Explainable and Robust Artificial Intelligence for Trustworthy Resource Management in 6G Networks

Nasir Khan, Sinem Coleri, Asmaa Abdallah, Abdulkadir Celik, and Ahmed M. Eltawil

The authors outline a range of explainable and robust AI techniques for feature visualization and attribution; model simplification and interpretability; model compression; and sensitivity analysis, then explain how they can be leveraged for RRM.

Abstract

Artificial intelligence (AI) is expected to be an integral part of radio resource management (RRM) in sixth-generation (6G) networks. However, the opaque nature of complex deep learning (DL) models lacks explainability and robustness, posing a significant hindrance to adoption in practice. Furthermore, wireless communication experts and stakeholders, concerned about potential vulnerabilities, such as data privacy issues or biased decision-making, express reluctance to fully embrace these AI technologies. To this end, this article sheds light on the importance and means of achieving explainability and robustness toward trustworthy AI-based RRM solutions for 6G networks. We outline a range of explainable and robust AI techniques for feature visualization and attribution; model simplification and interpretability; model compression; and sensitivity analysis, then explain how they can be leveraged for RRM. Two case studies are presented to demonstrate the application of explainability and robustness in wireless network design. The former case focuses on exploiting explainable AI methods to simplify the model by reducing the input size of deep reinforcement learning agents for scalable RRM of vehicular networks. On the other hand, the latter case highlights the importance of providing interpretable explanations of credible and confident decisions of a DL-based beam alignment solution in massive multiple-input multiple-output systems. Analyses of these cases provide a generic explainability pipeline and a credibility assessment tool for checking model robustness that can be applied to any pre-trained DL-based RRM method. Overall, the proposed framework offers a promising avenue for improving the practicality and trustworthiness of AI-empowered RRM.

Introduction

6G networks shall integrate emerging technology trends toward operating ultra-dense massive multiple-input multiple-output (MIMO) systems over broader bands at very high frequencies, leading to nontrivial orchestration of network functions and wireless resources. Nonetheless, the ever-increasing complexity of next-generation networks exceeds the modeling and optimization capabilities of conventional analytical methods, making the data-driven AI-based approaches an indispensable tool for designing and operating 6G networks. The AI can provide appealing performance in key performance indicators and robustness to unpredictable changes in the wireless environment due to its model-free nature. Despite AI's remarkable performance, lack of transparency and trust in the model decisions is a major bottleneck in its practical deployment for mission-critical services [1]. Developing quantifiable metrics to ensure the confidence and credibility of the results generated by AI-driven approaches is necessary to establish a reliable 6G network in practical scenarios.

Assessment of the AI models in a precise context is necessary to meet privacy, fairness, robustness, and explainability criteria. Privacy and fairness are crucial to ensure that AI models are not discriminatory or biased. Sensitive AI applications often require understanding how data is being processed and how the model makes decisions, allowing for transparency and accountability and assuring individuals' privacy is not compromised. However, when it comes to radio resource management (RRM), the focus shifts toward interpreting the inference process of AI models and their robustness against adversarial attacks. It is essential to understand the model's inner workings to improve its performance, enabling the use of simplified model structures with fewer parameters. This can also help improve the model's robustness to adversarial attacks, increasing the detection accuracy from both the model and data aspects. Robustness is a critical aspect of trustworthy AI, as it provides confidence in the decision outcomes. This confidence can be translated into trust metrics that are understandable to both technical and non-technical audiences. The use of trustworthy AI models in RRM can have significant impacts on performance and security, and the ability to interpret and explain their operations is crucial to their success.

AI-based RRM usually relies on complex deep learning (DL)-based models and algorithms that are hard to comprehend and interpret due to their black-box nature. One of the key drawbacks of DL-based models is the difficulty in understanding how input data features influence the model's output and the logical reasoning processes behind its decisions, causing a lack of interpre-

ISSN: 0163-6804
Digital Object Identifier: 10.1109/MCOM.001.2300172

Nasir Khan and Sinem Coleri are with Koc University, Turkey;
Asmaa Abdallah, Abdulkadir Celik and Ahmed M. Eltawil are with King Abdullah University of Science and Technology, Saudi Arabia.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

tation and transparency in the decision-making process. Moreover, AI-based models can be sensitive to biases in the training data, which can lead to poor performance or even discriminatory outcomes when the model is applied to new data that differs from the training data. Additionally, AI-based models can be uncertain about their predictions, making it challenging to understand the confidence level of a given decision. The deficiency of explainability and robustness in AI-based models poses a significant risk of losing control over decision-making beyond the comprehension of wireless system designers and end-users, ultimately resulting in a shortfall of human trust.

