

T-PRIME: Transformer-based Protocol Identification for Machine-learning at the Edge

Mauro Belgiovine, Joshua Groen, Miquel Sirera, Chinenye Tassie, Ayberk Yarkin Yıldız, Sage Trudeau, Stratis Ioannidis, Kaushik Chowdhury

Department of Electrical and Computer Engineering, Northeastern University Boston, MA
Email: {belgiovine.m, groen.j, sirera.m, tassie.c, yildiz.ay, trudeau.s, ioannidis, k.chowdhury}@northeastern.edu

Abstract—Spectrum sharing allows different protocols of the same standard (e.g., 802.11 family) or different standards (e.g., LTE and DVB) to coexist in overlapping frequency bands. As this paradigm continues to spread, wireless systems must also evolve to identify active transmitters and unauthorized waveforms in real time under intentional distortion of preambles, extremely low signal-to-noise ratios and challenging channel conditions. We overcome limitations of correlation-based preamble matching methods in such conditions through the design of T-PRIME: a Transformer-based machine learning approach. T-PRIME learns the structural design of transmitted frames through its attention mechanism, looking at sequence patterns that go beyond the preamble alone. The paper makes three contributions: First, it compares Transformer models and demonstrates their superiority over traditional methods and state-of-the-art neural networks. Second, it rigorously analyzes T-PRIME’s real-time feasibility on DeepWave’s AIR-T platform. Third, it utilizes an extensive 66 GB dataset of over-the-air (OTA) WiFi transmissions for training, which is released along with the code for community use. Results reveal nearly perfect (i.e. > 98%) classification accuracy under simulated scenarios, showing 100% detection improvement over legacy methods in low SNR ranges, 97% classification accuracy for OTA single-protocol transmissions and up to 75% double-protocol classification accuracy in interference scenarios.

Index Terms—deep learning, protocol classification, edge computing

I. INTRODUCTION

The increasing demand for wireless services has caused a scarcity of spectrum resources [1]. This results in congested wireless spectrum environments as various communication protocols coexist in the same frequency bands [2]. Unauthorized transmissions further raise security concerns, posing risks to critical operations [3]. Detecting diverse protocols in crowded spectrums allows for intelligent strategies to mitigate interference, improve spectral efficiency, and enhance overall wireless system performance, benefiting regulatory bodies, network operators, and researchers.

• **Problems with legacy methods of protocol classification.** Legacy methods require extensive prior knowledge of protocols, leading to challenges in detecting new protocols. Even with full protocol knowledge, correlation-based methods suffer from reduced detection accuracy in low SNR conditions and when multiple protocols overlap in frequency. To address these challenges, T-PRIME (shown in Fig. 1) adopts a Machine

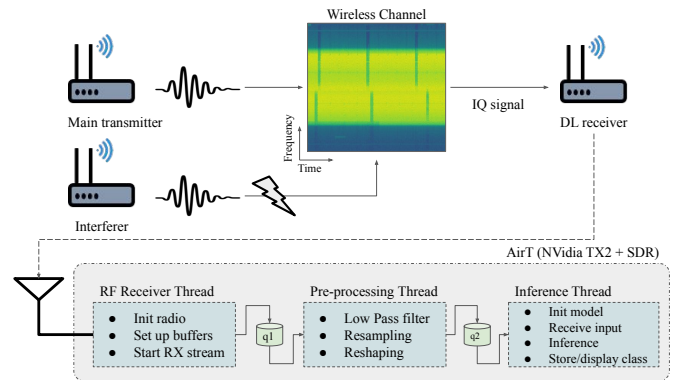


Fig. 1: T-PRIME system overview and receiver design.

Learning (ML) based approach, performing protocol classification without requiring specific channel boundaries or symbol rates, even with overlapping signals and low SNR conditions.

