2024 IEEE International Conference on Machine Learning for Communication and Networking (ICMLCN) | 979-8-3503-4319-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICMLCN59089.2024.10624980

# Exploring Explainable AI Techniques for Radio Frequency Machine Learning

Stephen Adams*, Mia Taylor†, Cody Crofford†, Scott Harper†, Whitney Batchelor†, William C. Headley*
*Virginia Tech National Security Institute and †Graf Research Corporation

***Abstract*—Deep learning models are increasingly being used to solve complex wireless radio frequency communications problems. These state-of-the-art machine learning models have demonstrated superior performance over traditional methods when signal and environmental parameters are unknown a priori. However, due to the complexity of the architecture and the number of parameters, deep learning models are difficult to interpret. This opacity can lead to difficulties during testing and a lack of trust by the user. Explainable artificial intelligence (XAI) techniques can provide estimates for the impact an input has on the output of a model. In this study, we apply a wide range of common attribution techniques, a subset of XAI that focuses on estimating the contribution of each input to an output of a model, to simple wireless communications problems over two different data modalities (raw IQ and spectrogram images) and show how estimates of attributions could be used for test and evaluation.**

***Index Terms*—explainability, attribution, machine learning, radio frequency**

## I. Introduction

In recent years, deep learning has increasingly been investigated for solving complex wireless radio frequency (RF) communications problems. The reasoning for this is two-fold. First, there has been a significant push towards cognitive radios, as well as 6G and Next-G systems [1], that can intelligently and proactively utilize increasingly congested RF spectrum resources (typically termed Dynamic Spectrum Access [2]). To achieve this, these systems must be able to accurately sense their spectral environment (be it through signal detection, signal parameter estimation, and/or signal identification). Given the multitude of real-world propagation effects, transmitter/receiver effects, and the large number of possible signal formats, among other effects, traditional approaches to solving these sensing problems do not generalize well, and thus researchers have turned to deep learning. Secondly, in military communications, there is an inherent lack of prior knowledge of the adversaries expected signals, environments, effects, etc. that make these traditional approaches fail [3]–[5].

Testing and evaluating (T&E) black-box machine learning models, such as the deep neural networks becoming increasingly popular in wireless communication, can be difficult due to the lack of interpretable parameters. Explainable AI (XAI) [6] can provide insight into the workings of these black-box models by 1) estimating the attributions of input feature to the model's output, 2) providing counterfactual examples that highlight the minimal information needed to change the output of a model, and 3) generating text that describe an explanation of the output. Global techniques construct a surrogate model that can be interpreted by users that approximates the general behavior of black-box machine learning model [6]–[9]. Similarly, local techniques construct surrogate models to understand the decision process of the black-box model for a specific observation [6]. Local interpretable model-agnostic explanations (LIME) is a common local XAI technique [10]. Generally, presenting explanations of black-box models to a user can also increase the user's trust in the model [11].

We believe that XAI techniques could be used during the T&E process of RF machine learning (RFML) models. Furthermore, we believe that XAI techniques could be used to improve users' trust in RFML models when deployed. However, to the best of our knowledge, XAI techniques have not been studied in the context of wireless communication and RFML models. The contribution of this study is two fold. First, we evaluate common XAI attribution techniques to RFML models and provide a demonstration of their usefulness for T&E of RFML models. Second, we link XAI techniques across modalities of common data types in the communications space, namely raw IQ data and spectrogram data. The attribution techniques utilized in this study are well established in the XAI field; however, they have primarily been implemented on computer vision and natural language tasks. One objective of this study is to evaluate these techniques on classification tasks in the communications space. Implementations of the attribution techniques that are compatible with PyTorch models are available through the Captum library [12].

This paper is organized as follows. Section II describes the synthetic data generation used in the numerical experiments. Section III briefly outlines the XAI techniques used in this study. Section IV outlines the numerical experiments, and Section V provides our conclusions.

## II. RF Data Generation

The first step in our RF data generation process is to define the range of parameters over which randomly generated synthetic datasets will vary for T&E. The set of adjustable parameters is listed in Table I. Given our focus on exploring XAI with respect to classifying the modulation format of RF signals in the spectral environment, these parameters are primarily focused on varying and generalizing the parameters

This material is based upon work supported by the Air Force Research Laboratory, Rome NY under Contract No. FA8750-22-C-0523. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Air Force Research Laboratory. Approved for release; distribution unlimited: Case Number: AFRL-2023-4852 20230929.