To incorporate explainability and robustness into AI-based RRM for trustworthy AI, the concept of explainable AI (XAI) has emerged as a crucial tool. XAI encompasses a set of processes and methods that can reveal the inner operations of AI algorithms and their potential strengths, weaknesses, and behaviors in future wireless networks. As 6G networks are anticipated to incorporate AI technologies for network automation and optimization heavily, XAI becomes essential to comprehend complex cognitive processes behind AI-driven decision-making, thereby overcoming challenges of their automated use in network management. Otherwise, the lack of explainability may hinder human trust and slow the adoption of AI-enabled 6G services. Besides, XAI can simplify AI models to reduce the computational and cognitive burden on the systems/engineers. XAI has been previously applied in various networking and communication scenarios, including interpreting DL-based network routing systems, video quality classification, traffic classification, wireless service provisioning, and resource reservation decisions [1]. However, these aforementioned works mainly focus on supervised learning-based approaches to derive explainability at the network/application layers and mostly emphasize the need for explainability for future 6G technologies. The applicability of XAI methods for RRM at the physical and MAC layer has been addressed minimally, for example, [2]. The literature lacks a systematic methodology for applying XAI methods to RRM problems, such as model simplification and credibility assessment of decision outcomes for robustness against adversarial inputs.

This article aims to motivate the usage of XAI techniques to address the limitations of AI, including their black-box nature, sensitivity to biases, and uncertainty in predictions, ultimately leading to more transparent and trustworthy AI-empowered RRM. First, we elucidate the advantages of employing XAI, such as developing a systematic methodology, reducing the model complexity and convergence time, and improving the interpretability and robustness of decisions. Second, we present the pertinent XAI methods for feature visualization, feature attribution, model simplification, sensitivity analysis, model compression and their applicability in improving the explainability and robustness of AI-driven methods. Next, we present two practical case studies of RRM in 6G networks; the first employs an XAI-based feature importance ranking approach to reduce the input size of the deep reinforcement learning (DRL) agents for scalable RRM of vehicular networks, whereas the second focuses on assessing the robustness of a DL-based beam alignment approach for a mmWave MIMO systems. Finally, we conclude the article with a few remarks and future research directions.

Benefits of Explainabilty and Robustness in AI-Based RRM

This section explains how the incorporation of XAI benefits AI-driven RRM by adjusting the model parameters and network policy to reduce the complexity, provide interpretations, and improve model robustness to unpredictable changes.

Systematic Methodology for AI-Based Algorithm Design

AI-Based RRM algorithms provide a very limited insight to comprehend the logic of how and why certain modeling decisions have been made. In conventional optimization theory-based approaches, we follow a systematic methodology where the problem is first formulated based on system model assumptions, then the optimality conditions are analyzed to simplify the problem, and finally, approximate or heuristic algorithms are proposed by exploiting the optimality conditions. However, AI-based approaches mainly determine a set of system parameters and their functions as inputs and a specific model architecture based on prior experience and expert knowledge about the model. Nevertheless, there does not exist any specific methodology to choose the inputs most contributing to the final RRM decisions and a certain modeling architecture. At this point, XAI can be used in developing such a systematic approach by providing methodologies for input feature selection and the development of interpretable models that accurately represent the original system behavior. XAI methods based on feature attribution analysis allow for identifying the features of the input data with the most significant impact on the model output. Utilizing simple yet inherently interpretable models allows for determining a locally reliable representation of the original black-box model behavior and analysis of the accuracy loss.

Reducing Model Complexity and Convergence Time

A naive and intuitive way to reduce the model complexity would be to simplify the model itself by either reducing the dimensionality of the input dataset or minimizing the model size. XAI can alleviate the model complexity and speed up the convergence time by offering methods for eliminating less important input features, aggregating identical states based on state abstraction in Markov decision processes, and utilizing model compression techniques. Feature importance analysis enables the identification of features most pertinent to RRM decisions, allowing for the removal of less important features while still preserving a reasonable level of accuracy. In the case of DRL-based RRM, the states of the underlying model can be aggregated using a compact abstraction of the original state space and ignoring irrelevant state information. XAI techniques combined with model compression methods can be utilized to implement architectural reforms, such as pruning the number of parameters (i.e., removing the number of connections/weights) and compressing a model with a comparable performance to the original model.

Improving Interpretability and Robustness of Decisions

Ensuring the resilience of wireless applications against input perturbation and adversarial attacks is

A naive and intuitive way to reduce the model complexity would be to simplify the model itself by either reducing the dimensionality of the input dataset or minimizing the model size.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

XAI can substantially improve the interpretability and robustness of AI decisions by providing methods for testing the system robustness, identifying potential outliers, and verifying the credibility and conformity of the decision outcomes.

