IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 April 2024, accepted 25 May 2024, date of publication 4 June 2024, date of current version 13 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3409471

RESEARCH ARTICLE

Sensitivity Analysis of RFML Applications

BRAEDEN P. MULLER[1,2], (Member, IEEE), LAUREN J. WONG[2,3], AND ALAN J. MICHAELS[1,2], (Senior Member, IEEE)

[1]National Security Institute, Virginia Tech, Blacksburg, VA 24060, USA
[2]Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA 24060, USA
[3]AI Laboratory, Intel Corporation, Santa Clara, CA 95054, USA

Corresponding authors: Braeden P. Muller (braedenm@vt.edu) and Alan J. Michaels (ajm@vt.edu)

This work was supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA).

ABSTRACT Performance of radio frequency machine learning (RFML) models for classification tasks such as specific emitter identification (SEI) and automatic modulation classification (AMC) have improved greatly since their introduction, especially when measured against simulated data. When using captured RF data in a real environment, the performance of these RFML-based models is inconsistent when the propagation environment of the training data significantly differs from testing data. In this work, the correlations between measurable variations in propagation environment, ambient interference, amplifier compression, and overall classification performance are investigated and quantified. Parametric variations are ranked by impact to predict how well models trained in one environment can support operation in a dissimilar environment. Consistent with previous work, almost every factor studied was shown to impact classification performance in some way, with the effect of interference being particularly severe even at low levels.

INDEX TERMS Specific emitter identification (SEI), automatic modulation classification (AMC), RF fingerprinting, radio frequency machine learning (RFML).

I. INTRODUCTION

Radio Frequency Machine Learning (RFML) is the application of Machine Learning (ML) techniques to solve problems in the RF domain. ML is particularly useful for higher-dimensional problem spaces. It can excel where manually-specified (expert) features are unable to capture all useful information, or when the relationship between inputs and desired output is unclear. These considerations make RFML well suited to applications such as the spectrum sensing task of automatic modulation classification (AMC) and the source identification task of specific emitter identification (SEI).

AMC is the task of determining the modulation scheme of a received RF signal with little to no *a priori* information about the signal or channel, such as signal power, carrier frequency, phase offset and timing information [1]. This is often used as an intermediate step between signal detection and demodulation [1]. RFML is commonly used for this task, due to ML strategies historically having high classification performance [2], [3], [4], [5], [6], [7], [8], [9]. SEI is the task of distinguishing individual radio emitter identities by comparing features of their RF fingerprint, unique and unavoidable imperfections in the hardware of the RF signal chain that are orthogonal to the data being transmitted [10], [11], [12], [13]. SEI has potential applications in wireless network security for Wi-Fi, VHF, IoT, and cellular networks, cognitive radio, self-organized networking, traffic analysis, and spectrum management [10], [11], [12], [14].

For both AMC and SEI, the overwhelming majority of new approaches are partially or completely based on ML strategies [2], [15]. Models that rely on RFML techniques are often preferred since they permit correlations to observed data, which may lead to the discovery of powerful non-expert-defined features. SEI benefits, in particular, given the difficulties of modeling nonlinearities in the transmit path that are caused primarily by manufacturing variation [10]. However, these models may also learn ill effects that are

The associate editor coordinating the review of this manuscript and approving it for publication was Qilian Liang.

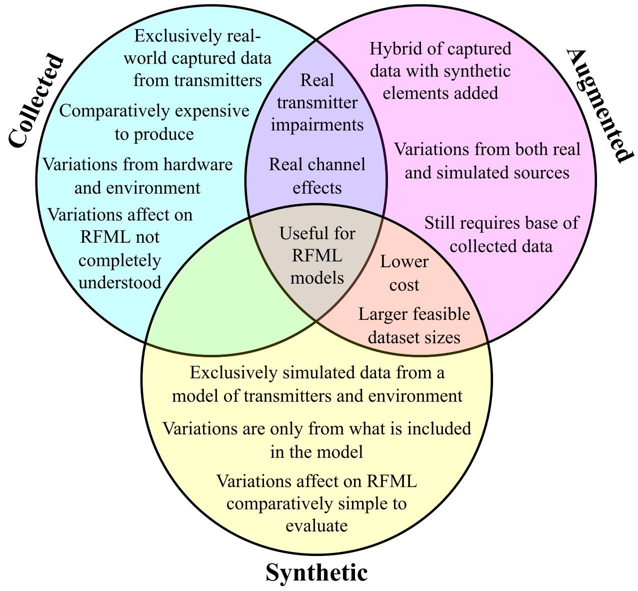
Collected
Augmented
Synthetic
Exclusively real-world captured data from transmitters
Comparatively expensive to produce
Variations from hardware and environment
Variations affect on RFML not completely understood
Hybrid of captured data with synthetic elements added
Variations from both real and simulated sources
Still requires base of collected data
Real transmitter impairments
Real channel effects
Useful for RFML models
Lower cost
Larger feasible dataset sizes
Exclusively simulated data from a model of transmitters and environment
Variations are only from what is included in the model
Variations affect on RFML comparatively simple to evaluate

FIGURE 1. A comparison of the types of data sources available for the development of RFML models: collected, synthetic, and augmented.

not necessarily applicable to general use cases, and therefore suffer severe performance penalties outside their ideal environment [16]. No model for AMC or SEI is complete, and each one makes assumptions, either implicitly or explicitly, about the hardware and propagation environment.

Channel environments can vary significantly with differences in path loss, shadowing, and multipath propagation. Attenuation due to path loss and shadowing intensifies at longer ranges or with obstacles present, while multipath propagation varies unpredictably based on the physical environment and the location of emitters within it. In addition, relative motion between transceivers can induce variance in Doppler and delay spread effects [17]. It is common for AMC and SEI models to suffer performance degradation when these channel effects deviate from the baseline seen in training or development [18], [19], [20], [21].

There are few studies on the impact of specific channel effects on model performance. One study by Al-Shawabka quantifies the impact of channel effects by analyzing existing models using transmissions data collected through a direct wired connection and over-the-air. This data was gathered in both an anechoic chamber and in a typical indoor environment [22]. Elmaghbub and Hamdaoui tested the performance of SEI models for LoRa devices in varying indoor and outdoor environments, software and hardware configurations, and physical locations [19]. Hauser et al. studies receiver distortion effects, such as errors in carrier frequency estimation and sample rate estimation [23]. Other prior sensitivity analyses are scattered across various works for AMC or SEI models that include a section evaluating sensitivity to one or more channel effects [13], [14], [21], [23], [24], [25].

When developing an RFML model, a large part of its ability to accurately classify signals depends on the quantity and quality of available data. For AMC and SEI applications, this data can come from three sources: simulation, collection, and augmentation [2], [26]. Prior work [26] has demonstrated varying levels of efficacy with training models using each of the three data types, while recent work in RF transfer learning [16] has developed quantified dataset similarity models for data collected in different RF environments. A visual comparison of these data sources is shown in Fig. 1.

Simulated or synthetic data is simple to obtain in large quantities, since it can be created quickly and inexpensively with open-source toolkits such as *GNURadio* [27] or *liquidDSP* [4]. Synthetic data is applicable for initial development and modeling of simplistic environments, but can be unsuitable as the only data source for models deployed in a real-world environment [2], [3]. Simulated data has historically proven effective for AMC [23]. For an application such as SEI, simulated data can be problematic for eventual real-world use due to the inability to accurately model non-idealities in a transmission from specific real-world hardware.

Collected or captured data from a real-world environment or physical testbed can potentially lead to superior performance, due to data being more similar in training and deployment. However, the process of collecting this data can be orders of magnitude more expensive in terms of time, effort, and hardware costs when compared to synthetic data [28]. In addition to increased cost, for SEI, collected data also carries the risk of incorrectly associating features of the channel environment with specific emitters rather than actual impairments in the hardware [3], [24]. Augmentation is an intermediate step between data collection and training where domain knowledge is used to expand the effective size of a dataset to improve model generalization and performance across more diverse scenarios [26], [29], [30], [31], [32]. Augmentation can be used to improve upon limited data to create more robust training datasets, yet has limitations particularly for controlled experiments to isolate specific channel effects.

Wide coverage of diverse channel conditions and scenario parameters in a training dataset is especially critical for RFML applications. For new observations that are on the interior of known cases, it is more likely that a correlation-based (interpolation) decision will be accurate, while new observations outside the contour of learned behaviors from known examples (extrapolation) may be more unpredictable [33], [34]. Increasing the range of conditions under which known examples are collected therefore serves to increase the likelihood that new observations lie within these learned behaviors, improving the robustness and predictability of the model.

Given the possible deployment of RFML models in military applications or otherwise congested spectral environments, the presence of non-cooperative interference is inevitable [35]. If it has not been accounted for in the development of a model, interference can significantly degrade performance. It is therefore needed to condition RFML-based decision agents to be robust against ambient interference whenever the intended application has the possibility of interference that is outside the control of the user. Co-channel interference can be especially damaging to performance, since is likely to overlap in the frequency-domain with symbols from the signal-of-interest – potentially obscuring critical features for classification and demodulation [36]. With the density of spectral environments increasing, experiencing co-channel interference is becoming increasingly likely even in commercial settings [37]. Adjacent-channel interference can also degrade performance of some RFML algorithms, but can be removed with analog filters if the center frequency and bandwidth of the signal-of-interest is known. This technique has the notable downside of invalidating RFML models that rely on out-of-band distortions for classification [38].

This work aims to identify and characterize the factors in testbed configuration and propagation environment for collected RF data that impact RFML model performance and generality across scenarios. We evaluate sensitivity to a broad set of parametric variations from a common baseline to better understand RFML algorithm performance under different training and evaluation conditions. Since propagation channels and RF signal chains can be highly diverse, the list of possible sensitivity parameters in this environment is near-inexhaustible, meaning comprehensive understanding can only be built up over time. The parametric variations of specific interest covered in this work are co-channel interference – both synthetic and real-world, transmitter high power amplifier (HPA) non-linearity, path loss, and transmitter displacement by fractions of a wavelength – referred to as micro-channel multipath variations in this work.

This paper is structured as follows. Section II-A discusses the AMC and SEI models used as baseline for experimentation. Section II presents the experimental setup for each of the variations of interest, the hardware and software configuration of the testbed collection system, and each of the data collection scenarios. Section III goes into detail about the methodology for dataset creation, RFML model training, and performance evaluation. The results of the testing, overall takeaways, and ranking of the impact of the tested parameters is given in Section IV. Finally, recommendations for future work and conclusions are discussed in Section V.

II. EXPERIMENTAL SETUP

A. BASELINE MODELS

This work aims to evaluate the relative impacts of parametric variations in the RF signal chain and in the propagation environment. Therefore, the training regimen and architecture for the RFML models are held the same across all experiments. Only the training data changes between scenarios; the model structure remains constant.

The model trained for each experiment is a Convolutional, Long-Short Term Memory (LSTM) Deep Neural Network (CLDNN) [39]. From the non-testing subset, 90% of available data from real-world captures is allocated for training, while the remaining 10% is allocated to verification. The CLDNN is trained for a maximum of 50 epochs on all allocated data or until loss during validation fails to decrease for 4 epochs. This CLDNN model was selected due to its proven success in SEI and AMC tasks across synthetic, real, and augmented datasets [13], [26], [39]. A variety of other RFML-based models have been considered for these AMC and SEI applications in related works, using Convolutional Neural Networks (CNN) [10], [40], [41], Deep Neural Networks (DNN) [42], Recurrent Neural Network (RNN) approaches like Long Short-Term Memory Networks (LSTM) [43], decision trees [44], ensemble methods such as Adaboost and Random Forest [45], and other network types [26], [46]. Our analysis [47] demonstrated that differences in performance are attributable more to the training and operational parameters than the chosen architecture. The sheer number of models to be trained in this experiment led us to choose the faster training CLDNN architecture. The structure of the CLDNN is shown in Fig. 2. This structure is most similar to the structure presented in Clark et al. [26], based on the work of Flowers et al. [48] and West and O'Shea [39].