Legacy protocol detection methods are integrated into the RF receiver network interface card, enabling fast preamble correlations in hardware. However, updating the system for new protocols leads to backward compatibility issues and potential detection errors in challenging wireless channels. On the other hand, software-defined radio (SDR) systems offer flexibility but introduce higher latencies due to data transfer and complexities in I/O and buffer management. This paper demonstrates that using edge devices with CPU and GPU on a system-on-a-module (SOM), along with careful software design, can overcome these limitations and enable real-time processing for complex ML wireless applications on an SDR-based edge device.

• **Problems with WiFi protocol classification.** We showcase the potential of SDR based, ML enabled architectures by focusing on a challenging wireless scenario involving protocol classification. Specifically, we classify different 802.11 WiFi standards, including 802.11b, 802.11g, 802.11n, and 802.11ax. These standards share similar preamble structures and can operate with the same MCS, frequency bands, channel bandwidths, and transmit power. Importantly, this classification is distinct from and more difficult than prior works [1]–[3], which classify diverse protocols like LTE, WiFi, and Bluetooth from a candidate pool and lack a real time deployment.

• **Proposed solution: T-PRIME.** We demonstrate the effective

use of Transformer neural networks [4] for the above WiFi protocol classification problem. Transformers have gained significant interest in the deep learning community for their ability to ingest long sequences of data and efficiently correlate temporal features far from each other, which is particularly useful for natural language processing tasks [5]. However, their application to wireless communications is still in a nascent stage, with some works focusing on modulation classification [6]–[8]. Although Transformer-based models excel at processing 1D sequential and time-dependent information, adapting them to work with 2D sequences of IQ samples is not straightforward. Additionally, it is crucial to evaluate ML models intended to be deployed as real-time classifiers on resource-constrained devices to understand the trade-offs between accuracy and delay. Our contributions are as follows:

- We create, and publicly release [9], a protocol classifier called T-PRIME based on Transformer architecture and compare it to both classical, preamble based processing and ML based modulation detectors using a large data set of simulated WiFi signals.
- We rigorously evaluate the performance of each of these protocol classification methods under various Single-Input-Single-Output (SISO) wireless channel settings, including randomized channel realization, different signal-to-noise ratio (SNR) levels, and multiple channel models. In all these settings, T-PRIME outperforms state-of-the-art legacy and ML solutions available for classification of raw IQ signals.
- We implement T-PRIME on DeepWave’s Artificial Intelligence Radio Transceiver (AIR-T) and evaluate performance in realistic scenarios. Drawing from this experience, we provide insights that will allow building a generic ML-based wireless signal processing pipeline on SOM architectures that operates in near real-time using sequences of time-domain IQ samples.
- We obtain and release the first-of-its-kind dataset comprising over-the-air (OTA) WiFi-signals for different protocol versions. This dataset includes transmissions collected in diverse environments, encompassing both single and overlapping transmission scenarios.

II. RELATED WORK

• **Preamble correlation Methods.** Frame preamble usage for protocol identification and synchronization involves detecting a predefined symbol sequence at the start of a transmitted signal. For example, the 802.11b long preamble consists of 128 scrambled 1s and 16 SFD marker bits, while the 802.11n preamble includes at least 9 symbols spread across all 52 OFDM sub-carriers.

Traditional signal processing techniques use energy detection and correlation [10], [11], requiring prior knowledge of the unique preamble sequences. However, they can be vulnerable to noise, interference, and channel variations. In Sec. IV-A, we demonstrate how legacy approaches fail at low SNR, while T-PRIME achieves highly accurate protocol classification under the same conditions.

• **ML-based methods.** Protocol classification via ML methods is still in a nascent stage. Zhang *et al.* [1] perform protocol detection between WiFi, LTE, and 5G signals using a multi-layer bidirectional RNN. They use 512 IQ pairs with a sliding window as inputs to their RNN. Similarly, Schmidt *et al.* use a CNN to classify between WiFi, Bluetooth, and ZigBee signals [2]. Jagannath and Jagannath [3] perform both modulation classification and signal classification via CNNs by distinguishing between a wide range of different protocols including AM Radio, Bluetooth, WiFi, and IEEE 802.15.4 (Zigbee). However, these approaches do not address the challenge of classifying similar protocols, such as different WiFi standards.