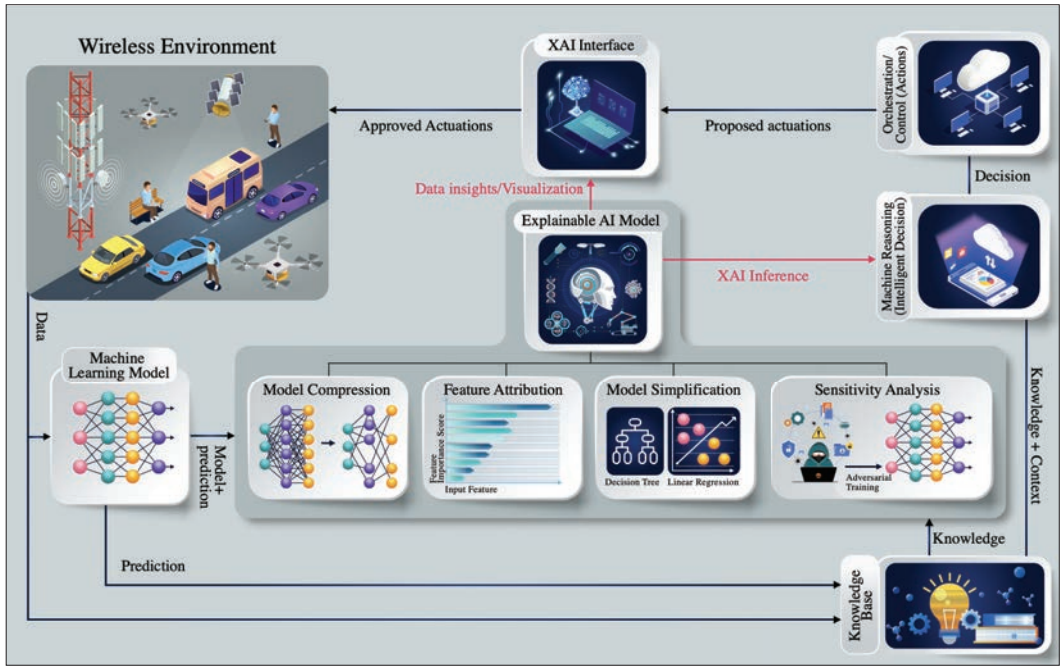


FIGURE 1. XAI-empowered zero-touch orchestration and management framework for explainable and robust RRM.

an important requirement of trustworthiness. Malicious actors can clandestinely inject erroneous data into a network, thereby causing the AI models to make false predictions with high confidence. XAI can substantially improve the interpretability and robustness of AI decisions by providing methods for testing the system robustness, identifying potential outliers, and verifying the credibility and conformity of the decision outcomes. Sensitivity analysis can be leveraged to assess the robustness of the system against adversarial attacks by quantifying the model output's sensitivity to changes in the input data by crafting adversarial input examples. Adversarial training, on the other hand, involves augmenting the training data with adversarial examples and retraining the model to improve its resistance to future attacks and the reliability of its decision outcomes. Furthermore, incorporating feature attribution techniques into intrusion detection systems can aid in identifying outliers or anomalies in the training data. Finally, the credibility and conformity of decision outcomes from complex models can be verified using adversarial training of the learning model. This approach involves crafting adversarial examples that enable the system to detect deviations from its expected behavior.

Techniques to Improve Explainability and Robustness of AI-empowered RRM

Figure 1 presents a generic XAI framework for explainable and robust AI-empowered RRM, which is compliant with the zero-touch network and service management (ZSM) architecture for closed-loop automation. We begin this section with the state-of-the-art techniques and tools that can be utilized to improve the explainability and robustness of AI-based RRM methods.

Feature Visualization and Attribution

Feature visualization techniques offer simple visual illustrations to explain the inner workings and interactions of AI-driven models. Visual augmentation and dimensionality reduction methods can be exploited to generate human-readable interpretations to provide meaningful insights about the model behavior and the decision outcomes. In the context of DRL and unmanned aerial vehicle (UAV) assisted 6G networks, feature visualization can help extract the most important input state features, such as the remaining battery power of the UAV, to better understand the UAV's decision to either move to a new hotspot or stay in its current position [3]. Feature visualization in RRM enables visualizing the characteristics of inputs strongly influencing the output and the derivation of input-output mapping produced by the model, assisting in avoiding conflicts or system-level failures.

Feature attribution methods can further enhance interpretability by assigning importance scores to individual features and assessing their impact on the model outcome and prediction performance. SHapley Additive exPlanations (SHAP) is a particularly effective approach providing a unified framework that assigns each feature an importance value for a particular prediction. Generating feature importance-based explanations for AI-driven RRM problems can reduce the state space for training the DRL agents by filtering the most relevant data to be fed to the AI model, thus simultaneously simplifying the model and improving its understandability.

Model Simplification and Interpretability

Deep neural networks (DNNs) have high predictive power but lack interpretability for wireless network operators. XAI methods offer a solution by simplifying DL-based models, maintaining performance, and providing human-understandable interpretations.