The input to the first convolutional layer is two channels of 256 raw RF floating-point samples, one for in-phase and another for quadrature. Each convolutional layer, with 50 output channels and 1 × 8 kernel, is succeeded by a layer with rectified linear unit (ReLU) activation and then a layer of 1-dimensional batch normalization. The output of the first layer set is both fed into the second convolutional layer and saved to concatenate with the output of the third layer set as an input to the LSTM. The output of the LSTM is flattened into a vector and then fed through a ReLU activation layer with output size of 256, then through 1-dimensional batch normalization.

B. PARAMETERS

The parameters studied here are variances in multi-path propagation, (co-channel) narrowband interference, HPA non-linearity/compression effects, and SNR. A non-exhaustive table of parameters that may affect RFML performance is given in Table 1. These specific parameters were chosen to balance the anticipated impact on classification accuracy and practicality of study with a real-world collection setup. Path loss can be studied simply by varying the distance between transmitter and receiver, but the resulting attenuation is not expected to have a large impact on general classification performance at the smaller scales evaluated in this work. Multi-path, due to the high variability of distortion, especially in indoor spaces, is expected to have a large impact on classification performance for both AMC and SEI; its effect can be studied by varying the relative position of transmitter and receiver within the same indoor space. The effect of co-channel interference can be studied by either augmenting a "clean" dataset with different levels

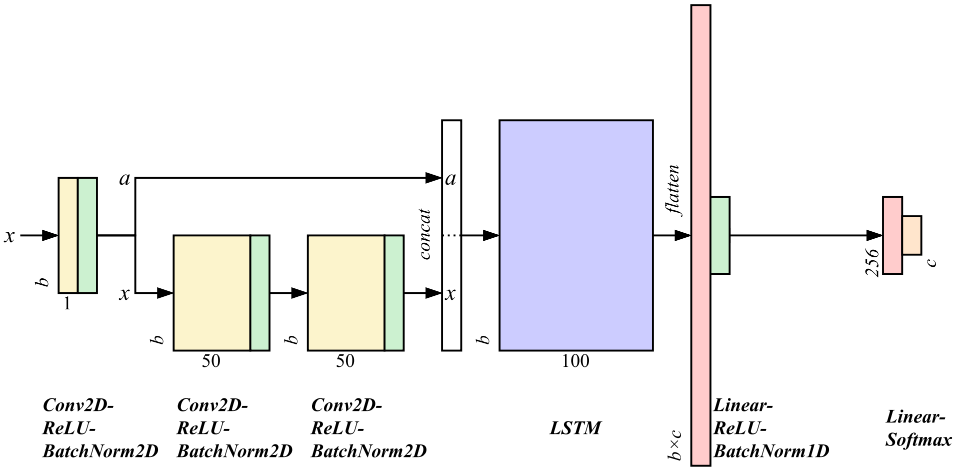

FIGURE 2. Structure of the CLDNN used for classification tasks. *b* represents the number of samples provided to the network at a time, in this case 256. *c* represents the number of classes, which is the number of modulation schemes for AMC and the number of unique radios for SEI. *x* represents the input and progression of data throughout the forwarding process and *a* represents the output of the first layer that is concatenated to the output of the third layer.

of simulated interference or by introducing an extra emitter in the collection environment. Co-channel interference may degrade classification performance by introducing energy on frequencies where a model may be expecting to only see out-of-band distortions that are consistent with a particular emitter or modulation scheme in the baseline environment. HPA compression effects are comparatively more difficult to induce and study, since manufacturers have an incentive to produce transmitters with amplifiers that can operate in the linear region across the range of supported power levels. This effect is studied by varying the power level of emissions from the device, with less compression distortion at lower power levels – decreasing the intensity of HPA compression artifacts. If compression artifacts were features learned by an SEI model, reducing these effects should result in degraded classification performance.

Dynamic channel effects, temperature, mixing and filtering errors, and clock jitter were among the parameters with higher anticipated impact that were not chosen for practicality reasons. Previous work showed that channel effects that vary from day-to-day affect classification performance, but this non-stationarity is impossible to predict or control [49]. Measuring and controlling for transmitter hardware temperature would be impractical for the number of devices used in this work [50], since it would require modification of each device with specific measuring, heating, and cooling components. The chosen parameters are discussed further in Section II-D.

TABLE 1. Table comparing various parameters of the transmitter, receiver, and channel environment that could affect AMC and SEI classification performance by anticipated impact on performance and anticipated feasibility of evaluation using captured RF data on a real-world testbed.

Parameter	Type	AMC/SEI Impact	Feasibility
Path Loss	Fading	Low / Low [17], [51], [52]	Med
Shadowing	Fading	Low / Low [17], [52], [53]	Med
Multipath Propagation	Fading	High / High [17], [30], [54]	High
Dynamic Channel Effects	Time-Variance	High / High [49], [55], [56]	Low
Background Noise / SNR	Sample Clarity	Med / Med [54], [57], [58]	High
Co-Channel Interference	Interference	High / High [57], [59]	High
Wideband Interference	Interference	Med / Med [57], [59]	High
HPA Compression Effects	TX Impairment	Low / High [21], [55], [60]	Med
HPA Temperature	TX Impairment	Low / Med [49], [55]	Med
Mixing & Filtering Errors	TX Impairment	Low / Med [19], [21], [60]	Low
Clock Jitter	TX Impairment	Low / High [21], [49], [55]	Low
TX Hardware Construction	TX Impairment	Low / High [21], [55], [60]	High
RX Hardware Construction	RX Impairment	Low / Med [55], [56], [61]	High
Receiver Phase Errors	RX Impairment	Med / Med [1], [23], [62]	Low
Receiver LNA Non-linearity	RX Impairment	Low / Med [55], [62]	Low
Sample Rate Mismatch	RX Impairment	Low / Low [7], [23], [53]	Low
Frequency Offset	RX Impairment	Med / Med [1], [23], [57]	Low

C. EXPERIMENTAL TESTBED

The experimental data collection strategy uses the Blind User Reconfigurable Platform (BURP) that is described in detail in a previous work [63]. The experimental hardware and software setup consists of a number of transmitters connected to a transmitter host machine, several co-located receivers connected to "collection node" (CN) machines, and a software control back-plane to coordinate the entire system. The system's quick reconfigurability facilitates

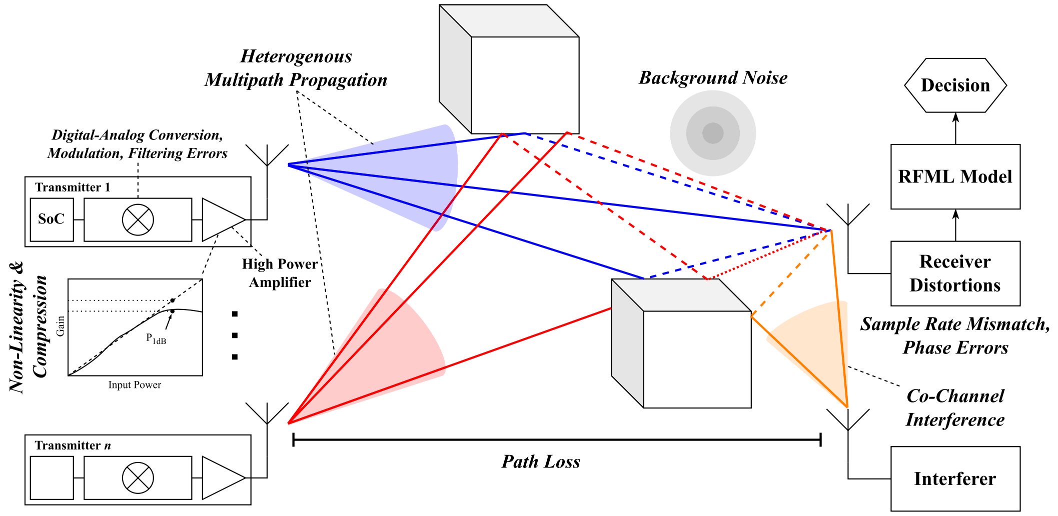

FIGURE 3. System-level diagram of the transmitters and receiver in the propagation environment with labeled non-idealities used as parameter variations.

TABLE 2. Components of the transmitter host and each collection node in the experimental setup.

Component	Selection
Collection Node	
Processor	Intel i5-11600K (6 cores)
Motherboard	Asus ROG STRIX Z590-I
Memory	64 GB DDR4-2400
Storage	2 TB M.2 SSD, 14 TB SATA HDD
Networking	Intel X520-DA2 10/1 GbE SFP+
Power Supply	750 W
Operating System	Ubuntu 22.04
Receiver (Type 1)	USRP X310; SBX-120 (400-4400MHz)
Receiver (Type 2)	USRP B210
Transmitter Host	
Processor	Intel i7-12700 (8p+4e cores)
Motherboard	Asus Prime Z690-P
Memory	64 GB DDR4-2400
Storage	2 TB M.2 SSD, 16 TB SATA HDD
Networking	ST1000SPEX2 1GbE RJ45
USB Hub	12x 10 Port USB 2.0
USB Controller	3x PEXUSB3S44V 4-Port USB 3.2 Gen 1
Relay	Numato Lab 16 Channel USB Relay
Power Supply	1000 W
Aux. Power Supply	300 W
Operating System	Ubuntu 20.04
Transmitter	60x YARD Stick One

adjustments to data collection for varying specific parameters of interest [63].

1) RECEIVE HARDWARE

The chosen receivers are 2 USRP X310s with SBX-120 daughterboards and 1 USRP B210. Each receiver radio is connected to a dedicated CN machine, which provides control and data storage functionality. To minimize any possible differences resulting from antenna positioning, all receivers share a common antenna though an RF splitter/combiner network. The upper half Table 2 gives a list of significant components of the collection nodes. Fig. 4 shows a photograph of this portion of the collection setup.

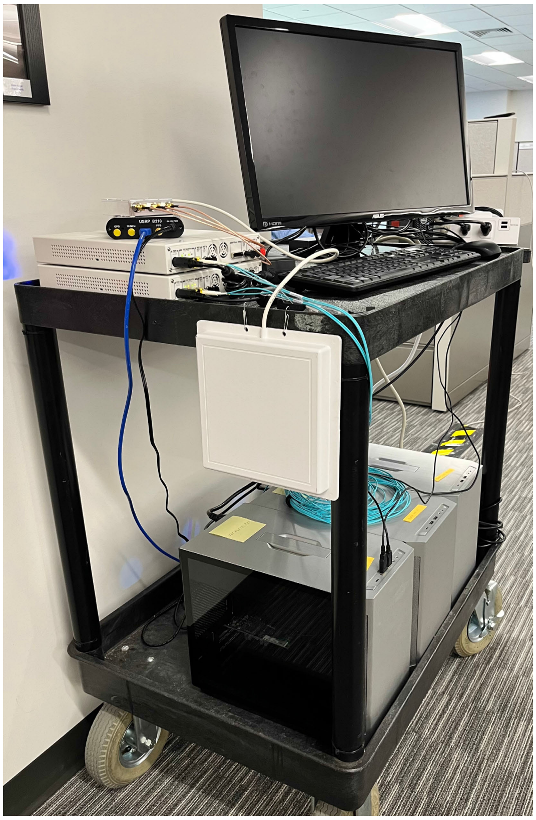

FIGURE 4. The receive-side hardware setup with collection nodes and their attached radios. The collection nodes are located on the lower shelf while the radios are placed on the top of the cart.