There is a growing body of work that uses ML for the different problem of modulation classification. These include CNN-based architectures [12]–[17], CNNs with a self-attention mechanism [18] and RNNs [19]. There is a smaller number of prior works that introduce Transformer based ML models for the modulation classification task. For example, Zhen *et al.* create spectrograms from IQ data and use an image Transformer [7]. Hamidi-Rad and Jain [8] first use a CNN to transform the synthetically generated IQ samples to a 1 dimensional latent space input to the Transformer. They utilize this input CNN layer to help extract features in a location independent way because their IQ samples are not synchronized and start and stop at random locations. Cai *et al.* [6] state that IQ data can not be directly applied to the Transformer because it represents only two sequences (I and Q) and a Transformer requires several sequences. They utilize a four step pre-processing technique that creates “patches” of IQ samples that are temporally close to each other. Finally, they use a linear projection embedding layer with additional sequence and positional encoding before the Transformer encoder layer. In contrast to these works that use complicated signal pre-processing and show results on offline inference only, T-PRIME shows how IQ samples can be directly provided as the input to a Transformer and performs well in realistic environments on an edge platform.

III. T-PRIME DESIGN

Transformers, as detailed in [20], employ attention mechanisms, or activation masks, which filter information within a layer for transmission to the subsequent layer. In *self-attention*, the network generates this mask based on input and/or preceding layers. Transformers excel in modeling varied dependencies among sequence elements, independent of their positions in input or output sequences, unlike RNNs. Crucially, these adaptable dependencies in Transformers are shaped by and learned from input data.

The general Transformer architecture, as described by Vaswani *et al.* [4], is composed of several stacked transformer layers shown in Fig. 2, typically forming an encoder-decoder structure. Each layer has two sub-layers: a multi-head attention mechanism and a feed-forward network. The input data, along with positional encoding, is fed into the Transformer. Next, it is passed to the multi-head attention mechanism. There are residual connections and normalizations around the two

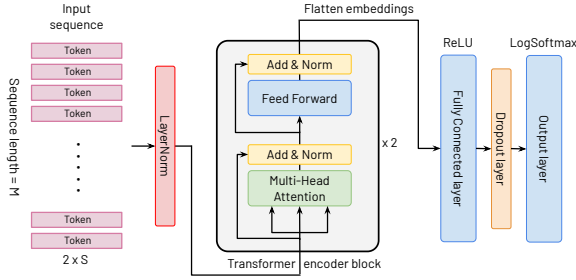


Fig. 2: Transformer-based model architecture.

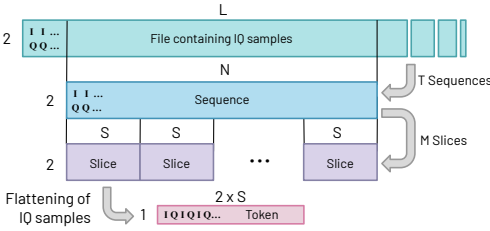


Fig. 3: Transmission split into sequences, slices and, tokens.

sub-layers. Implementations generally differ in where the normalization occurs and in the specific implementation of the feed-forward layer.

• **Adapting Transformer for IQ Input.** The classical Transformer architecture requires a token sequence as input, typically achieved in NLP tasks through word embeddings and positional information. However, feeding IQ data to the Transformer is not straightforward, leading to complex approaches in prior work [6]–[8]. Initially, we used spectrograms and a vision Transformer [21], but this yielded lower performance compared to a baseline CNN [12].

To address this, we adopted an NLP-inspired approach of breaking the IQ samples into smaller “sentences” and “words”. Our dataset initially consists of $2 \times L$ samples, which are partitioned into sequences of dimension $2 \times N$ (sentences), then further sliced into M slices of size $2 \times S$ (words). Instead of defining a linear embedding for these “words” as in traditional NLP Transformers and since I and Q real values are defined in the $[0,1]$ interval, we treat the $2S$ -dimensional space defined by the set of all possible “words” as our embedding space. Thus, we simply flattened them by interleaving the I and Q samples into a vector of dimension $1 \times 2S$, eliminating the need for linear embeddings (see Fig. 3).