Surrogate models, for example, linear regression or decision trees, can simplify the perception of the complex DL-based models and facilitate the interpretability of AI-empowered RRM methods. The Local Interpretable Model-agnostic Explanations (LIME) is an XAI method that creates a locally faithful representation of the original model behavior by approximating it with a simple, interpretable model. The LIME controls the trade-off between explanation accuracy and its interpretability by introducing

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

a complexity measure. In [4], the LIME is exploited to simplify the DRL-based user traffic offloading in a generic wireless network by dividing the user load demand locally and linearly between the ultra-reliable low-latency communication and enhanced mobile broadband load demands. Then, a feature extraction and visual augmentation-based XAI approach is used to show that the linear model decisions are human-understandable and easy to interpret. The usage of surrogate models in RRM makes model decisions easy to interpret while decreasing decision-making latency and resource consumption.

Knowledge Distillation based approaches utilizing a teacher-student framework can be used to design a simpler and more interpretable model. The teacher model, developed by domain-specific algorithms or human engineers, acts as a transparent system to guide the decision-making of the student model. In the context of DRL-based wireless networking systems, the framework proposed in [5] involves a teacher (e.g., classical buffer-based algorithm) and a student (DRL agent). This framework enhances the robustness of the DRL-based student system by providing guidance against incorrect and risky decisions. It can be adapted to other RRM problems to improve both robustness and interpretability.

Model Compression

Model compression techniques aim to reduce the size and complexity of models while preserving their accuracy, making them more interpretable and easier to understand. Model compression reduces the energy and memory requirements of the model, which can be critical for resource-constrained devices and systems.

Compressing DNNs by exploiting sparsity within the network can be achieved by combining model compression techniques (e.g., model pruning, parameter quantization, low-rank decomposition, lightweight model design) with XAI methods. [6] uses Deep Learning Important FeaTures (DeepLIFT), a back-propagation-based XAI approach, for obtaining the importance of neurons for the pruning and quantization of DNNs. [7] simulates a multichannel orthogonal frequency-division multiple access power allocation network using a simple two-layered DNN to show that a two to three times energy reduction can be achieved by compressing the DNN from twenty to five neurons per layer. Compressing DNNs in DL-based RRM schemes through XAI methods reduces the communication overhead for the periodic broadcast of model parameters and increases the convergence speed.

Sensitivity Analysis

With the ever-increasing adoption of AI-based models for RRM, it is certain that attackers will increasingly seek out new methods of attacking the underlying AI-empowered systems, which has become a critical challenge in many real-world applications. Attackers can deliberately design input data to manipulate the model output and produce inaccurate decisions, posing a significant risk to the integrity of AI models. In the context of RRM, a novel RL-based network slicing framework is proposed to enable the base station to allocate resource blocks (RBs) based on the user's request [8]. However, the victim RL algorithm is shown to be vulnerable to adversarial attacks, where an attacker can observe the spectrum RBs and build a surrogate model to maximize the number of failed requests, thus compromising the algorithm's reward. To address this issue, sensitivity analysis can be employed to develop various reactive and proactive defense mechanisms to understand the vulnerabilities of the DNN models and data aspects.

Backpropagation approaches, which involve computing the gradient of the prediction output with respect to the features of the input data, can be applied in RRM solutions to unveil the rationale behind AI predictions and improve their robustness. Backpropagation-based techniques such as integrated gradients and layer-wise relevance propagation can be used in AI-based RRM to analyze the impact of changing network parameters, such as the number of antennas, modulation schemes, and transmission power levels on the key performance indicators.

Explainable and Robust RRM Case Studies

In this section, we consider two case studies focusing on explainability and robustness in RRM problems. The former case focuses on RRM of vehicular networks and aims at reducing DRL agents' input size by using XAI feature importance methods. The latter case concentrates on DRL-based beam codebook design for MIMO systems and boosts robustness by assessing the credibility of model decisions with the goal of outlier and adversarial input detection.

DRL-Based RRM in Vehicular Networks

Figure 2 showcases a vehicular network scenario, where vehicles communicate under a single roadside unit (RSU) assistance and optimize real-time power control by utilizing a centralized training and distributed execution DRL-based framework [9]. However, designing an optimal set of input states for the DRL agent is not straightforward, requiring a careful selection of states/features relevant to the agent's action. To this end, an intuitive approach is adopted, considering the following features:

- Direct channel gains among vehicles in the current and previous time slots.
- Interference channel gains in the current and previous time slot.
- The transmit power of the vehicle transmitter in the previous time slot.
- The received signal-to-interference in the previous time slot.

The selected states yield a state space cardinality of $2K + 2$ per vehicle-to-vehicle (V2V) agent for a total of K V2V pairs in the network. The DRL agent's ultimate goal is to optimize its action (i.e., transmit power) based on these input states to maximize its long-term reward. To define the reward function, we incorporate unsatisfied constraints of transmit power, reliability, and latency as penalties into the worst-case decoding error probability of the vehicular network. The resulting problem's complexity increases with the size of the state space, posing a significant challenge in developing a scalable network architecture and robust algorithm design.