TABLE I: RF Dataset Adjustable Parameter Set.

| Parameter Type | Options |
|---|---|
| **Signal Modulation Format** | Amplitude Shift Keying<br>Phase Shift Keying<br>Quadrature Shift Keying<br>Frequency Modulated<br>Carrier Wave |
| **Pulse Shaping** | Square<br>Raised-Cosine<br>Root-Raised-Cosine<br>Filter Roll-off<br>Filter Span |
| **Signal Extent** | Center Frequency<br>Bandwidth<br>Start Time<br>Duration |
| **Channel Propagation** | Signal-to-Noise Ratio |
| **Aggregate Spectrum** | Total Observation Time<br>Max Number of Signals<br>Noise Floor<br>Signal Collision Flag |
| **Dataset** | Number of Training Examples<br>Number of Testing Examples |

of the signals such as their frequency extent (i.e. center frequency and bandwidth), their time extent (i.e. start time and duration), and their format (i.e. modulation format and pulse shaping). Additionally, the user can define parameters of the propagation environment such as the noise floor, signal-to-noise (SNR) ratios of each signal, and the expected number of neighboring signals.

The adjustable parameters can either be selected by the user or chosen stochastically for each training and evaluation example. The ground truth for each example is in the form of a metadata "burst list." A visual representation of an example burst list can be seen in Figure 1, and the spectrogram of the simulated RF data generated from this burst list can be seen in Figure 2 (with the burst metadata superimposed for reference).

Spectrogram images have demonstrated superior performance for solving signal detection problems, particularly for higher SNR environments. Additionally, spectrogram images can be used for the classification of signals that have distinct frequency properties. Therefore, the data generation tool provides the option of generating spectrogram images where the ground truth for each example image is a modified burst list representing the extent of the signals (time/frequency) with respect to image pixels. In Figure 2, note that visually there is a clear distinction of the frequency modulated continuous wave (FMCW) signals from the linear digital amplitude-phase modulation (LDAPM) signals (e.g., phase shift keying and quadrature amplitude modulation) that can be learned by RFML algorithms and the extent of the signals in time frequency can be clearly seen for high signals (red-orange in color). However, traditionally, these visual approaches have poorer performance for low SNR (yellow-green in color) signals or for classifying the type of LDAPM signal (the rectangular looking signals appear identical in the spectrogram).

In cases where the spectrogram images fail (due primarily

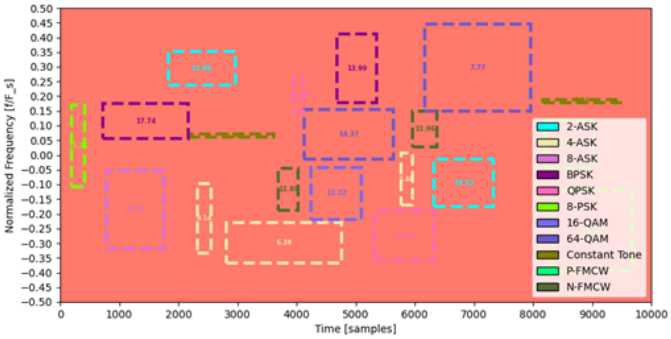


Fig. 1: Visual representation of a burst list.

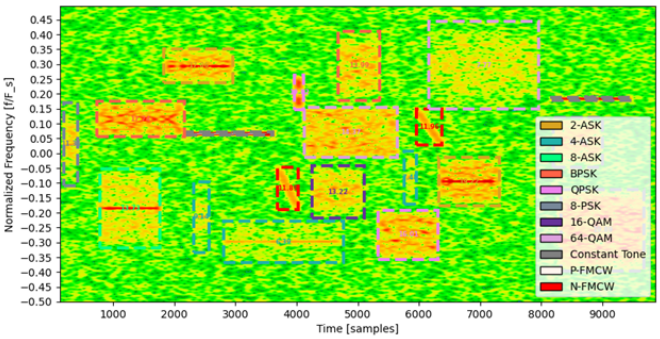


Fig. 2: Spectrogram of the burst list with the metadata overlaid to visually validate correct generation.

to loss of phase information), leveraging the raw IQ data can make up for these shortcomings as it contains all the information of the received signal. In these cases, the raw IQ data is typically used. While IQ-based models have been used to detect low SNR signals and more accurately classify signals, they have the limitation of not being well-suited to bounding the extent of signals in wideband data. Therefore, an approach that merges image and raw data analysis is likely to provide the best solution for many RFML applications. Given this, for this work we primarily investigate image-based machine learning for frequency-based modulation classification in RF spectrum consisting of a multitude of signals and RF-based machine learning for more general modulation classification in RF spectrum consisting of a single signal.

## III. XAI Methods

Many attribution XAI techniques are implemented in the Captum Python library [12]. Generally, attribution techniques can be divided into two categories. Perturbation-based methods augment the input and observe the effect on the output, and gradient- or backpropagation-based techniques use the gradients generated by the deep learning model. Both categories require a baseline input for comparison when estimating the attribution. Table II contains the methods used in this study and a brief description of each method. For a more detailed description of the techniques, see the Captum documentation and the publications cited in Table II.