• **Transformer Architecture Design.** Our proposed Transformer architecture (Fig. 2) omits positional encoding as it did not affect accuracy. Since our task is classification, we utilize only the encoder part of the Transformer. Contextualized embeddings from the encoder layers flow into a Fully-Connected layer and an Output layer representing the four classes in our classification. The architecture includes an initial normalization layer, *LogSoftmax* as the final activation

Standard	Bandwidth	MCS	Modulation	#PPDU/burst	#IQ/burst
802.11b*	11MHz	QPSK	DSSS	1	18112
802.11g	20MHz	16-QAM (1/2 & 3/4)	OFDM	4 & 4	32960
802.11n	20MHz	16-QAM (1/2 & 3/4)	OFDM	5 & 6	31340
802.11ax	20MHz	16-QAM (1/2 & 3/4)	OFDMA	5 & 6	31640

TABLE I: Synthetic dataset specifications. 802.11b* is upsampled from 11MHz to 20MHz for consistency with other signals. PPDU per burst are reported for each coding rate.

function and *NLLLoss* as loss function.

We evaluate T-PRIME using a systematic two-phase approach. In the first phase, we generate a large synthetic dataset to compare T-PRIME with the legacy correlation-based WiFi preamble detection algorithm and ML models from prior work. T-PRIME achieves superior results on synthetically generated datasets compared to these other approaches. In the second phase, we conduct an in-depth evaluation of T-PRIME on a more realistic dataset consisting of over-the-air transmissions collected in diverse and complex RF environments. Additionally, we implement T-PRIME in the AIR-T testbed to demonstrate its accuracy and real-time inference latency.

IV. OFFLINE EVALUATION OVER SYNTHETIC DATA

1) *Dataset Description:* We used MATLAB’s WLAN Waveform Generator to generate 802.11b, 802.11g, 802.11n and 802.11ax waveforms. This dataset contains baseband IQ samples according to the specifications shown in Table I. For each protocol, we generate 2000 Physical layer Protocol Data Units (PPDU) bursts, each comprising of multiple packets carrying a 1000 bit random payload, saved into individual files. As 802.11b does not support a bandwidth/sampling rate of 20MHz, the waveform is upsampled from 11MHz to 20MHz to ensure waveforms have the same channel bandwidth and ensure the problem is challenging for the classifier.

2) *Training Procedure:* Once the burst signals have been generated, we use a PyTorch data generator to train our models that operates on a per-training batch basis to dynamically load into memory each file individually and simulate transmission through a channel model at an SNR level randomly chosen. We consider the following options for (a) channel models: TGn, TGax, Rayleigh, no channel impairment; and (b) a random SNR level defined between -30 dB and 30 dB, applied using Additive White Gaussian Noise (AWGN). We develop a Python wrapper for Matlab’s channel models standard compliant implementation and employ Model-B multipath delay profile for TGn and TGax with device distance of 3m, while for Rayleigh model we use a single Non-Line Of Sight (NLOS) path with an average path gain of -3 dB and delay of $1.5e-9$ seconds.

Input sequences are pre-generated before training procedure for the complete set of baseband signals in the dataset. Once sequence indexes are generated for each dataset burst, they are randomly shuffled and then separated into training and validation set with a split of 80% and 20%. When a mini-batch of sequences is generated at training time, the complete burst signal is retrieved and passed through a random channel model and AWGN. Finally, the input sequence is retrieved from the

complete distorted signal and divided into slices as described in Fig. 3. Multi-path distortions thus impact the complete set of IQ samples in a batch on a per-sequence basis, simulating realistic transmission.