2) TRANSMIT HARDWARE

The BURP machine is unique in the literature for the number of real-world transmitters, with support for up to 120 emitters [63]. For this work, the chosen transmitters are 60 Great Scott Gadgets YARD Stick Ones (YS1s) [64]. The transmitters communicate with the transmitter host over

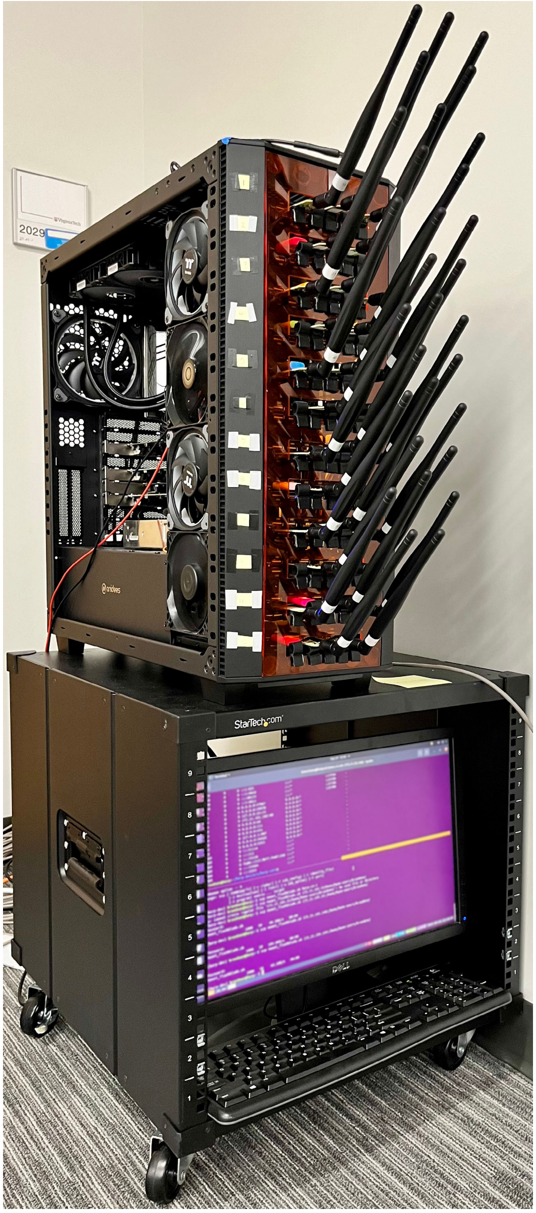

FIGURE 5. **The transmit-side hardware setup of the transmitter host and attached YARD Stick One radios on the front USB hub array.**

the USB protocol and are installed in an array of USB hubs on the side of the machine oriented towards the receivers. These transmitters are low cost and are based on the Texas Instruments CC1111Fx MCU [64]. Being an off-the-shelf device, the YS1 is both highly available and similar in construction to IoT devices that operate in the sub-GHz ISM bands. With the quantity of low-cost emitters in the same weight-class as the typical IoT transmitter, this platform is well suited to collecting data for RFML algorithms intended for IoT applications – security, cognitive radio, etc. the lower half of Table 2 gives a list of significant components of the transmitter host. Fig. 5 shows a photograph of the transmitter host used for this portion of the collection setup.

D. PARAMETERS

The choice of real-world over-the-air collections with real hardware enables the study of sensitivity to the differences in the propagation environment and emitter impairments. Fig. 3 gives an overview of these real-world effects on the collection setup. The experimental setup of this work seeks to isolate and study the impact of power amplifier gain compression, co-channel interference, path loss, and micro-channel variations of multipath fading differences due to variable transmitter location.

1) POWER AMPLIFIER GAIN COMPRESSION

Impairments in the RF signal chain, particularly HPA nonlinearities, contribute to distortions in the emitted signal that lead to a transmitter's specific RF fingerprint, whose distinctness is important for the performance of SEI algorithms [19]. These distortions become very significant when the HPA enters compression at higher power levels, i.e., when the gain of the HPA begins to near the 1 dB compression point, P_{1dB}, decreasing gain 1 dB from its previously constant value. This effect is studied by proxy by varying the target output power of emissions. For the chosen emitter, limits imposed by the manufacturer prevented over-driving of the output to fully enter the compression region, but non-linearities still contribute to varying output characteristics when modulating the target power level.

2) CO-CHANNEL INTERFERENCE

Co-channel interference can significantly degrade the quality of data for received signals by introducing energy at unexpected frequencies. This signal, that is not part of the signal-of-interest, may be incorrectly interpreted as useful features by AMC or SEI algorithms, leading to undefined behavior and decreased classification accuracy. Adding a notch filter to remove the interference will remove the unexpected signal, but any out-of-band features that may have been useful for classification will be unavailable to the RFML models, also possibly decreasing classification accuracy. This work studies real interference introduced by another emitter.

The choice to differentiate between synthetic interference and real-world interference is due to the expectation that real-world interference, by existing in the same environment as the signal-of-interest, will also be influenced by similar multipath and phase noise effects. Synthetic interference is much easier to generate at scale, since it can be injected as a post-processing step, but it is unknown whether using it during training significantly improves performance of an RFML algorithm when faced with real interference.

3) MICRO-CHANNEL VARIATION

Differences in transmitter placement within the environment induce a heterogeneous multipath fading effect for each transmitter. Transmissions from radios located close together may have similar, but not identical, multipath fading – referred to in this work as channel micro-variations.

RFML works dealing with captured data generally agree that variations multipath propagation may have some role to play in classification performance, but it is unknown how intense these variations can be before performance is affected. Depending on placement in the hardware setup, with emissions at 915.25MHz, the position of each transmitter can vary from fractions of a wavelength ($<1\lambda$) to multiple wavelengths ($\sim5.4\lambda$) between the top and bottom slots. This effect is explored by scrambling the positions of most emitters, maintaining a subset stationary, and comparing the model's classification performance across these groups.

4) SIGNAL-TO-NOISE RATIO

Background noise is present in all channel environments, typically modeled as additive white Gaussian noise (AWGN). The signal-to-noise ratio (SNR) of a signal determines the clarity of raw data available to a model, influencing the degree to which subtle features are discernible to RFML algorithms. SNR is the most common system parameter over which the existing literature evaluates performance (i.e. classification accuracy).

E. COLLECTION SCENARIOS

We examine RFML model sensitivity through six scenario sets, each designed to isolate specific environmental and hardware effects that may affect classification performance. Each of the scenarios is a variation upon the baseline scenario that is intended to isolate a particular effect commonly cited in RFML works. These effects include: co-channel interference, both synthetically generated and produced by real emitters within the environment, variable transmission power – intended to elicit transmit signal chain nonlinearities, micro-variations in transmitter location, and line-of-sight path loss.

1) BASELINE SCENARIO

The baseline scenario is intended to represent a benign lab collection environment. The transmitter host and the receivers are indoors and placed 30m apart, with the receivers sharing a common antenna through an RF combining network. Each device transmits a full-power (10 dBm), 1024 byte burst of randomized data at 31,250 bits per second. The transmissions are at a chosen center frequency of 416.4 MHz and use one of the OOK, MSK, 2-FSK, 4-FSK, or 2-GFSK modulation schemes. The receiver sample rate and bandwidth are both 250 kHz.

The testbed setup is discussed in detail in Section II-C. A diagram for the arrangement of transmitters and receivers within the collection environment for the baseline scenario is shown in Fig. 6, and pictures of the receive-side and transmit-side setup are shown in Fig. 4 and Fig. 5, respectively.

As described in Subsection II-C, each transmitter is positioned on a bank of USB hubs on a transmitter host. To increase generality and reduce the likelihood of the models learning patterns associated with specific transmitter locations, the transmitter positions within the bank are scrambled at regular intervals.

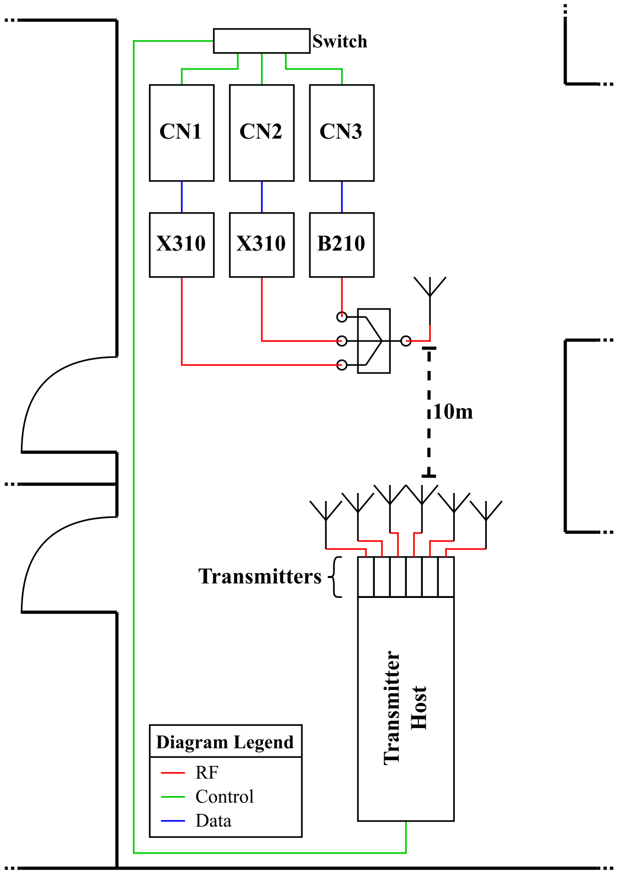

FIGURE 6. The arrangement of the collection setup for the baseline scenario (not to scale). Each of the receivers shares a common antenna through a 3-way RF splitter and stores data on their respective CNs. The transmitter host and CNs are connected through a control back-plane to coordinate accurate data labeling.

2) SIGNAL-TO-NOISE RATIO

This scenario assesses the model's performance across different SNR levels. Understanding the impact of SNR is important because a model trained on high SNR data may learn subtle features that are not discernible at low SNRs. Learning these features may cause the model to have high performance on its own dataset, but worse performance in noisier conditions. Documenting the inflection point where this begins to occur is crucial for optimizing future dataset generation efforts.

No changes occur in the data collection process from the baseline scenario, except for the inclusion of varying-intensity additive white Gaussian noise (AWGN) introduced as a post-processing step on each example. The channel effects and noise in the original baseline collections is still present along with the augmented noise. Typically, signal SNR refers to the ratio in power of a symbol to the noise floor. In this scenario, SNR instead refers to the decision SNR – the ratio of the power of every symbol in an example to the

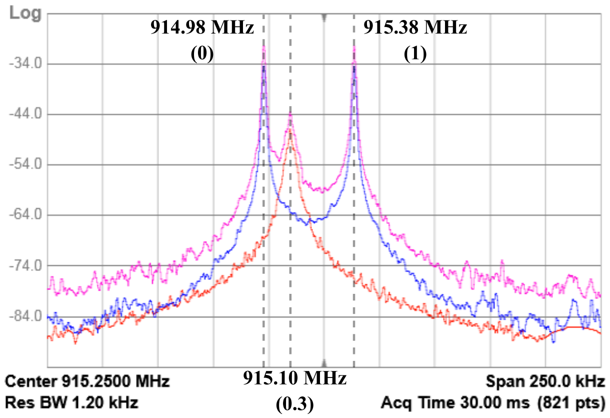

FIGURE 7. A composite image from a spectrum analyzer of a sample 2-FSK transmission with interference of 1 kHz bandwidth positioned 30% of the way between the symbol tones. The signal is visible in blue, the interference in red, and the resulting signal+interference in magenta.

background noise. In this case, the model makes a decision based on an observation window 8 symbols in length.

3) REAL-WORLD INTERFERENCE

The real-world interference scenario assess the model's performance with adjacent emitters introducing co-channel narrowband interference. This is intended to mimic a deployment environment with many devices communicating on similar frequencies and to reflect the fact that real-world interference is subject to the same channel conditions and receiver characteristics as the signal-of-interest.

An emitter is placed between the transmitter and receiver as in the baseline scenario. Data collections are conducted with the emitter producing Gaussian interference with bandwidths of 1 kHz and 10 kHz (equivalent to 1.4% and 13.8% of the modulated bandwidth) and at random center frequencies within the bandwidth of the signal of interest. The interfering emitter's gain was tuned to create interference with signal-to-interference ratio (SIR) levels targeted around 30, 20, and 10 dB. In reality depending on the actual power of signal-of-interest.

4) TRANSMISSION POWER

Especially for SEI performance, a considerable amount of the noticeable variation between transmitters is attributed to HPA non-linearity distortions [9], [65], [66], [67]. Using real integrated circuit hardware limits direct control over HPA distortion intensity. Instead, varying transmission power is used as a proxy for attenuating this effect – decreasing transmission power will move the transmitter's HPA further into its linear region, potentially reducing the distortion that typically contributes to an RF fingerprint.