Attribution techniques assign weight to each input of a model. The IQ data consists of time-series data so each

TABLE II: XAI Methods

| Method | Type | Description |
|---|---|---|
| Integrated Gradients [13] | Gradient-based | Creates points between the input and the baseline and then calculates attributions using the integral of gradients over the set of points. |
| Guided GradCAM [14] | Gradient-based | Combines guided backpropagation with gradient-weighted class activation maps (Grad-CAM). |
| Saliency [15] | Gradient-based | Estimates attributions by comparing the gradient of the input with the gradient of the baseline. |
| Deconvoluation [16] | Gradient-based | Similar to saliency but treats the gradients in the ReLU layer differently. |
| Deep Lift [17] | Gradient-Based | Estimates attributions by comparing the activations of the input at different layers of the network with the activations of the baseline. |
| Guided backpropagation [18] | Gradient-based | Estimates attributions using the gradient calculated with respect to the output. |
| Input $\times$ gradient [19] | Gradient-based | Estimates attributions by multiplying the gradient by the input. |
| Feature ablation | Perturbation-based | Estimates attributions by replacing a component of the input with the corresponding component in the baseline. |
| Feature permutation | Perturbation-based | Estimates attributions by permuting components of the input. |
| Kernel SHAP [20] | Perturbation-based | Estimates attributions by combining LIME [10] with Shapley values. |
| Shapley value sampling [21] | Perturbation-based | Estimates attributions by permuting the input and adding to the baseline. |

observation in a signal is assigned an attribution value. Positive attribution represents evidence that the observation contributed positively to a prediction. Negative attribution represents evidence against a prediction. We define the total attribution of a signal as the sum of the attributions over the length of the signal. Therefore, each signal can be represented by a single total attribution for each component of the input (the real and imaginary parts of the signal). The application of the attribution techniques to the spectrogram images is straight forward because of the similarity to computer vision tasks.

## IV. Numerical Experiments

This section outlines the numerical experiments performed on the raw IQ data and the spectrogram images in this study.

### A. IQ Numerical Experiments

A binary modulation classification problem between binary phase shift keying (BPSK) and quadrature phase shift keying (QPSK) is selected because these modulation classes are able to be distinguished by information on the imaginary part of the signal (under the assumption of time/frequency/phase synchronization for this introductory example). Specifically, QPSK has information on the imaginary part of the signal and BPSK does not. GradCAM is not applied here because it was designed and implemented for convolution neural networks.

*1) Model:* The binary classifier has a single gated recurrent unit (GRU) layer with a hidden state size of 5. The output of the GRU layer is passed through a sigmoid activation function, then a linear layer and a softmax function are used to map to two output nodes representing the classes. Table III contains the data generation and model training parameters. The training set contains an equal number of observations for each modulation type. The trained model produces a perfect classification accuracy on a validation set.

*2) XAI:* Generally, the mean of the total attribution (Table IV) for the real part of the signal for both BPSK and QPSK is near zero except for Kernel Shap on QPSK (for visual representations of the total attributions see Figures 3 and 4). This indicates that the real part of the signal does not contribute to modulation classification for this model. For the gradient-based methods, the mean of the total attributions on the imaginary part of the signal is close to zero for BPSK. However, the mean of the total attributions on the imaginary part of the signal using Integrated Gradients is significantly greater than the other attribution techniques for QPSK. This indicates that information in the imaginary part of the signal is being used to classify QPSK. Shapley Value Sampling and Kernel Shap also have an average total attribution on the imaginary part of the signal for QPSK that is significantly higher than the other perturbation methods. This conclusion aligns with our hypothesis that the model should be keying on the imaginary part of the signal for this classification problem.

TABLE III: Data generation and model training parameters.

| | Parameter | Value |
|---|---|---|
| **Data Generation** | Number of Training Observations | 1000000 |
| | Number of Validation Observations | 200000 |
| | Observation Sequence Length | 250 |
| | Center Frequency | 0 |
| | Bandwidth | 2 |
| | Signal-to-Noise Ratio | 5 |
| **Model Training** | Learning Rate | 0.05 |
| | Training Epochs | 500 |
| | Batch Size | 50 |
| | Loss Function | Cross Entropy |

There are numerous sources of randomness in this set of numerical experiments - data generation, model parameter initialization, and model training. Figure 5 displays the total attributions using Integrated Gradients on a different model with the same architecture, which also has perfect classification accuracy on a validation set. The training data set is different, but the training data was generated using the same parameters. The total attributions appear to be the mirror image of the prior results. There is strong positive attribution on the real component and strong negative attribution on the imaginary component for BPSK. These two models are converging to different points in the parameter space but have different