To decrease the complexity of the agent network by reducing the input size for the DRL agent model, we devise an input feature selection strategy utilizing the SHAP-based feature importance ranking. Figure 2 presents the workflow of the proposed method, where a pre-trained DRL agent approximates the optimal power allocation based on the wireless environment observations as input. The DRL input states and corresponding

With the ever-increasing adoption of AI-based models for RRM, it is certain that attackers will increasingly seek out new methods of attacking the underlying AI-empowered systems, which has become a critical challenge in many real-world applications. Attackers can deliberately design input data to manipulate the model output and produce inaccurate decisions, posing a significant risk to the integrity of AI models.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

The SHAP values are utilized to rank the input states/features of the DRL agent based on their global contribution to the agent output, that is, the transmit power.

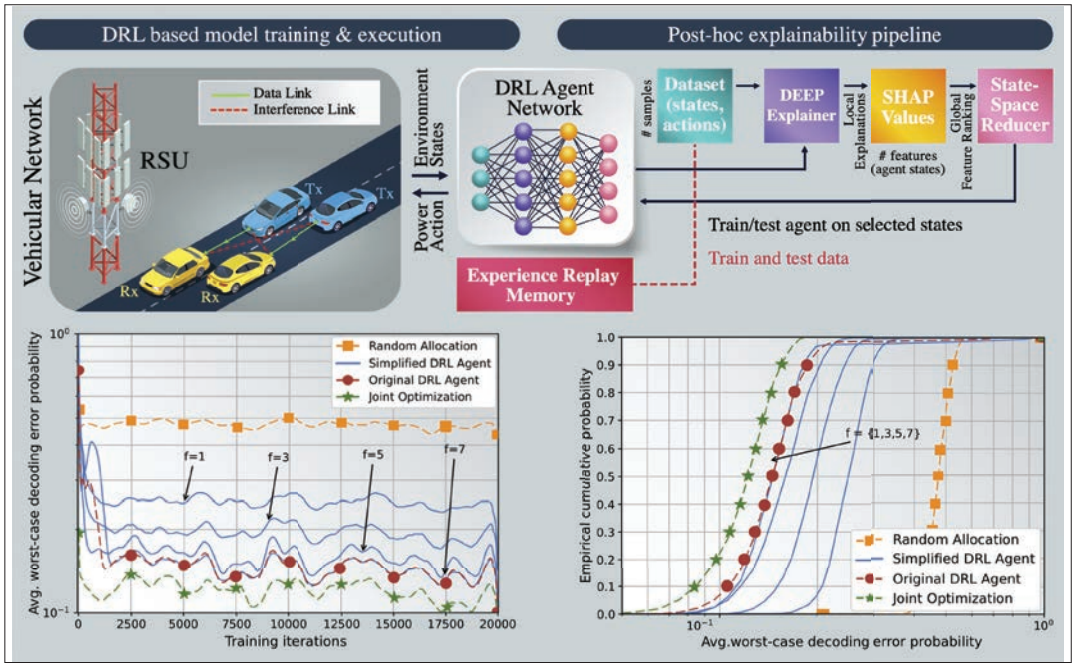


FIGURE 2. XAI-based model simplification using post-hoc feature importance ranking for DRL-based RRM in vehicular networks.

actions are collected from the simulated environment and stored in the DRL agent's replay memory, which are then utilized to create the testing and training datasets for post-hoc explainability. The Deep-Explainer, an explainability technique built upon DeepLIFT [10], is used to approximate the SHAP values for the different input states of the DRL agent. One major obstacle associated with the Deep-Explainer based feature importance scoring is the considerable computational complexity, which increases exponentially with the number of features, and linearly with the number of background data samples required for approximating the SHAP values [10]. To strike a balance between achieving high accuracy in SHAP value estimates and keeping computation time reasonable, we limit the background data size to 20 percent of the initial dataset size, which is sufficient to accurately assign feature importance to different input states without incurring the full computational burden of calculating SHAP values for all possible feature combinations and background data samples.

The SHAP values are utilized to rank the input states/features of the DRL agent based on their global contribution to the agent output, that is, the transmit power. We then select the f most important features for simplifying the model and retrain it using the selected input states.

Figure 2 also presents the training and testing results of the DRL agent based on different numbers of states utilized to optimize the power allocation policy. For a vehicular network with $K = 10$ vehicular pairs, the simplified DRL agent can remarkably achieve the same convergence rate as the original model utilizing all 22 input states by selecting the first seven most important ones, significantly reducing the model complexity. The simplified DRL agent effectively reduces the number of input states by $\sim$ 68 percent, the number of training parameters to broadcast by $\sim$ 81.5 percent, and the average training time by $\sim$ 49.1 percent , while maintaining the performance gap (in absolute values) less than one percent.