Transmission power is varied from the baseline scenario and data is collected for emissions at 5, 7, 8, 9, 10 dBm. The chosen power levels are concentrated near the upper end of the transmitters capabilities to attempt to excise the nonlinear region where HPA compression occurs. It is expected that the greatest observed differences will lie between increments near the devices' upper limit, though speculation also exists that the manufacturer's software limits prevent driving the device into full-on saturation.

5) STATIC SUBSET

Variations in multipath propagation commonly hinder SEI algorithms' ability to generalize across scenarios [4], [24], [31]. This is because RFML algorithms trained on real world data have been shown to have a tendency to associate the patterns of variations in the propagation channel with particular labels instead of real differences in an emitter's RF fingerprint or modulation scheme [15], [23], [24], [25]. The purpose of this scenario is to test if small variations of radio positioning and multipath affect classification accuracy.

In this scenario, 12 transmitters are held at constant positions on the bank of USB hubs while the rest are scrambled at regular intervals. RFML models trained on this data are evaluated on classification performance comparing radios with constant position versus those that are scrambled.

6) COMMON ANTENNA

This scenario assesses RFML model performance by eliminating propagation environment differences to isolate the effect of micro-channel variations in multipath propagation.

In this scenario, all transmitters from the baseline scenario use a common antenna, connected through an RF combining network. We use passive Wilkinson-based combiners since they are the cleanest for large-scale signal combining. Transmitter positions within the bank are no longer scrambled.

III. METHODOLOGY

A model is trained for each scenario based on datasets reflecting each specific parametric variation. Performance is then evaluated against the respective testing sets. The model is a CLDNN trained on real-world captured RF data, whose structure is described in Section II-A. The performance for each of these models is communicated through a *confusion matrix* rasterized as an image – a visualization of how frequently examples of a particular class are associated with each label and how often the classification is correct. In a confusion matrix, the horizontal position corresponds to the predicted class, and the vertical position corresponds to the true class, with the entries along the diagonal corresponding to when the predicted class matches the true class. Overall classification accuracy, the average *precision*, average *recall*, and the average *F1 score* for each class (macro F1) are also provided. The F1 score is a commonly-used metric to assess classification performance; it is defined as the harmonic mean of *precision* and *recall*. Precision measures the accuracy of predictions; it is defined as the ratio of true positive predictions to all predicted positives. Recall measures the frequency at examples of a particular class are able to be correctly identified as that class; it is defined as the ratio of true positive predictions to all occurrences of that class in the dataset (true positive predictions and false negative predictions).

For experiments such as this static subset, where only the relative classification accuracy of a subset is compared to the rest of the group, only the confusion matrix is necessary. For the experiments with multiple variations, such as different SNR levels, overall performance is compared between trials. Developing different datasets for each parametric variation is important since it allows the creation of a model from a specific set of parameters that can then be evaluated in performance on examples in different channel environments, thus isolating performance difference resulting from a change in parameters.

For SEI, each dataset consists of 720,000 (12,000 per class) training, 60,000 (1,200 per class) validation, and 60,000 (1,200 per class) testing examples. For AMC, each dataset consists of 60,000 (12,000 per class) training, 6,000 (1,200 per class) validation, and 6,000 (1,200 per class) testing examples. This quantity is believed to be sufficient from previous work on this architecture, which observed that marginal performance gains decreased as data quantity increased, with similar AMC classification accuracy between 5,000 and 10,000 training examples per class [3]. From raw captures, examples are generated by isolating transmissions of 1024 symbols (31,250 samples) by comparing timestamps of detected energy at the receiver and those recorded in the groundtruth during data collection and then partitioning into individual examples. Each example in the dataset consists of 256 raw IQ samples in a window that contains about 8 symbols, meaning each isolated transmission yields around 122 usable examples. The model for each scenario has the same architecture as the baseline model, which has been trained on its corresponding dataset from collected data in the baseline scenario, discussed in Section II-E1.

IV. RESULTS AND ANALYSIS

Classification models for each parametric variation were created on the data collected in each of the scenarios outlined in Section II-E according to the training regimen outlined in Section III. Classification performance of the models in these scenarios are compared to the established baseline performance. Some factors were found to cause performance degradation, such as decreased SNR or the presence of interference. Other scenarios, such as variable transmission power, showed patterns of classification accuracy between circumstances that suggested different levels of model interoperability depending on the specific pairings of training and evaluation datasets.

A. BASELINE

The overall classification performance in the baseline scenario is given in Table 3, showing an 89% baseline accuracy for AMC tasks and 34% baseline accuracy for the 60-class SEI task. The confusion matrix for AMC is shown in Fig. 8, which is acceptable on its own for a single input.

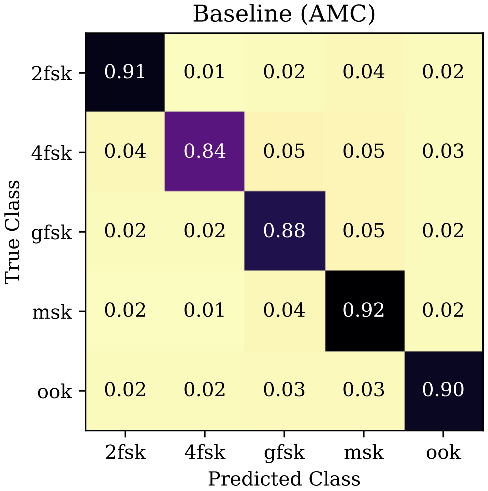

FIGURE 8. Baseline confusion matrix for AMC.

TABLE 3. Classification performance of the model in the baseline scenario.

Task	Accuracy	Avg. Precision	Avg. Recall	Avg. F1 Score
AMC	89%	0.89	0.89	0.89
SEI	34%	0.35	0.34	0.33

B. DECISION AGGREGATION

Before proceeding further to the quantitative performance analyses, it is important to briefly introduce a previously published method [68] to aggregate CNN-based classifier decisions into an improved overall decision. To use this multinomial-based method, the core assumptions are that the inputs are drawn from independent, identically distributed (*iid*) sets and that for the brief window of observation that the signal classifications should be the same. In our scenario, temporal edge detections are used to isolate the burst, which we know to not have any significant co-channel effects (except where intended), so we can confidently aggregate the classification decisions of successive input frames during that isolated burst into a single decision. Our CNN input frame size is on the order of 8 symbols, so aggregating successive decisions by a factor of 10x or 100x is a reasonable range over which to combine decisions. The SEI confusion matrices for the base case (without applying multinomial decision synthesis), and the cases for $n = 10$, and $n = 100$ are shown in Fig. 10.

Under these *iid* assumptions, the multinomial decision aggregation approach will lead to a better overall classification as long as the correct decision (i.e., the diagonal elements of the confusions matrix) is in fact the dominant one. Moreover, the rate of the increase in performance, which naturally converges to a fixed value, is relative to the ratio of the correct decision and the most likely error(s). When any of the error cases exceeds the correct decision, then aggregation will actually decay performance, making the resulting decision [incorrectly] more confident in the error. As such, the net effect of aggregation is a benefit as long as the correct decision exceeds random guessing and none of

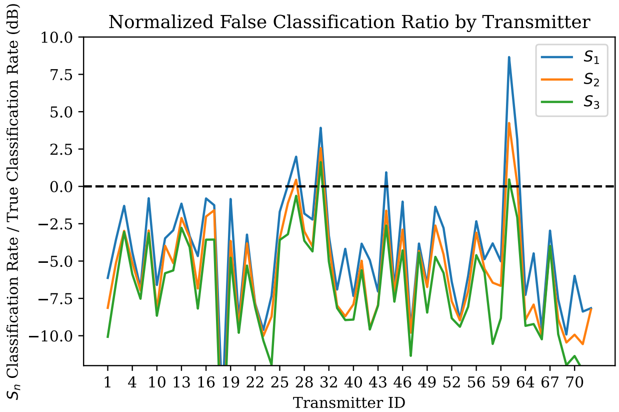

FIGURE 9. The next-highest classification rate normalized by the true classification rate for a given transmitter ID expressed in dB. S_1 represents the highest classification rate of an incorrect class for a true class of a particular transmitter ID. S_2 represents the next highest false-classification rate and S_3 represents the next highest after that. Values below 0 indicate that classification accuracy would improve with aggregation, while values above 0 indicate that classification accuracy would degrade for that specific emitter.

the error cases are more prevalent. For the case of a 60-class SEI algorithm, where random guessing is only 1.4%, that means that a 10 − 20% baseline accuracy from a single input can actually be quite positive and grow into a highly confident decision over multiple successive input frames. To demonstrate this more concretely, a visual comparing the confusion matrices of the 60-class SEI classifier output based upon a single input (left), 10 inputs (center) and 100 inputs (right) is shown in Fig. 10. As more decisions are aggregated, the aggregated decisions have substantially higher accuracy, with the notable exceptions where the baseline trained CNN had errors that exceed the correct decisions for a given emitter. A deeper inspection of that condition is shown in Fig. 9, where the top-3 conditional errors are depicted, normalized as a ratio with the true decision probability. As such, a value of 0 dB for S1 corresponds to a equal probability of the true decision and the worst case of potential SEI classification error; S2 and S3 correspond to the next two ordered error cases, respectively. The emphasis in these further analyses of the classification decision is to explore how well aggregating multiple successive decisions into a higher confidence decision will perform, with guaranteed improvements whenever {S1, S2, S3} are smaller ($<$ 0dB) than the true decision probability [68]. These results show that decision aggregation approach leads to emitters {26, 27, 31, 44, 61, 62} each decaying in performance, while the other 66 emitter classes improve. The remainder of the paper makes use of this decision aggregation method, with a consolidation of 10 successive outputs for AMC algorithms and 100 successive outputs for SEI decisions.

C. INTERFERENCE

Interference tests focused on center frequencies within the signal-of-interest's bandwidth. Results are separated into cases where interference of two bandwidths was introduced with three levels of intensity. Evaluation results reflect the performance of models trained on data with this interference present. Interference presents a particularly challenging problem for the current approach to both AMC and SEI with the observed level of performance degradation. Successfully mitigating these effects could substantially recover performance.

TABLE 4. Classification performance with different intensities of 10kHz interference.

Task	Intensity	Accuracy	Avg. Precision	Avg. Recall	Avg. F1 Score
AMC	*Minor*	71%	0.71	0.71	0.70
	Medium	51%	0.51	0.50	0.50
	Severe	25%	0.29	0.26	0.26
SEI	*Minor*	25%	0.25	0.25	0.24
	Medium	10%	0.10	0.10	0.09
	Severe	7%	0.09	0.07	0.06

TABLE 5. Classification performance with different intensities of 1kHz interference.

Task	Intensity	Accuracy	Avg. Precision	Avg. Recall	Avg. F1 Score
AMC	*Minor*	48%	0.48	0.48	0.48
	Medium	38%	0.41	0.38	0.37
	Severe	22%	0.23	0.22	0.19
SEI	*Minor*	8%	0.08	0.06	0.05
	Medium	13%	0.23	0.08	0.05
	Severe	3%	0.05	0.03	0.02

1) 10KHZ BANDWIDTH

AMC model performance declined with increasing interference intensity at 10kHz bandwidth, as given in Table 4. In Fig. 11, it is shown that at the highest level of interference, the model became much more likely to classify examples as 4-FSK, appearing to incorrectly interpret the interferer as an extra tone in the FSK symbol set. This likely occurred because 4-FSK's wider occupied bandwidth, similar to that of the 10kHz interference, led the model to mis-classify narrower-bandwidth modulation schemes as 4-FSK.

As can be observed from Table 4, SEI performance also decreased as interference intensity increased. Fig. 12 shows the pattern in more detail, where SEI performance is somewhat preserved in the low interference case and still somewhat present in the medium case, but completely useless when high interference is present. Vertical bands appear in the confusion matrices, where the model appears to pick from a small subset of classes that it prefers to pick over others as interference intensifies.