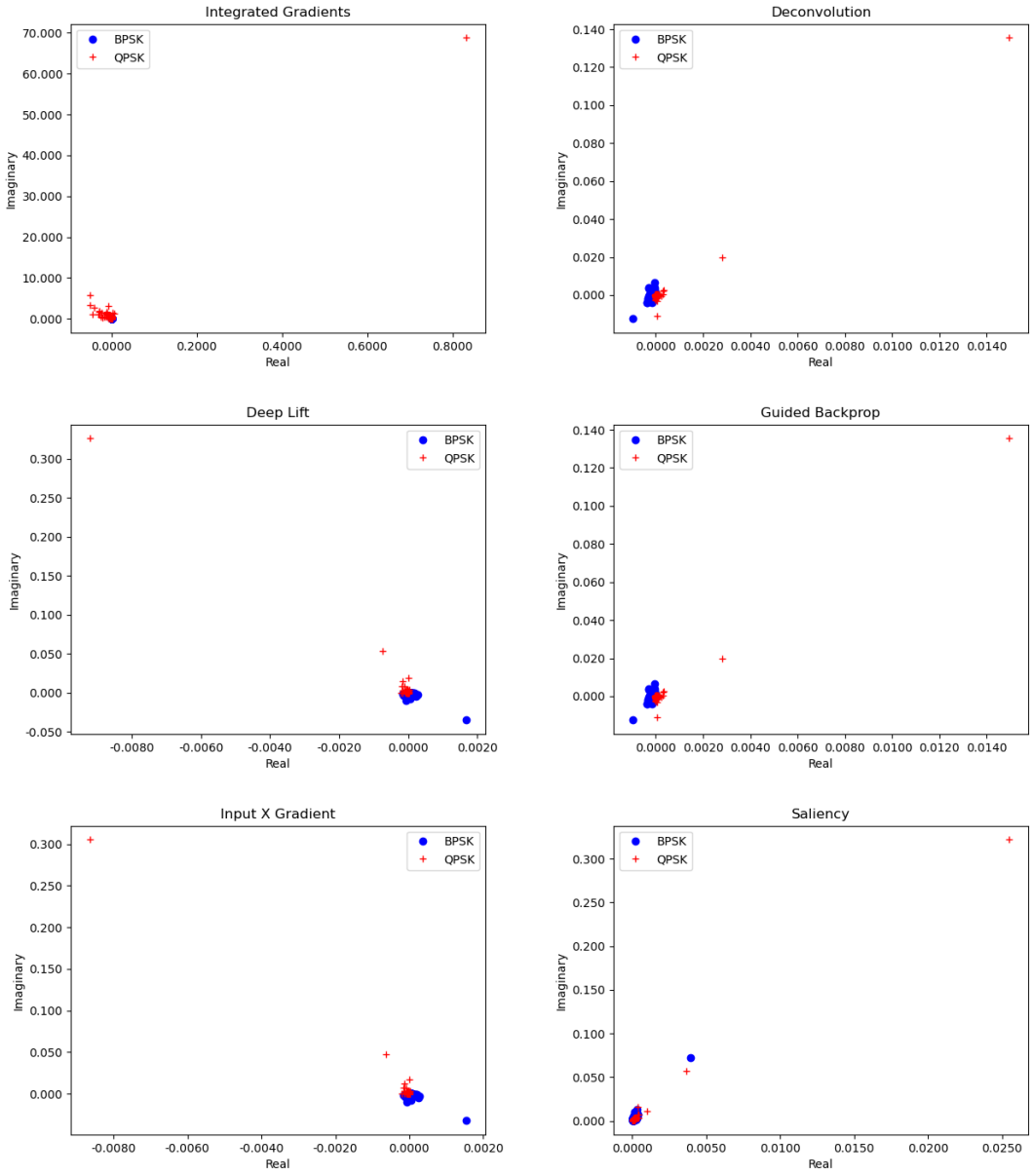


Fig. 3: Gradient-based attributions for IQ data.