While dynamic signal augmentation is performed for both train and validation data, to evaluate the performance of our models in similar conditions we use an additional static test set, composed of 500 bursts for each protocol following the configuration in Table I. Such test transmissions are pre-processed by creating multiple copies of the bursts that are passed through a random channel instance for all the considered channel models, for a total of an additional 2000 test packet bursts. Then, at test time, we create multiple sequences from each signal copy and evaluate them on discrete SNR levels ranging between -30 and 30 dB in increments of 5 dB.

The proposed T-PRIME design was finalized using the same training procedure above and a hyperparameters tuning phase based on a grid search on the validation set. The chosen training configuration for transformers architectures is 5 training epochs, batch size of 122 and learning rate $2e-4$. We found that a stack of 2 transformer layers is optimal, along with two distinct sets of values for sequence length and slice length that provided nearly perfect results. The first model, named Small (SM) Transformer, utilized an input sequence of $M = 24$ and tokens with a length of $S = 64$. The second, Large (LG) model, employed a sequence of $M = 64$ tokens with a length of $S = 128$ each. The resulting number of parameters of SM and LG architectures are 1.6M and 6.8M, respectively.

A. Comparison with state-of-the-art legacy and ML methods

We first show how T-PRIME approach can effectively improve the state-of-the-art over legacy approaches. Specifically, we use the same static test set described in Sec. IV-1 for a simulation study that processes each test transmission using the correlation-based procedure specified in the standard [10]. We employ a set of multiple decoders provided by Matlab that use the legacy method to detect the following preambles formats: 1) VHT (i.e. 802.11n), 2) HE-SU (i.e. 802.11ax), 3) Non-HT (i.e. 802.11g/b). We then process our test transmissions through these functions for each channel model considered. For each test signal, we simulate 100 transmissions over random channel instances for each channel model and SNR levels considered in our evaluation procedure described before. To report the accuracy of legacy procedures, we consider the inverse Packet-Detection-Rate (i.e. $1 - \text{PDR}$), defined as the ratio of correct preamble format detections over the total number of simulated packets transmissions, and compare it with accuracy of proposed architectures under the same statistical channel and SNR conditions. We evince from Fig. 4a that both our T-PRIME models drastically outperform the legacy detection mechanism, achieving near-perfect detection accuracy already at $\text{SNR} = 0$ dB, while traditional methods completely fail to operate under such circumstances. This demonstrates how these models are still able to extract and learn powerful time-domain dependencies from received

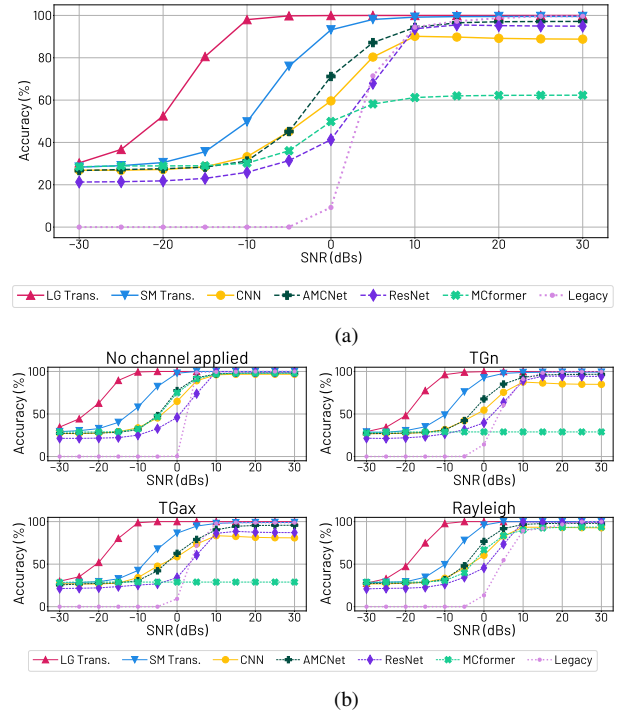


Fig. 4: (a) Comparison between SM, LG Transformer-based architectures and state-of-the-art signal classification models tested on all channel models and different SNR conditions. The LG Transformer achieves the best overall accuracy and it retains high accuracy (i.e. $> 98\%$) for SNR as low as -10 dB. All models improve under optimal hyperparameter search (see Appendix), but the LG Transformer still outperforms all other methods. (b) Performance comparison for each individual channel model tested during simulation.

signals even in presence of strong noise or low reception power.