2) 1KHZ BANDWIDTH

With 1kHz bandwidth interference, performance of the AMC model was impaired even at the lowest level of interference, further decreasing as intensity increased, as shown in Table 5. In Fig. 13, it can be seen that as the level of interference increases, the model begins to confuse examples

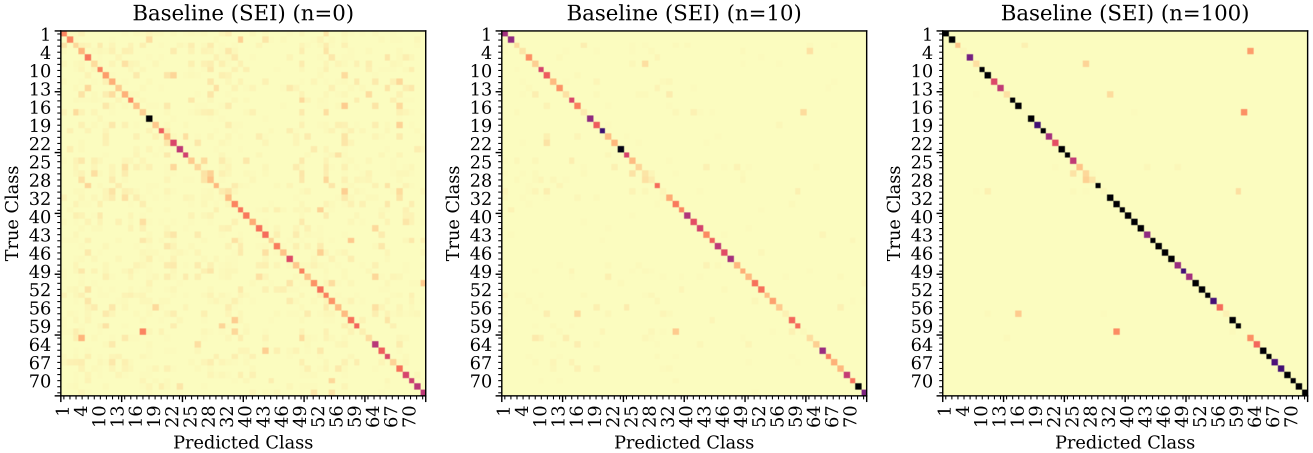

FIGURE 10. Confusion matrices of the SEI performance in the baseline case. *n* represents the number of iteration of the multinomial sampling technique applied, with the leftmost matrix (*n* = 0) representing the raw performance before the statistical algorithm is applied and the rightmost matrix (*n* = 100) showing the result of 100 iterations.

for 2-FSK and GFSK. This is possibly because of the narrow bandwidth interference having a more distinct energy "spike" at a particular frequency, a characteristic consistent with transmissions of those modulation schemes. It is not known why the model prefers these two over MSK, but it should be noted that predictions of MSK are still much more frequent than the remaining modulation schemes.

Under the same circumstances, Table 5 shows that the performance of the SEI model became extremely deteriorated. From Fig. 14, it can be seen that at lower and medium levels of interference, the model was still able to make inferences with very minor success. The classification accuracy in the medium case is slightly higher than in the low case, however the F1 score is identical, suggesting that the model is not actually any more reliable. It can also be seen that the output is also not completely random, with the model seemingly preferring to predict from a small subset of emitters whenever it had low confidence – a pattern especially discernible in the severe interference case.

3) PERFORMANCE COMPARISON

Fig. 15 shows the performance of the models trained on each variation of interference evaluated on each other variant. Each model seems to have the best performance only on the exact type of interference they were trained on, without much evidence of natural generalization to even different intensities of same-bandwidth interference.

The level of observed performance degradation was found to be a function of both SIR and the interference bandwidth. For both AMC and SEI, the classification performance of the model was found to have a positive relationship with SIR, with performance decreasing in tandem with decreasing SIR level. Controlling for the same SIR level, the narrower bandwidth interference was found to have a greater impact than the wider bandwidth interference. For SEI, this is likely because the interference at each SIR level was calibrated to a specific total integrated power instead of peak power spectral density, meaning the narrower bandwidth interference had the same amount of total power as the wider bandwidth interference, but with a higher power spectral density concentrated around its center frequency. The model had an easier time coping with the more distributed nature of the wider bandwidth interference, where it was able to more easily ignore the excess information and discern the critical components of the signal-of-interest. On the other hand, the more concentrated peak power of the narrower bandwidth interference completely dominated the local region around its center, making unrecoverable components of the signal that were critical for classification.

For AMC, the specific patterns of mis-classification also depended on the bandwidth, with wider bandwidth interference contributing to a dominant decision of the wider bandwidth modulation scheme 4FSK and narrower bandwidth interference contributing to a dominant decision of the narrower bandwidth modulation schemes FSK and GFSK.

D. TRANSMISSION POWER

Overall classification performance for both AMC and SEI at every level of transmission power is given in Table 6. From these observations, there is no clear linear relationship between transmission power and performance. The power levels 8dBm and 0 dBm have noticeably lower classification performance for both AMC and SEI, suggesting that the same factor that is degrading performance impacts both tasks in a similar manner. It can be seen in the confusion matrices that there is no particular pattern of favoring any particular modulation scheme or transmitter ID, unlike what was observed in the case of interference. It is unlikely that this factor lies in the data collection setup, since examples for every power level were collected in serial with only the transmission power adjusted in software.

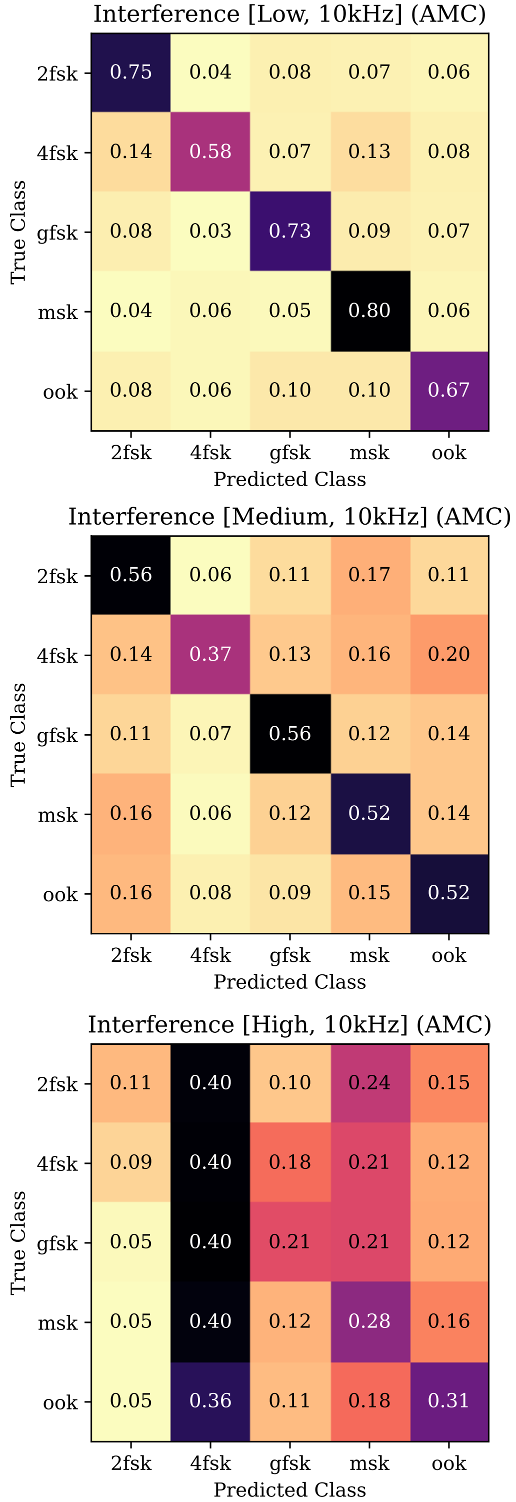

FIGURE 11. Confusion matrices for AMC with 10kHz bandwidth interference.

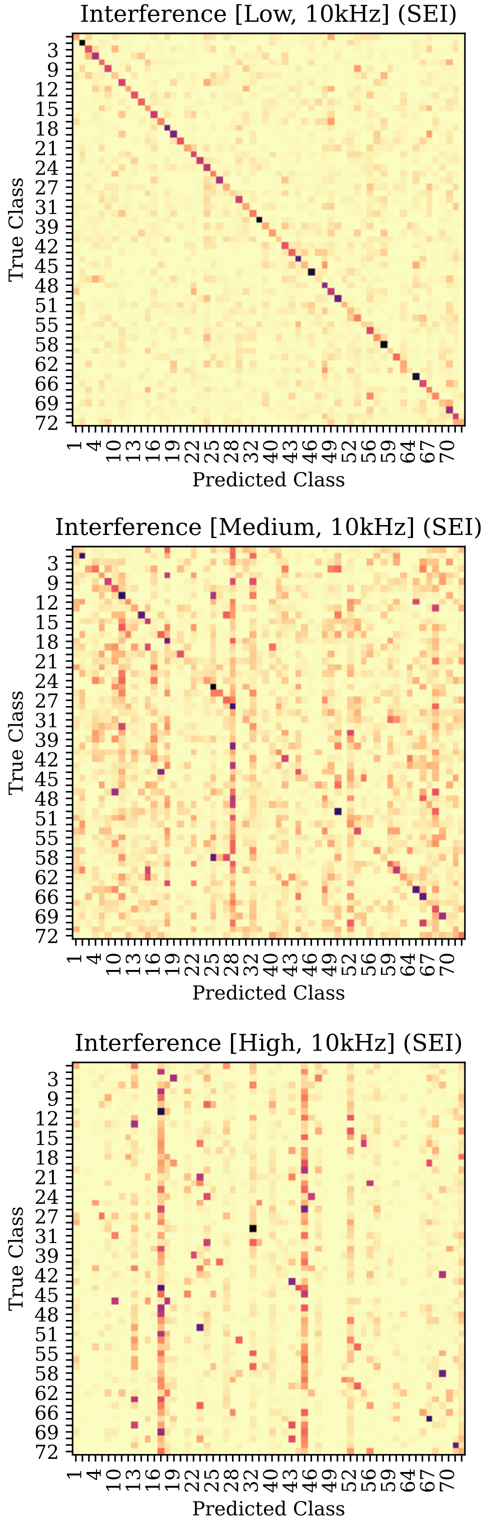

FIGURE 12. Confusion matrices for SEI with 10kHz bandwidth interference.

It is similarly unlikely that this is due to an error in the detection and isolation phase of dataset creation, given that scenarios of both higher and lower power levels still had high performance. It is possible that this pattern arises from behavior that is consistent across all transmitters at these specific power levels, such as a consistent pattern of attenuation or garbled output that makes the emitted signal less recognizable as a well-formed transmission.

Fig. 16 shows a comparison of the performance of every model trained on variable transmission power on the testing

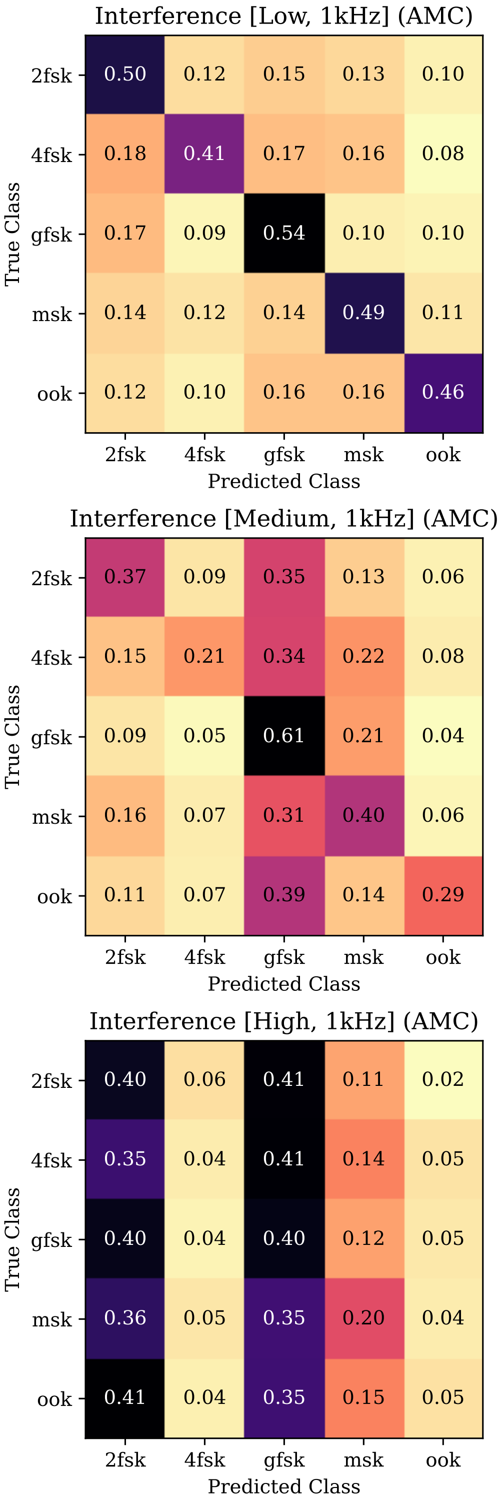

FIGURE 13. Confusion matrices for AMC with 1kHz bandwidth interference.