TABLE IV: Mean and standard deviation of attributions for BPSK and QPSK.

| | BPSK | | | | QPSK | | | |
|---|---|---|---|---|---|---|---|---|
| **Attribution Method** | **Real Mean** | **Real Std** | **Imag Mean** | **Imag Std** | **Real Mean** | **Real Std** | **Imag Mean** | **Imag Std** |
| **Integrated Gradients** | 0.00001 | 0.00005 | -0.00027 | 0.00068 | 0.00059 | 0.04216 | 0.29230 | 0.00068 |
| **Saliency** | 0.00011 | 0.00020 | 0.00209 | 0.00377 | 0.00014 | 0.00129 | 0.00158 | 0.00377 |
| **Deep Lift** | 0.00004 | 0.00009 | -0.00087 | 0.00200 | -0.00004 | 0.00046 | 0.00159 | 0.00200 |
| **Input X Gradient** | 0.00004 | 0.00009 | -0.00086 | 0.00187 | -0.00004 | 0.00043 | 0.00143 | 0.00187 |
| **Guided Backprop** | -0.00009 | 0.00007 | 0.00025 | 0.00116 | 0.00005 | 0.00076 | 0.00026 | 0.00116 |
| **Deconvolution** | -0.00009 | 0.00007 | 0.00025 | 0.00116 | 0.00005 | 0.00076 | 0.00026 | 0.00116 |
| **Feature Ablation** | 0.00002 | 0.00004 | -0.00521 | 0.09919 | -0.00007 | 0.00112 | 0.00332 | 0.09919 |
| **Feature Permutation** | 0.00000 | 0.00006 | 0.12787 | 0.37803 | -0.00007 | 0.00137 | 0.00299 | 0.37803 |
| **Shapley Value Sampling** | -0.00008 | 0.00194 | -0.00014 | 0.06856 | -0.02147 | 0.05455 | 1.01302 | 0.06856 |
| **Kernel Shap** | -0.00014 | 0.03054 | -0.00009 | 0.03970 | 0.48202 | 0.05049 | 0.50953 | 0.03970 |

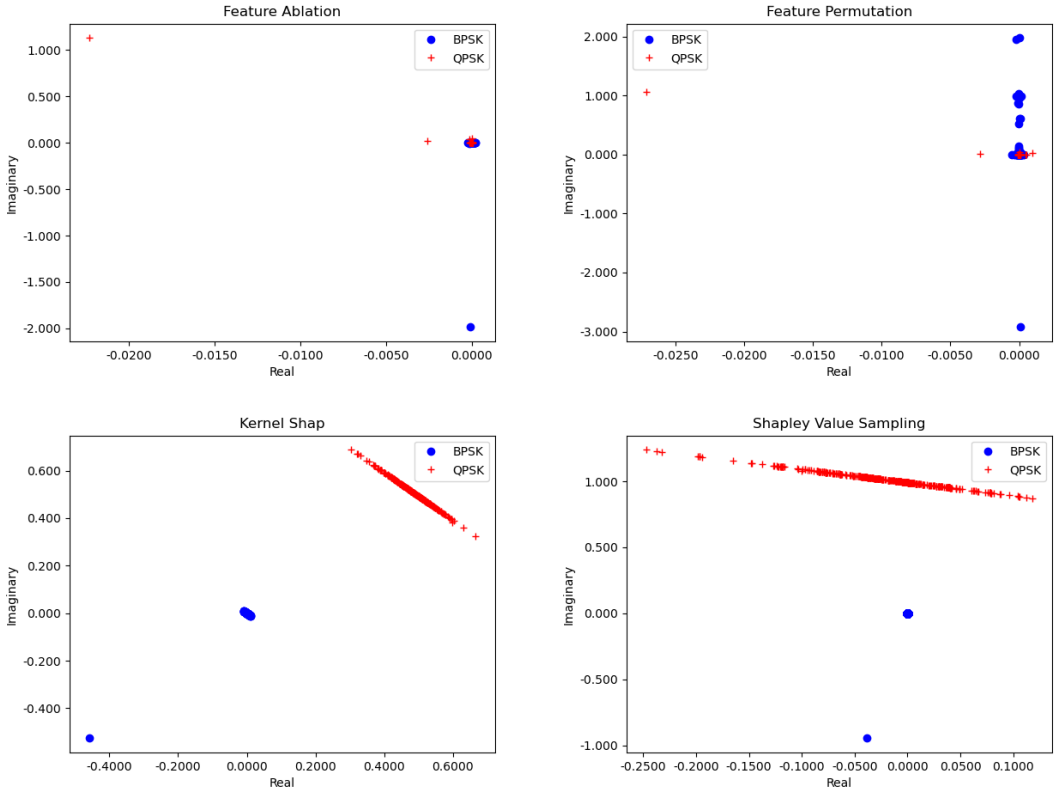


Fig. 4: Perturbation-based attributions for IQ data.

