33, 2023.

[12] S. B. Saad, B. Brik, and A. Ksentini, “A trust and explainable federated deep learning framework in zero touch b5g networks,” in *GLOBECOM 2022-2022 IEEE Global Communications Conference*. IEEE, 2022, pp. 1037–1042.

[13] A. Renda, P. Ducange, F. Marcelloni, D. Sabella, M. C. Filippou, G. Nardini, G. Stea, A. Virdis, D. Micheli, D. Rapone, and L. G. Baltar, “Federated learning of explainable ai models in 6g systems: Towards secure and automated vehicle networking,” *Information*, vol. 13, no. 8, p. 395, 2022.

[14] C. Sandeepa, T. Senevirathna, B. Siniarski, M.-D. Nguyen, V.-H. La, S. Wang, and M. Liyanage, “From opacity to clarity: Leveraging xai for robust network traffic classification,” in *Asia Pacific Advanced Network*, D. Herath, S. Date, U. Jayasinghe, V. Narayanan, R. Ragel, and J. Wang, Eds. Cham: Springer Nature Switzerland, 2024, pp. 125–138.

[15] A. Thantharate, R. Paropkari, V. Walunj, and C. Beard, “Deepslice: A deep learning approach towards an efficient and reliable network slicing in 5g networks,” in *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2019, pp. 0762–0767.

[16] H. Gupta, A. V. Dastjerdi, S. K. Ghosh, and R. Buyya, “ifogsim: A toolkit for modeling and simulation of resource management techniques in the internet of things, edge and fog computing environments,” *Software: Practice and Experience*, vol. 47, no. 9, pp. 1275–1296, 2017.

[17] A. A. Khan, M. Zakarya, and R. Khan, “Energy-aware dynamic resource management in elastic cloud datacenters,” *Simulation modelling practice and theory*, vol. 92, pp. 82–99, 2019.

[18] P. Domingues, F. Araujo, and L. Silva, “Evaluating the performance and intrusiveness of virtual machines for desktop grid computing,” in *2009 IEEE International Symposium on Parallel & Distributed Processing*. IEEE, 2009, pp. 1–8.

[19] J. E. Preciado-Velasco, J. D. Gonzalez-Franco, C. E. Anias-Calderon, J. I. Nieto-Hipolito, and R. Rivera-Rodriguez, “5g/b5g service classification using supervised learning,” *Applied Sciences*, vol. 11, no. 11, p. 4942, 2021.

[20] G. Dutta, “Network slicing recognition,” 2021, accessed: 2024-08-09. [Online]. Available: https://www.kaggle.com/datasets/gauravduttakiit/network-slicing-recognition

[21] V. Vasilev, J. Leguay, S. Paris, L. Maggi, and M. Debbah, “Predicting qoe factors with machine learning,” in *2018 IEEE International Conference on Communications (ICC)*. IEEE, 2018, pp. 1–6.

[22] D. Chicco and G. Jurman, “The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation,” *BMC genomics*, vol. 21, no. 1, pp. 1–13, 2020.