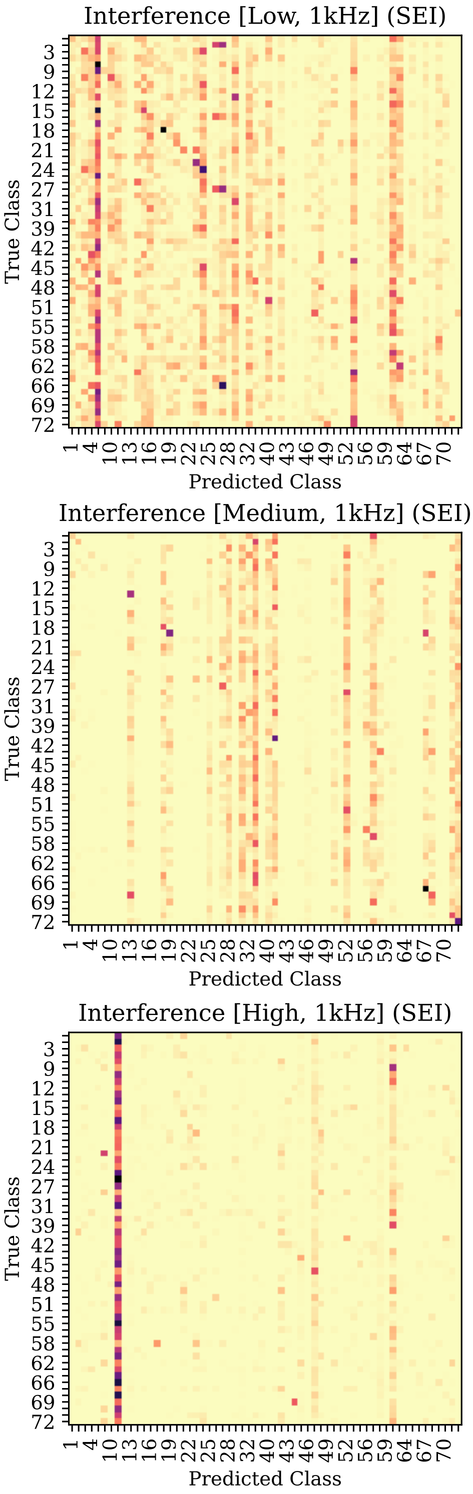

FIGURE 14. Confusion matrices for SEI with 1kHz bandwidth interference.

sets for each transmission power. For AMC, there is a clear pattern of each power level having a particular evaluation accuracy, regardless of the power level the model was trained on. This suggests that the models learned very similar features for each power level and that the factors impacting classification accuracy remained consistent between devices within each power level. SEI performance shows that models were most successful on the datasets they were trained on and very unsuccessful on datasets they were not trained on, with minor spillover into power levels one step higher or lower. The differences in the patterns between AMC and

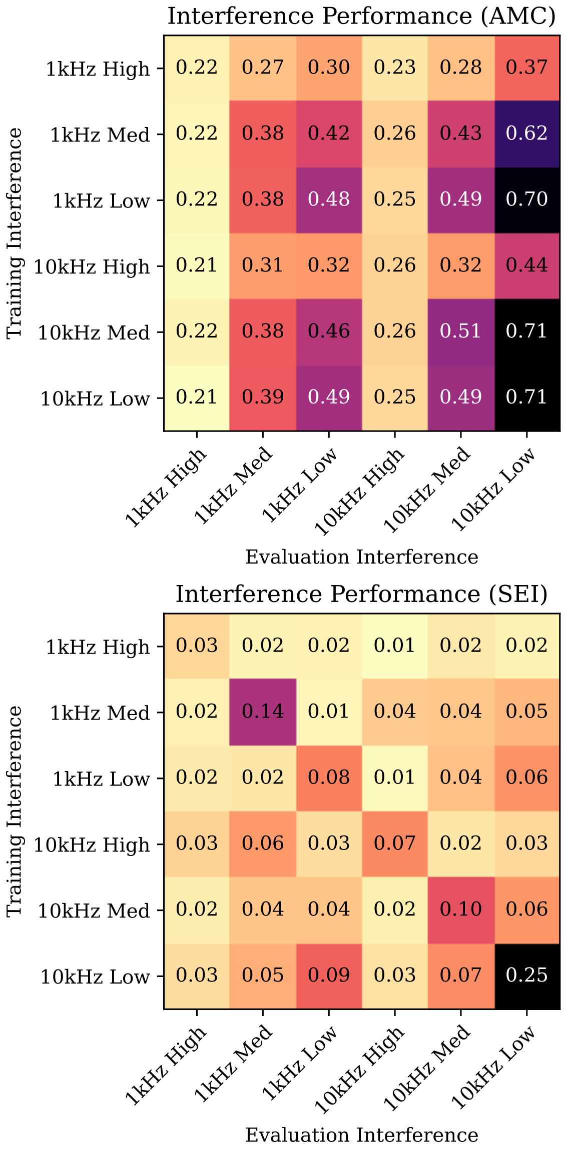

FIGURE 15. Performance matrix comparing AMC and SEI classification accuracy of models trained on one type of interference evaluated on every other type of interference.

TABLE 6. Classification performance with different levels of transmission power.

Task	TX Power (dBm)	Accuracy	Avg. Precision	Avg. Recall	Avg. F1 Score
AMC	9	89%	0.89	0.88	0.88
	8	74%	0.73	0.72	0.72
	7	98%	0.98	0.98	0.98
	5	94%	0.94	0.94	0.94
	0	77%	0.77	0.76	0.76
	-10	93%	0.92	0.92	0.92
SEI	9	39%	0.39	0.38	0.38
	8	25%	0.26	0.25	0.24
	7	46%	0.47	0.47	0.47
	5	46%	0.47	0.46	0.46
	0	27%	0.28	0.28	0.27
	-10	34%	0.34	0.34	0.33

SEI classification accuracy suggest that whatever factors are contributing to differences in AMC performance do not contribute to the RF fingerprints of the transmitters.

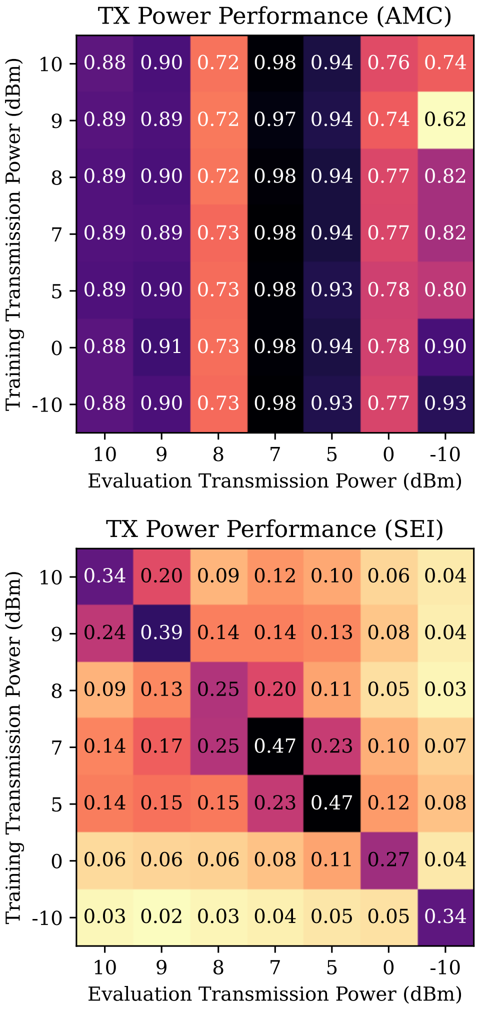

FIGURE 16. Performance matrix comparing classification accuracy for AMC and SEI of models trained on one transmission power level evaluated on every other transmission power level.

TABLE 7. Classification performance comparison between models trained on static radios, shuffled radios, and on the overall dataset.

Task	Set	Accuracy	Avg. Precision	Avg. Recall	Avg. F1 Score
AMC	Overall	86%	0.85	0.85	0.85
	Static	75%	0.76	0.74	0.74
	Shuffled	88%	0.89	0.88	0.88
SEI	Overall	46%	0.45	0.45	0.44
	Static	60%	0.61	0.60	0.60
	Shuffled	43%	0.44	0.42	0.43

E. STATIC SUBSET

The performance of the collection scenario where a subset of radios is held in constant position is given in Table 7. The confusion matrix is shown in Fig. 17. The overall AMC performance was similar to the baseline scenario, with the static subset having slightly lower performance. The overall SEI performance was measurably higher, with the shuffled subset having similar performance to the baseline and the static subset having considerably higher classification

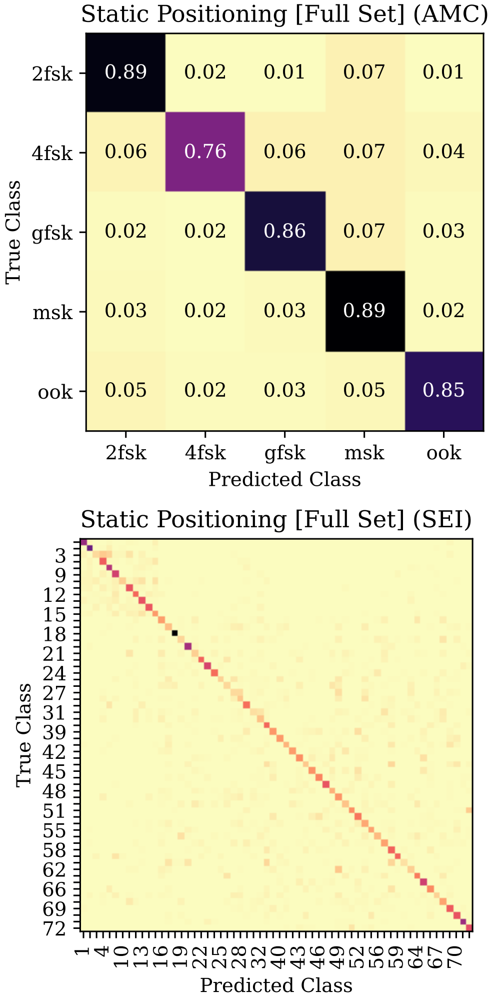

FIGURE 17. Confusion matrices for AMC and SEI where emitters with IDs 1-15 and below were held in a constant position.