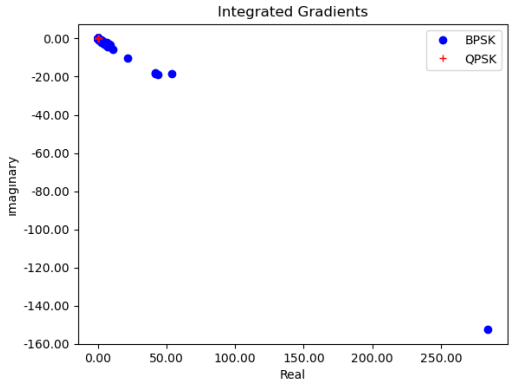


Fig. 5: Attributions using Integrated Gradients.

interpretations when using XAI techniques. This implies that multiple versions of an accurate model exist in the parameter space. This needs to be considered during T&E for these models and taken into account when presenting the results from an XAI evaluation to a user.

### *B. Spectrogram Numerical Experiments*

For this set of numerical experiments, a binary classifier is trained to recognize if an amplitude-shift keying (ASK) modulation type is present in the spectrogram.

*1) Model:* A 50-layer ResNet model [22] was used as the architecture. The ASK binary classifier was trained on a data set containing 6,538 spectrogram images where approximately half of the spectrogram images contained ASK signals. The model used 75% of the data set for training, 15% for validation, and 10% for testing. The ResNet ASK model reached a training accuracy of 96%, and a test accuracy of 93% within 6 training epochs. The ASK classifier also had a training F1 score of 0.96, and a test F1 score of 0.92.

*2) XAI:* An example spectrogram and a sample of attributions are shown in Figure 6 for the best-trained ASK classifier. Positive attributions are shown in blue. Negative attributions are shown in red. For example, the GradCam attribution uses the 4-ASK and bigger 2-ASK signals to explain why the classifier predicts the spectrogram contains an ASK signal. The Shapley Value Sampling method takes significantly longer to generate than the other methods.

The GradCAM attribution technique was concluded to be the best XAI technique for the best performing ASK classifier. Further analysis using GradCAM was conducted and it was determined that the attributions do not favor a certain type of ASK signal (2-ASK, 4-ASK, 8-ASK). As expected, high SNR ASK signals provide for more distinct visual cues for positive attribution. The FMCW and constant tone distinct features (which also have distinct "lines" in the spectrum) may confuse the classifier. These features are shown respectively by the arrows in Figure 7 where the images on the top are individual spectrograms and the images on the bottom show the GradCam attribution overlaid onto a monotone version of the spectrogram. Conversely, the following spectrogram