The proposed explainability pipeline is model-agnostic and applicable to any other type of pre-trained AI-based RRM system. Furthermore, the DRL agent is trained using static states of the environment, and the post-training inference process is fast enough so that the environment does not change in the meantime. It is noteworthy that the state selection process can be automated to comply with the XAI-empowered zero-touch orchestration and management framework depicted in Fig. 1, where XAI inferences in the form of SHAP scores are used for variable state selection based on repeated state randomization and model evaluations. The idea is to select the state inputs with the highest scores out of the available inputs and re-evaluate the model progressively discarding the least important inputs in an automated fashion without manual intervention.

DL-Based Robust Beam Alignment for MIMO Networks

We consider the problem of beam alignment in mmWave MIMO system using beamforming codebooks for both initial access and data transmission, as shown in Fig. 3. We propose a DL-based beam alignment method that uses the sweeping beam measurements of a sensing codebook to predict the optimal narrow beam. The proposed approach reduces beam sweeping overhead by using a compact sensing beams codebook, avoiding exhaustive search in the oversampled discrete Fourier Transform (DFT) codebook. Convolutional neural networks (CNNs) are employed to develop robust and interpretable learning-based inference and decision strategies. The small-sized probing beam codebook is generated using the DRL framework of [11] based on received signal strength indicators (RSSIs). In the simulations, we consider the outdoor scenario "O1_60" from the DeepMIMO dataset [12] to generate 10,000 channel realizations, based on which we compute RSSIs over $B = 16$ sensing beams. For each channel realization, we also find the index of the best narrow beam with the highest RSSI out of 128 DFT codewords. The generated dataset is further split into training, testing, and calibration

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

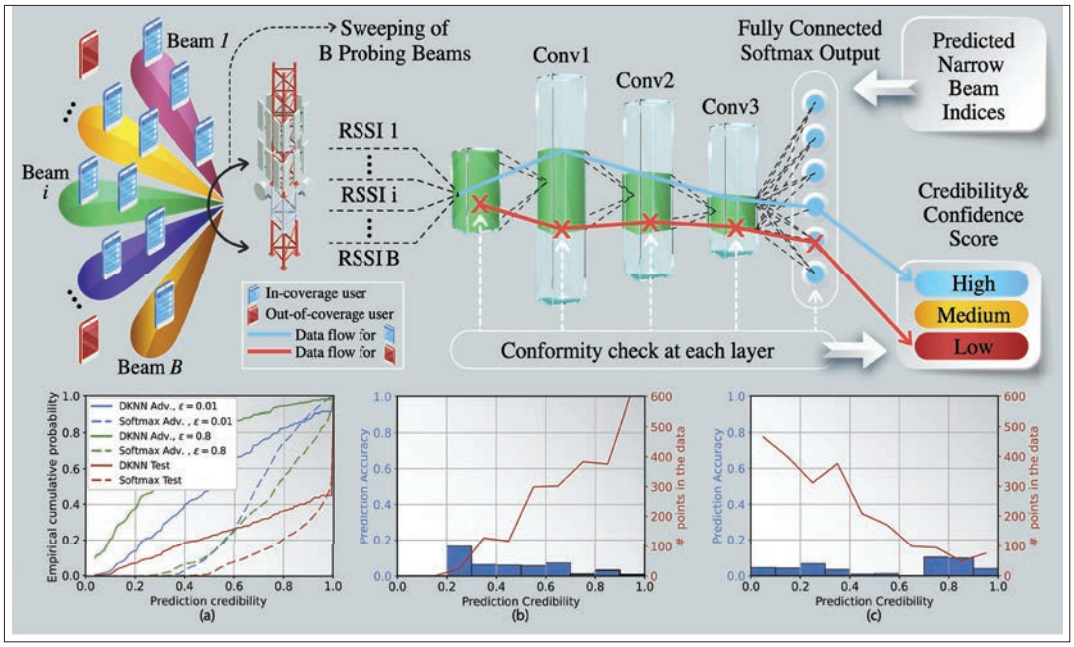


FIGURE 3. A DkNN-based credibility assessment of a DRL-based beam alignment for mmWave MIMO systems.

Determining the confidence of a CNN can be challenging due to the lack of calibration in the softmax layer's output probabilities. The softmax layer often overestimates confidence when making predictions on out-of-distribution inputs.

subsets using a 70:20:10 ratio. The proposed CNN takes $B = 16$ sensing beam RSSIs and the best beam indices as input and output, respectively.

Determining the confidence of a CNN can be challenging due to the lack of calibration in the softmax layer's output probabilities. The softmax layer often overestimates confidence when making predictions on out-of-distribution inputs. As a remedy, we utilize the Deep k-Nearest Neighbors (DkNN) algorithm from [13], which incorporates concepts from conformal prediction to measure the discrepancy between labeled inputs and past observations, providing a more reliable measure of confidence. The DkNN algorithm is model-agnostic and can be applied to any pre-trained DL model, with manageable computational complexity, especially for models with shallow hidden layers as our 3-layered CNN.