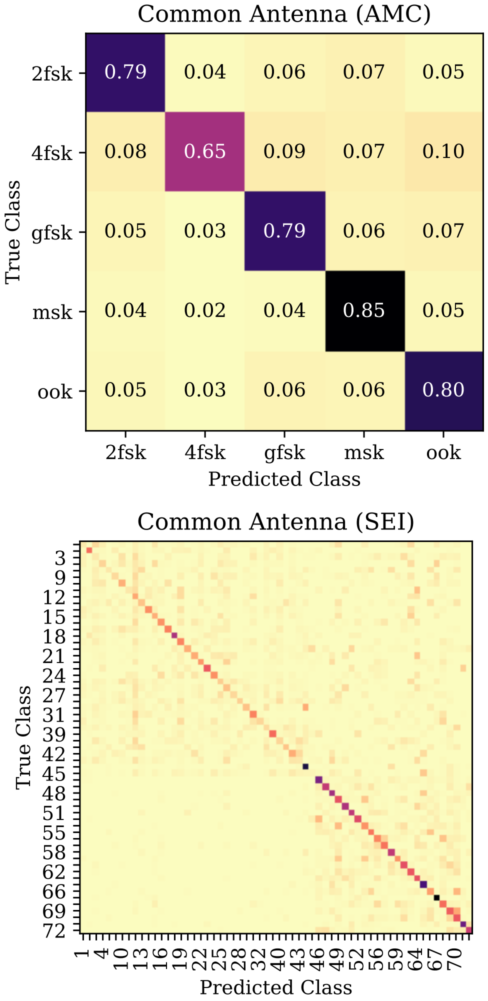

FIGURE 18. Confusion matrices for AMC and SEI where all emitters shared a common antenna.

TABLE 8. Classification performance of the model with all transmitters using a common antenna.

Task	Accuracy	Avg. Precision	Avg. Recall	Avg. F1 Score
AMC	79%	0.78	0.78	0.78
SEI	34%	0.31	0.31	0.29

performance. When looking at the confusion matrix, there are two clearly-defined regions split between the static and shuffled set, where emitters are not often confused between the two regions. This suggests that consistency in radio positioning plays some part in classification performance.

F. COMMON ANTENNA

Performance of the model in the common antenna case is given in Table 8 and the confusion matrices are shown in Fig. 18. The performance for AMC was somewhat inferior to that in the baseline scenario, possibly caused by the attenuation introduced by the passive RF combining network. For SEI, performance was about the same. Some interesting artifacts of the collection process are visible in the confusion matrix for the transmitters with an ID greater than 45. Data collections took place over the course of two days with devices from one day almost never being confused for devices whose collections took place on the other day. This pattern would suggest that while no changes occurred in the actual experimental setup, dynamic channel conditions were different enough to make the two sets appreciably different.

G. SIGNAL-TO-NOISE RATIO

A performance matrix comparing every pairing of training SNR and evaluation SNR for both AMC and SEI for indoor collections is shown in Fig. 19. For AMC, a pattern emerges where models seem to be able to generalize quite easily to higher SNRs than what they were trained on, but have some more difficulty generalizing to lower SNRs. This matches previously obtained results in related research. For SEI, models tend to have the highest performance on the SNR they were trained on and slightly lower performance with higher SNR levels, but seem to be limited to the accuracy they are

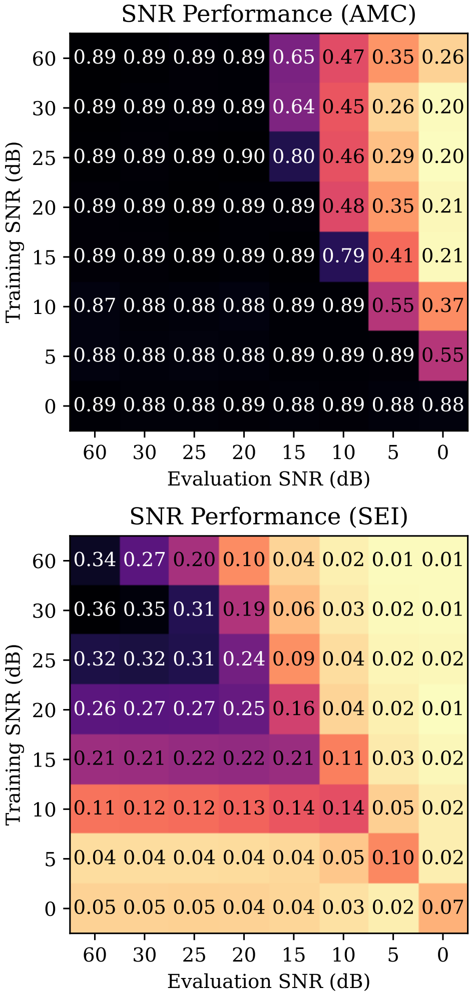

FIGURE 19. Performance matrix comparing AMC and SEI classification accuracy of models trained on one SNR evaluated on every other SNR level of data collected in the baseline scenario.

able to achieve with their native dataset. Models seem to have considerable difficulty with data at a higher SNR than what they were trained on. This would suggest that increasing noise levels obscure features necessary for a reliable RF fingerprint, with higher noise levels obscuring more features. Models trained at lower SNR levels do not seem to learn certain features that are discernible at higher SNR levels, but are still able to discern the features they have learned even when the SNR increases.

H. SUMMARY

From the results of these experiments, the studied factors are ranked in order of their observed impact:

1) Interference
2) SNR
3) Transmitter HPA non-linearity
4) Multipath propagation

When creating datasets from real-world captured data, it is necessary to label these conditions in experimental setup, because a model trained under one set of circumstances may not necessarily be applicable to circumstances with different conditions. This is especially true for high-impact factors such as interference and SNR, which have the potential to drastically diminish model performance. As such, detailed analyses of performance over SIR are warranted.

V. FUTURE WORK

This work investigated only a small fraction of the factors that could affect RFML performance, with many more not yet investigated. Even of the parameters investigated here, factors such as differences in multipath propagation and dynamic channels remain difficult to characterize.

There are too many variations of possible types of interference to address in a introductory work such as this one looking for a common way to compare sensitivity to parametric variations for RFML models. Interference can vary in terms of fractional bandwidth of the transmission, from extremely narrow to wide bandwidths. While a rudimentary relationship between bandwidth and AMC classification patterns was observed in this work, the full impact of interference bandwidth requires further investigation. In-band and out-of-band interference can also be present; while this work only studied interference within the bandwidth of the signal-of-interest, different effects may be observed when the bulk of the interference energy lies outside the transmission band. The type of interference can also vary, with types and prevalence varying depending on the frequency and circumstances. This could take the form of spectral congestion in commonly-used communications bands, patterns of emissions from electronic equipment, or deliberate jamming attempts.

Effects of transmitter construction and operating circumstances also require further investigation. Specific components within the RF signal chain and their operation, especially the HPA, are thought to have partial responsibility for the RF fingerprint that plays a part in SEI performance. However, it is not known how factors such as fatigue or temperature affect the nature of their influence.

VI. CONCLUSION

This work presented a series of experiments to investigate the real-world considerations for RFML applications of interference, multipath propagation, transmitter HPA non-linearity, and SNR. By varying data collection conditions, links between these conditions and the associated impacts of RFML model performance were evaluated. Results from existing works that have documented changes in performance due to SNR and dynamic channel conditions have been reproduced. The observed trends established here prompt the further study of how RFML models can be adapted to function under wider variety of conditions and different hardware configurations.

ACKNOWLEDGMENT

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. government. The U.S. government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

REFERENCES

[1] O. A. Dobre, A. Abdi, Y. Bar-Ness, and W. Su, "A survey of automatic modulation classification techniques: Classical approaches and new trends," *IET Commun.*, vol. 1, no. 2, pp. 137–156, 2007.

[2] L. J. Wong, W. H. Clark IV, B. Flowers, R. M. Buehrer, A. J. Michaels, and W. C. Headley, "The RFML ecosystem: A look at the unique challenges of applying deep learning to radio frequency applications," 2020, *arXiv:2010.00432*.

[3] W. H. Clark and A. J. Michaels, "Quantifying dataset quality in radio frequency machine learning," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, Nov. 2021, pp. 384–389.

[4] L. J. Wong, S. McPherson, and A. J. Michaels., "An analysis of RF transfer learning behavior using synthetic data," 2022, *arXiv:2210.01158*.

[5] Y. Tu, Y. Lin, C. Hou, and S. Mao, "Complex-valued networks for automatic modulation classification," *IEEE Trans. Veh. Technol.*, vol. 69, no. 9, pp. 10085–10089, Sep. 2020.

[6] A. Tsakmalis, S. Chatzinotas, and B. Ottersten, "Automatic modulation classification for adaptive power control in cognitive satellite communications," in *Proc. 7th Adv. Satell. Multimedia Syst. Conf.*, Sep. 2014, pp. 234–240.

[7] S. Ramjee, S. Ju, D. Yang, X. Liu, A. El Gamal, and Y. C. Eldar, "Fast deep learning for automatic modulation classification," 2019, *arXiv:1901.05850*.

[8] R. Lin, W. Ren, X. Sun, Z. Yang, and K. Fu, "A hybrid neural network for fast automatic modulation classification," *IEEE Access*, vol. 8, pp. 130314–130322, 2020.

[9] F. Mkadem and S. Boumaiza, "Physically inspired neural network model for RF power amplifier behavioral modeling and digital predistortion," *IEEE Trans. Microw. Theory Techn.*, vol. 59, no. 4, pp. 913–923, Apr. 2011.

[10] L. Ding, S. Wang, F. Wang, and W. Zhang, "Specific emitter identification via convolutional neural networks," *IEEE Commun. Lett.*, vol. 22, no. 12, pp. 2591–2594, Dec. 2018.

[11] Y. Lin, J. Jia, S. Wang, B. Ge, and S. Mao, "Wireless device identification based on radio frequency fingerprint features," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2020, pp. 1–6.

[12] G. Huang, Y. Yuan, X. Wang, and Z. Huang, "Specific emitter identification based on nonlinear dynamical characteristics," *Can. J. Electr. Comput. Eng.*, vol. 39, no. 1, pp. 34–41, Winter. 2016.

[13] L. J. Wong, W. C. Headley, S. Andrews, R. M. Gerdes, and A. J. Michaels, "Clustering learned CNN features from raw I/Q data for emitter identification," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, Oct. 2018, pp. 26–33.

[14] O. Ureten and N. Serinken, "Wireless security through RF fingerprinting securite sansfil parune empreinte digitale RF," *Can. J. Elect. Comput. Eng.*, vol. 32, no. 1, pp. 27–33, 2007.

[15] A. Jagannath, J. Jagannath, and P. S. P. V. Kumar, "A comprehensive survey on radio frequency (RF) fingerprinting: Traditional approaches, deep learning, and open challenges," *Comput. Netw.*, vol. 219, Dec. 2022, Art. no. 109455.

[16] L. J. Wong, S. McPherson, and A. J. Michaels, "Quantifying raw RF dataset similarity for transfer learning applications," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 2076–2086, 2022.

[17] X. Yin and X. Cheng, *Propagation Channel Characterization, Parameter Estimation, and Modeling for Wireless Communications*. Hoboken, NJ, USA: Wiley, 2016.

[18] B. Hamdaoui, A. Elmaghbub, and S. Mejri, "Deep neural network feature designs for RF data-driven wireless device classification," *IEEE Netw.*, vol. 35, no. 3, pp. 191–197, May 2021.

[19] A. Elmaghbub and B. Hamdaoui, "LoRa device fingerprinting in the wild: Disclosing RF data-driven fingerprint sensitivity to deployment variability," *IEEE Access*, vol. 9, pp. 142893–142909, 2021.

[20] T. J. O'Shea, J. Corgan, and T. C. Clancy, "Convolutional radio modulation recognition networks," in *Engineering Applications of Neural Networks*. Cham, Switzerland: Springer, 2016, pp. 213–226.

[21] S. Riyaz, K. Sankhe, S. Ioannidis, and K. Chowdhury, "Deep learning convolutional neural networks for radio identification," *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 146–152, Sep. 2018.

[22] A. Al-Shawabka, F. Restuccia, S. D'Oro, T. Jian, B. C. Rendon, N. Soltani, J. Dy, S. Ioannidis, K. Chowdhury, and T. Melodia, "Exposing the fingerprint: Dissecting the impact of the wireless channel on radio fingerprinting," in *Proc. IEEE Conf. Comput. Commun.*, Jul. 2020, pp. 646–655.