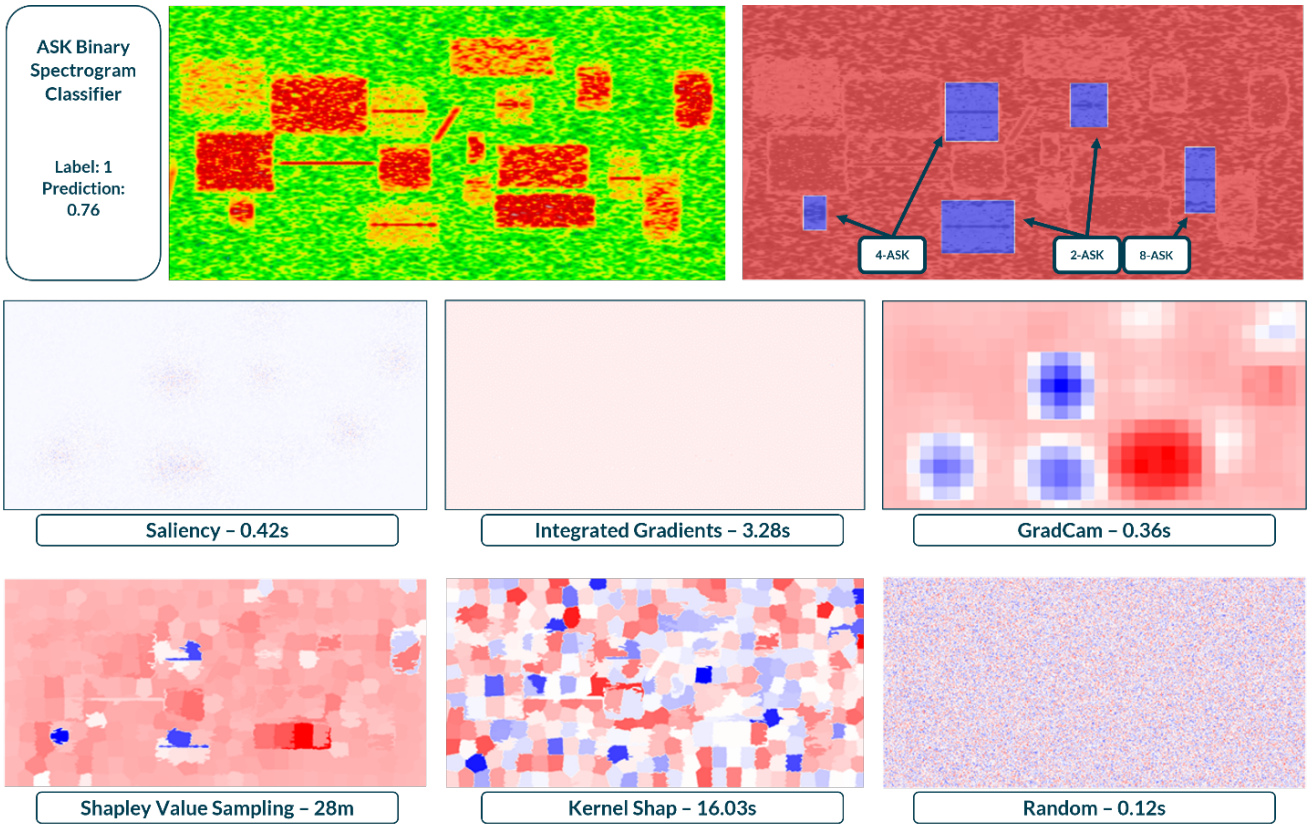


Fig. 6: The estimated attributions of the best performing binary spectrogram classifier are shown. The model predicts that the spectrogram has a 76% chance of containing an ASK signal. The spectrogram is shown at the top left. A labeled segmentation mask is shown to its right. A perfect attribution would match the segmentation mask. The bottom two rows show each attribution and calculation time.

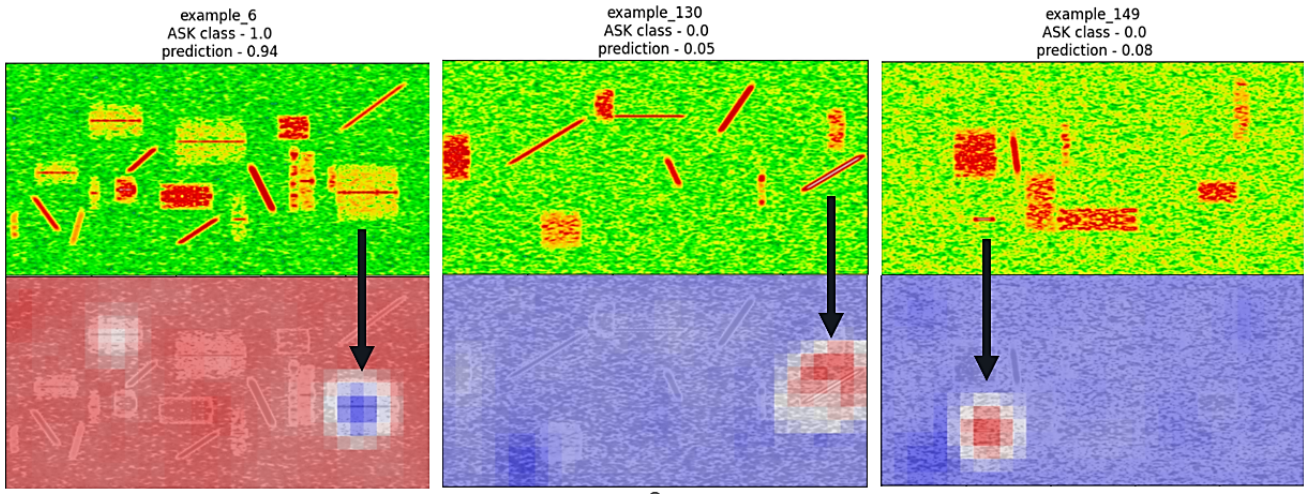


Fig. 7: Visual spectrogram features that contribute positively to an ASK classification (towards a prediction score of 1). The left example shows a high signal-to-noise ratio ASK signal. The middle example shows how a very high signal-to-noise ratio FMCW signal is associated with a higher prediction. The right example shows how a constant tone is also associated with a higher prediction.

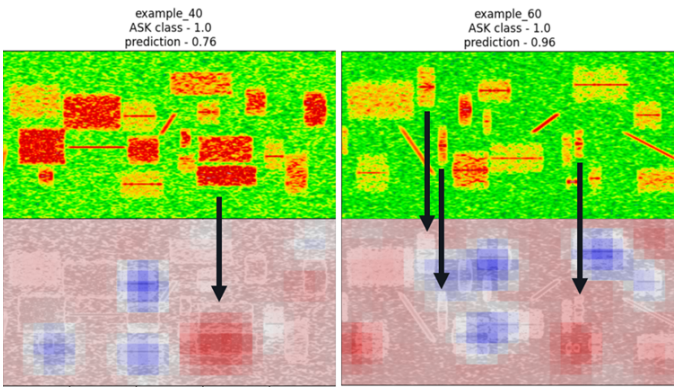


Fig. 8: Visual spectrogram features that contribute negatively to an ASK classification (towards a prediction score of 0). The left example shows that two adjacent channel non-ASK signals with roughly the same duration contribute to a negative attribution. The right example shows how short duration ASK signals are not contributing to a positive attribution.

features contribute to negative ASK classification: adjacent channel signals with roughly similar durations (where the visual cue is a green bar) and short duration signals (in terms of duration/x-axis). An example of each is shown in Figure 8. Overall, the GradCam attributions corresponded to the spectrogram classification as expected.