By leveraging DkNN, we develop a method of making predictions with an associated measure of prediction credibility, which quantifies the appropriateness of the training data for making the predictions. To evaluate the model's resilience against adversarial/outlier inputs, we generate the out-of-training data (adversarial) examples by the Fast Gradient Sign Method (FGSM) [13] using different values of input perturbations, that is, $\varepsilon = \{0.01, 0.8\}$ as shown in Fig. 3a, and assess the robustness of the DkNN-based algorithm by comparing it with the softmax-DNN on a separate calibration dataset. The prediction credibility score is defined as the fraction of nonconformity measures for the calibration data that are larger than that of test data, serving as a quantifiable metric for trusting the DkNN-based algorithm's decision-making process. In contrast, the softmax-DNN utilizes the softmax layer as a proxy for DNN confidence.

Figure 3 showcases our DkNN-based credibility assessment framework for DRL-based beam alignment in mmWave MIMO systems. Figure 3a illustrates the empirical cumulative probability for the DkNN and softmax-DNN against the prediction credibility for different values of input perturbations. Obviously, softmax-DNN lacks calibration as it consistently exhibits high confidence on test and adversarial data, making it ineffective in identifying outliers. Figures 3b and c compare the credibility of softmax-DNN and DkNN on a calibration dataset. While the reliability diagrams are histograms presenting accuracy as a function of credibility estimates, the red line illustrates data density across bins. Figure 3b shows that softmax-DNN misclassifies most adversarial examples with confidence above 0.8. On the other hand, Fig. 3c highlights DkNN's capability to assign low credibility to outliers, indicating a lack of support from the training data. In terms of the number of adversarial beam indices assigned credibility below a certain threshold, the DkNN algorithm improves the robustness by a factor of ~ 2.9 and ~ 17.2 for credibility thresholds of 0.4 and 0.2, respectively.

Conclusions and Future Research Directions

In this article, we present an overview of the explainable and robust AI techniques for RRM. We explain how XAI methods can provide a systematic methodology for interpreting the decisions made by the black-box AI models, and improve the robustness of the decisions and performance of the algorithms by reducing the model complexity and convergence time. Besides, we outline the core explainability and robustness techniques, including feature visualization, feature attribution, model simplification, sensitivity analysis, and model compression. We also provide two practical case studies that illustrate the application of these techniques for model simplification and improving the robustness of RRM. The applied XAI techniques for achieving model simplification and robustness are based on post-hoc explainability and are designed to be model-agnostic. Albeit significant advancements in XAI systems, there are still many obstacles to overcome in the context of XAI-based RRM including lack of transferable XAI models, performance and explainability trade-off, lack of standardization, real-time constraints and privacy concerns of XAI-based models, for which

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

The applied XAI techniques for achieving model simplification and robustness are based on post-hoc explainability and are designed to be model-agnostic.

we highlight potential directions in the sequel.

Transferability of the Post-Hoc Explainability

While transfer learning-based approaches have shown good performance in training based on small datasets using pre-trained models for similar RRM problems, the transfer of knowledge (criterion for selection of layers/ learned weights) from the pre-trained source model to build the target model has not been thoroughly explored [14]. Leveraging insights obtained through XAI can assist in reapplying the solution to several AI-empowered applications. Post-hoc explainability techniques that simplify models are usually closely linked to the specific ML model and network architecture used. There is a need for generalized XAI techniques that can be applied to a variety of RRM-related AI/ML systems and applications, as well as evaluation metrics to quantify how faithfully a given explanation mimics the behavior of the underlying model.

Twin Systems for Performance and Reasoning

There is a need for intelligent XAI-enabled systems that can capture the current wireless environment for decision-making while also being capable of reasoning and executing autonomous decisions. Network architectures need to be designed to incorporate explainable AI into wireless resource management systems in the near future, which can be achieved by developing a parallel/twin XAI system working alongside the primary AI system. This low-complexity XAI-based twin model can then replace the primary AI system over time to enhance the performance and trustworthiness of RRM algorithms.

Defence Mechanisms

Post-hoc explanation techniques such as LIME and SHAP themselves are prone to adversarial attacks and can easily be fooled by adversarial classifiers. Adversarial machine learning has successfully provided robustness and resilience against adversarial attacks. Therefore, it is crucial to define and incorporate similar defense mechanisms in XAI-enabled RRM solutions for protection against adversarial attacks.

Real-Life Assessments

Since XAI methods currently envisioned for RRM problems are evaluated in simulated or controlled environments, their performance may not reflect their efficacy in real-world environments. XAI-based models must be rigorously tested in practical wireless environments to facilitate real-time validation of different AI-based RRM solutions. Additionally, existing XAI-empowered algorithms lack automation in their inference process and suffer from performance-explainability trade-off. Striking the right balance between explainability versus performance and minimizing manual intervention is a cornerstone for driving autonomous network functions.