[23] S. C. Hauser, W. C. Headley, and A. J. Michaels, "Signal detection effects on deep neural networks utilizing raw IQ for modulation classification," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, Oct. 2017, pp. 121–127.

[24] K. Sankhe, M. Belgiovine, F. Zhou, S. Riyaz, S. Ioannidis, and K. Chowdhury, "ORACLE: Optimized radio classification through convolutional neuraL networks," in *Proc. IEEE Conf. Comput. Commun.*, Apr. 2019, pp. 370–378.

[25] T. James O'Shea, T. Roy, and T. Charles Clancy, "Over-the-air deep learning based radio signal classification," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 1, pp. 168–179, Feb. 2018.

[26] W. H. Clark, S. Hauser, W. C. Headley, and A. J. Michaels, "Training data augmentation for deep learning radio frequency systems," *J. Defense Model. Simulation, Appl., Methodol., Technol.*, vol. 18, no. 3, pp. 217–237, Jul. 2021.

[27] T. O'Shea and N. West, "Radio machine learning dataset generation with GNU radio," in *Proc. GNU Radio Conf.*, vol. 1, 2016, pp. 12–16.

[28] S. Apfeld and A. Charlish, "Recognition of unknown radar emitters with machine learning," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 57, no. 6, pp. 4433–4447, Dec. 2021.

[29] C. Comert, M. Kulhandjian, O. M. Gul, A. Touazi, C. Ellement, B. Kantarci, and C. D'Amours, "Analysis of augmentation methods for RF fingerprinting under impaired channels," in *Proc. ACM Workshop Wireless Secur. Mach. Learn.*, May 2022, pp. 3–8.

[30] L. Huang, W. Pan, Y. Zhang, L. Qian, N. Gao, and Y. Wu, "Data augmentation for deep learning-based radio modulation classification," *IEEE Access*, vol. 8, pp. 1498–1506, 2020.

[31] N. Soltani, K. Sankhe, J. Dy, S. Ioannidis, and K. Chowdhury, "More is better: Data augmentation for channel-resilient RF fingerprinting," *IEEE Commun. Mag.*, vol. 58, no. 10, pp. 66–72, Oct. 2020.

[32] P. Wang and M. Vindiola, "Data augmentation for blind signal classification," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, Nov. 2019, pp. 305–310.

[33] K. Wursthorn, M. Hillemann, and M. Ulrich, "Comparison of uncertainty quantification methods for CNN-based regression," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 43, pp. 721–728, May 2022.

[34] S. Ma, R. Bassily, and M. Belkin, "The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning," in *Proc. 35th Int. Conf. Mach. Learn.*, vol. 80, 2018, pp. 3325–3334.

[35] T. Tuukkanen and J. Anteroinen, "Framework to develop military operational understanding of cognitive radio," in *Proc. Int. Conf. Mil. Commun. Inf. Syst. (ICMCIS)*, May 2015, pp. 1–9.

[36] K. Kim, C. M. Spooner, I. Akbar, and J. H. Reed, "Specific emitter identification for cognitive radio with application to IEEE 802.11," in *Proc. IEEE Global Telecommun. Conf.*, Dec. 2008, pp. 1–5.

[37] K. Gulati, A. Chopra, B. L. Evans, and K. R. Tinsley, "Statistical modeling of co-channel interference," in *Proc. IEEE Global Telecommun. Conf.*, Dec. 2009, pp. 1–6.

[38] A. Elmaghbub, B. Hamdaoui, and A. Natarajan, "WideScan: Exploiting out-of-band distortion for device classification using deep learning," in *Proc. IEEE Global Commun. Conf.*, Dec. 2020, pp. 1–6.

[39] N. E. West and T. O'Shea, "Deep architectures for modulation recognition," in *Proc. IEEE Int. Symp. Dyn. Spectr. Access Netw. (DySPAN)*, Mar. 2017, pp. 1–6.

[40] Y. Wang, G. Gui, H. Gacanin, T. Ohtsuki, O. A. Dobre, and H. V. Poor, "An efficient specific emitter identification method based on complex-valued neural networks and network compression," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2305–2317, Aug. 2021.

[41] L. J. Wong, W. C. Headley, and A. J. Michaels, "Specific emitter identification using convolutional neural network-based IQ imbalance estimators," *IEEE Access*, vol. 7, pp. 33544–33555, 2019.

[42] Y. Pan, S. Yang, H. Peng, T. Li, and W. Wang, "Specific emitter identification based on deep residual networks," *IEEE Access*, vol. 7, pp. 54425–54434, 2019.

[43] B. He and F. Wang, "Cooperative specific emitter identification via multiple distorted receivers," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 3791–3806, 2020.

[44] J. Zhang, F. Wang, O. A. Dobre, and Z. Zhong, "Specific emitter identification via Hilbert–Huang transform in single-hop and relaying scenarios," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 6, pp. 1192–1205, Jun. 2016.

[45] Z. Zhang, Y. Li, C. Wang, M. Wang, Y. Tu, and J. Wang, "An ensemble learning method for wireless multimedia device identification," *Secur. Commun. Netw.*, vol. 2018, pp. 1–9, Oct. 2018.

[46] X. Zha, H. Chen, T. Li, Z. Qiu, and Y. Feng, "Specific emitter identification based on complex Fourier neural network," *IEEE Commun. Lett.*, vol. 26, no. 3, pp. 592–596, Mar. 2022.

[47] B. Olds, E. Maas, and A. J. Michaels, "Evaluation of confusion behaviors in SEI models," *IEEE Access*, vol. 12, 2024.

[48] B. Flowers, R. M. Buehrer, and W. C. Headley, "Communications aware adversarial residual networks for over the air evasion attacks," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, Nov. 2019, pp. 133–140.

[49] F. Restuccia, S. D'Oro, A. Al-Shawabka, M. Belgiovine, L. Angioloni, S. Ioannidis, K. Chowdhury, and T. Melodia, "DeepRadioID: Real-time channel-resilient optimization of deep learning-based radio fingerprinting algorithms," in *Proc. 20th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, Jul. 2019, pp. 51–60.

[50] B. Olds and A. J. Michaels, "Temperature sensitivity of RFML algorithms," in *Proc. IEEE ICC Workshop Mach. Learn. Deep Learn. Wireless Secur.*, 2024, pp. 1–19.

[51] M.-W. Liu and J. F. Doherty, "Specific emitter identification using nonlinear device estimation," in *Proc. IEEE Sarnoff Symp.*, Apr. 2008, pp. 1–5.

[52] K. Tekbiyik, Ö. Akbunar, A. R. Ekti, A. Görçin, and G. Karabulut Kurt, "Multi–Dimensional wireless signal identification based on support vector machines," *IEEE Access*, vol. 7, pp. 138890–138903, 2019.

[53] M. Kulin, T. Kazaz, I. Moerman, and E. De Poorter, "End-to-end learning from spectrum data: A deep learning approach for wireless signal identification in spectrum monitoring applications," *IEEE Access*, vol. 6, pp. 18484–18501, 2018.

[54] H. Xu and X. Xu, "A transformer based approach for open set specific emitter identification," in *Proc. 7th Int. Conf. Comput. Commun. (ICCC)*, Dec. 2021, pp. 1420–1425.

[55] B. Chatterjee, D. Das, S. Maity, and S. Sen, "RF-PUF: Enhancing IoT security through authentication of wireless nodes using in-situ machine learning," *IEEE Internet Things J.*, vol. 6, no. 1, pp. 388–398, Feb. 2019.

[56] N. Basha, B. Hamdaoui, K. Sivanesan, and M. Guizani, "Channel-resilient deep-learning-driven device fingerprinting through multiple data streams," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 118–133, 2023.

[57] A. Swami and B. M. Sadler, "Hierarchical digital modulation classification using cumulants," *IEEE Trans. Commun.*, vol. 48, no. 3, pp. 416–429, Mar. 2000.

[58] V. Brik, S. Banerjee, M. Gruteser, and S. Oh, "Wireless device identification with radiometric signatures," in *Proc. 14th ACM Int. Conf. Mobile Comput. Netw.*, Sep. 2008, p. 116.

[59] P. Triantaris, E. Tsimbalo, W. H. Chin, and D. Gündüz, "Automatic modulation classification in the presence of interference," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Jun. 2019, pp. 549–553.

[60] S. Guo, R. E. White, and M. Low, "A comparison study of radar emitter identification based on signal transients," in *Proc. IEEE Radar Conf.*, Apr. 2018, pp. 0286–0291.

[61] Q. Xu, R. Zheng, W. Saad, and Z. Han, "Device fingerprinting in wireless networks: Challenges and opportunities," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 94–104, 1st Quart., 2016.

[62] S. U. Rehman, K. Sowerby, and C. Coghill, "Analysis of receiver front end on the performance of RF fingerprinting," in *Proc. IEEE 23rd Int. Symp. Pers., Indoor Mobile Radio Commun.*, Sep. 2012, pp. 2494–2499.

[63] B. P. Müller, L. J. Wong, W. H. Clark, and A. J. Michaels, "A real-world dataset generator for specific emitter identification," *IEEE Access*, vol. 11, pp. 110023–110038, 2023.

[64] (2023). *Great Scott Gadgets*. [Online]. Available: https://web.archive.org/web/20230418020802/https

[65] A. Ali and G. Fischer, "The phase noise and clock synchronous carrier frequency offset based RF fingerprinting for the fake base station detection," in *Proc. IEEE 20th Wireless Microw. Technol. Conf. (WAMICON)*, Apr. 2019, pp. 1–6.

[66] L. Bowler and A. J. Michaels, "Evaluating HPA effects on chaotic sequence spread spectrum detectability," *IEEE Access*, vol. 11, pp. 104155–104164, 2023.

[67] P. Draxler, J. Deng, D. Kimball, I. Langmore, and P. M. Asbeck, "Memory effect evaluation and predistortion of power amplifiers," in *IEEE MTT-S Int. Microw. Symp. Dig.*, Sep. 2005, pp. 1549–1552.

[68] A. J. Michaels and L. J. Wong, "Multinomial-based decision synthesis of ML classification outputs," in *Modeling Decisions for Artificial Intelligence*. Cham, Switzerland: Springer, 2021, pp. 156–167.

BRAEDEN P. MULLER (Member, IEEE) received the B.S. and M.S. degrees in computer engineering from Virginia Tech, Blacksburg, VA, USA, in 2021 and 2023, respectively.

In 2021, he participated in the Virginia Microelectronics Consortium (VMEC) Summer Scholar Program performing research with the University of Virginia, Charlottesville, VA, USA. In 2022 and 2023, he was a Product Engineering Intern and a Digital Design Intern in analog signal chain with Texas Instruments. He is currently a Graduate Research Assistant with the Virginia Tech National Security Institute, Blacksburg. His research interests include design and automated test of complex hardware systems.

LAUREN J. WONG received the B.A. degree in computer science and in mathematics from Oberlin College, Oberlin, OH, USA, in 2016, and the M.S. and Ph.D. degrees in electrical engineering from Virginia Tech, Blacksburg, VA, USA, in 2018 and 2023, respectively.

From 2018 to 2020, she was a Research Associate with the Hume Center for National Security and Technology, Virginia Tech. Since 2020, she has been a Research Scientist with Intel Corporation, AI Laboratory. Her research interests include various aspects of RFML, from applications to operational considerations and user assurance, and data efficient learning and AI robustness.

ALAN J. MICHAELS (Senior Member, IEEE) received the degree in electrical and computer engineering, in applied mathematics, and in operations research from Georgia Tech, Atlanta, GA, USA, and the M.B.A. degree from Carnegie Mellon, Pittsburgh, PA, USA.

Previously, he worked a decade as a Systems Engineer, a Researcher, and the Department Head of Harris Corporation, specializing in secure communication. He is currently a Professor with the Virginia Tech's Bradley Department of Electrical and Computer Engineering (ECE) and the Director of the National Security Institute's Spectrum Dominance Division, Blacksburg, VA, USA.

Dr. Michaels is a Professional Engineer in the Commonwealth of Virginia, has 45 issued U.S. patents, and is a fellow of the National Academy of Inventors.