## V. Conclusion

This study provides the first demonstration of XAI attribution techniques to common RFML model architectures. Furthermore, we demonstrate this capability across two data types - raw IQ data and spectrogram images. Generally, we find that Integrated Gradients provide explanations that align with our prior knowledge about the simple BPSK/QPSK classification problem. Over the course of the numerical experiments, we discovered that RFML models can converge to different local minimums during the training process, and that this can be identified using attribution techniques. The interpretations of these different models can be seen by observing the mirrored patterns of the total attributions. The GradCam attribution technique was selected as the top performer for the spectrogram model. This is not surprising due to the fact that GradCam was specifically designed for computer vision tasks.

Future work at the intersection of XAI and RFML should investigate the impact of more difficult classification tasks (e.g. multi-class) and more complex model architectures. The data generation and training processes utilized in this study are stochastic in nature. The impact of the randomness should be further investigated in future studies. In particular, removing randomness should give more insight into the causes of the models converge to two different minimums. Finally, this study used a single architecture for each data modality.

## References

[1] Nokia, "6g explained," https://www.nokia.com/about-us/newsroom/articles/6g-explained/ (accessed May 19, 2023).

[2] Q. Zhao and B. M. Sadler, "A survey of dynamic spectrum access," *IEEE Signal Processing Magazine*, vol. 24, no. 3, pp. 79–89, 2007.

[3] JHUAPL, "Applications of machine learning for electronic warfare emitter identification and resource management," https://https://secwww.jhuapl.edu/techdigest/content/techdigest/pdf/V36-N02/36-02-Casterline.pdf (accessed May 19, 2023).

[4] M. Embedded, "Improving the capabilities of cognitive radar and EW systems," https://militaryembedded.com/radar-ew/rf-and-microwave/improving-the-capabilities-of-cognitive-radar-and-ew-systems (accessed May 19, 2023).

[5] A. International, "Cognitive electronic warfare: Radio frequency spectrum meets machine learning," https://interactive.aviationtoday.com/avionicsmagazine/august-september-2018/cognitive-electronic-warfare-radio-frequency-spectrum-meets-machine-learning/ (accessed May 19, 2023).

[6] R. Confalonieri, L. Coba, B. Wagner, and T. R. Besold, "A historical perspective of explainable artificial intelligence," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 1, p. e1391, 2021.

[7] Y. Coppens, K. Efthymiadis, T. Lenaerts, A. Nowé, T. Miller, R. Weber, and D. Magazzeni, "Distilling deep reinforcement learning policies in soft decision trees," in *Proceedings of the IJCAI 2019 workshop on explainable artificial intelligence*, 2019, pp. 1–6.

[8] Y. Hua, S. Ge, C. Li, Z. Luo, and X. Jin, "Distilling deep neural networks for robust classification with soft decision trees," in *2018 14th IEEE International Conference on Signal Processing (ICSP)*. IEEE, 2018, pp. 1128–1132.

[9] X. Liu, X. Wang, and S. Matwin, "Improving the interpretability of deep neural networks with knowledge distillation," in *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2018, pp. 905–912.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, ""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*, 2016, pp. 1135–1144.

[11] W. Samek and K.-R. Müller, "Towards explainable artificial intelligence," *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 5–22, 2019.

[12] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan *et al.*, "Captum: A unified and generic model interpretability library for pytorch," *arXiv preprint arXiv:2009.07896*, 2020.

[13] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[15] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: visualising image classification models and saliency maps," in *Proceedings of the International Conference on Learning Representations (ICLR)*. ICLR, 2014.

[16] A. Mahendran and A. Vedaldi, "Salient deconvolutional networks," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*. Springer, 2016, pp. 120–135.

[17] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *International conference on machine learning*. PMLR, 2017, pp. 3145–3153.

[18] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[19] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not just a black box: Learning important features through propagating activation differences," *arXiv preprint arXiv:1605.01713*, 2016.

[20] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," *Advances in neural info. processing systems*, vol. 30, 2017.

[21] J. Castro, D. Gómez, and J. Tejada, "Polynomial calculation of the shapley value based on sampling," *Computers & Operations Research*, vol. 36, no. 5, pp. 1726–1730, 2009.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.