XAI Integration in Open Radio Access Networks (O-RAN)

AI/ML algorithms in O-RAN framework are implemented as external applications, that is, xApps, deployed on the near real-time RAN Intelligent Controller (RIC). To reveal AI decision logic and harness such insights to optimize specific components of the AI lifecycle in O-RAN environments, it is important to integrate XAI techniques into the existing architectures [15]. In this regard, XAI twin systems can act as a fundamental enabler for O-RAN environments in future 6G wireless networks.

In conclusion, XAI for RRM is still in its nascent stages. Nevertheless, the insights provided in this article can serve as a fundamental guide for the gradual enhancement of XAI-based solutions for RRM.

Acknowledgment

This research is partially funded by the Scientific and Technological Research Council of Turkey-Ford Otosan Grant #119C058 and KAUST CRG Grant #ORA-2021-CRG10-4704.

References

[1] T. Zhang *et al.*, "Interpreting AI for Networking: Where We Are and Where We Are Going," *IEEE Commun. Mag.*, vol. 60, no. 2, 2022, pp. 25–31.
[2] A.-D. Marcu *et al.*, "Explainable Artificial Intelligence for Energy-Efficient Radio Resource Management," *Proc. IEEE WCNC*, 2023, pp. 1–6.
[3] W. Guo, "Partially Explainable Big Data Driven Deep Reinforcement Learning for Green 5G UAV," *Proc. IEEE ICC*, 2020, pp. 1–7.
[4] Guo *et al.*, "Explainable Artificial Intelligence for 6G: Improving Trust Between Human and Machine," *IEEE Commun. Mag.*, vol. 58, no. 6, 2020, pp. 39–45.
[5] Y. Zheng *et al.*, "Enabling Robust DRL-Driven Networking Systems via Teacher-Student Learning," *IEEE JSAC*, vol. 40, no. 1, 2022, pp. 376–92.
[6] M. Sabih *et al.*, "Utilizing Explainable AI for Quantization and Pruning of Deep Neural Networks," arXiv:2008.09072, 2020.
[7] Z. Du *et al.*, "Green Deep Reinforcement Learning for Radio Resource Management: Architecture, Algorithm Compression, and Challenges," *IEEE Veh. Technol. Mag.*, vol. 16, no. 1, 2021, pp. 29–39.
[8] Y. Shi *et al.*, "How to Attack and Defend Nextg Radio Access Network Slicing With Reinforcement Learning," *IEEE Open J. Veh. Technol.*, vol. 4, 2022, pp. 181–92.
[9] N. Khan *et al.*, "Multi-Agent Reinforcement Learning Based Joint Resource Allocation for Ultra-Reliable Low-Latency Vehicular Communication," 2022; available: https://doi.org/10.36227/techrxiv.21685046.v1.
[10] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Adv. Neural Inf. Process. Syst.* 30, 2017, pp. 4765–74.
[11] A. Abdallah, A. Celik, M. M. Mansour and A. M. Eltawil, "Multi-Agent Deep Reinforcement Learning for Beam Codebook Design in RIS-aided Systems," in IEEE Transactions on Wireless Communications, doi: 10.1109/TWC.2023.3347419.
[12] A. Alkhateeb, "DeepMIMO: A Generic Deep Learning Dataset for Millimeter Wave and Massive MIMO Applications," *Proc. Info. Theory and Appl. Workshop*, Feb. 2019, pp. 1–8.
[13] N. Papernot *et al.*, "Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning," arXiv:1803.04765, 2018.
[14] A. Thantharate and C. Beard, "Adaptive6G: Adaptive Resource Management for Network Slicing Architectures in Current 5G and Future 6G Systems," *J. Network and Systems Management*, vol. 31, no. 1, 2023, p. 9.
[15] B. Brik *et al.*, "A Survey on Explainable AI for 6G O-RAN: Architecture, Use Cases, Challenges and Research Directions," arXiv preprint arXiv:2307.00319, 2023.

Biographies

Nasir Khan (nkhan20@ku.edu.tr) is currently pursuing his Ph.D. in electrical and electronics engineering at Koc University, Istanbul, Turkey.

Sinem Coleri received her Ph.D. in electrical and computer sciences from the University of California, Berkeley, in 2005. She is currently a Full Professor at Koc University.

Asmaa Abdallah received her Ph.D. in electrical and computer engineering from the American University of Beirut, Beirut, Lebanon, in 2020. She is currently a post-doctoral fellow at King Abdullah University of Science and Technology (KAUST).

Abdulkadir Celik received his Ph.D. in electrical and computer engineering from Iowa State University, Ames, Iowa, in 2016. He is currently a Research Scientist at KAUST.

Ahmed M. Eltawil received his Ph.D. in electrical engineering from the University of California, Los Angeles, in 2003. He is currently a Full Professor at KAUST.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/