After establishing superiority of our approach over legacy methods, we compare our model to other state-of-the-art ML methods for wireless signal classification. While our problem formulation differs from modulation classification, it serves as a relevant example of architectures processing sequences of raw IQ samples for similar tasks. We compare T-PRIME against three state-of-the-art models in this domain: ResNet [15], AMCNet [18], and MCformer [8], using the hyperparameters reported as optimal by their respective authors; we also present supplementary findings obtained through additional hyperparameter exploration in Appendix. For further comparison, we trained a 1D CNN model inspired by [12] with a wider input size and slightly larger set of parameters. We used a common training configuration for all models with a batch size of 512, $M = 1$ for input generation, *CrossEntropyLoss* as the loss function, and a learning rate of 0.001 with exponential decay to 0.0001 when a plateau in the validation loss is reached. Table II lists all the models we compared, along with their input size and total number of parameters.

Fig. 4 illustrates that both T-PRIME LG and T-PRIME SM consistently outperform all other methods across the entire range of evaluated signal-to-noise ratios (SNRs). Notably, T-

Model	Input size ($M \times S$)	# Parameters
T-PRIME LG	64×128	6.8M
T-PRIME SM	24×64	1.6M
1D CNN [12]	1×512	4.1M
ResNet [15]	1×1024	162K
AMCNet [18]	1×128	462K
MCformer [8]	1×128	78K

TABLE II: Comparison between proposed architectures and state-of-the-art MCS classification architectures.

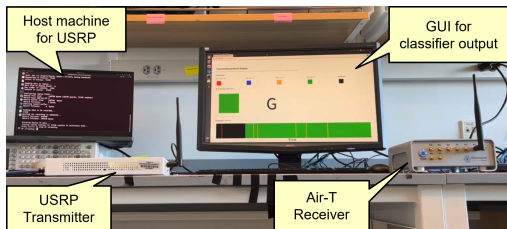


Fig. 5: T-PRIME running real-time classification in an AIR-T receiver. 802.11g protocol is shown to be correctly detected in this instance.

PRIME LG achieves a remarkable $> 60\%$ improvement over all other models at $\text{SNR} = -10$ dB, achieving 98% accuracy. From these experiments, two key points can be deduced: 1) T-PRIME’s performance benefits from a larger parameter space, following the trend of Large-Language-Models (LLMs) that use substantial parameter capacity to learn complex relationships [22]; 2) Our problem significantly differs from MCS classification, as the results suggest that larger amounts of temporal sequences are required to accurately identify the complex and less localized patterns in real-world wireless protocols, especially when protocols share the same MCS configurations. Finally, hyperparameter search improves all models (see Appendix); while, in this case, AMCNet surpasses the SM Transformer for low SNR, nevertheless the LG Transformer still consistently outperforms all competitors.

V. REAL-TIME IMPLEMENTATION ON AIR-T SDR

SDR-based systems offer flexibility in the receiver chain and seamless integration with AI/ML modules. However, they can introduce higher latency due to challenges like I/O and buffer management, software dependencies, and distribution of computing resources for signal processing tasks. To explore these systems’ challenges and showcase T-PRIME’s real-time efficiency, we implemented it on Deepwave Digital’s AIR-T model AIR7101, as shown in Fig 5. The AIR-T, built on NVIDIA’s Jetson TX2 module, includes a CPU with 4 ARM A-57 cores, 2 ARM Denver2 cores, and a Pascal 256-core GPU with 8 GB of shared memory [23]. This design resolves traditional SDR system issues by integrating resources on one platform, including SDR with FPGA, networking, periphery connections, an Ubuntu-based OS, and open source APIs.

For real-time protocol classification, low latency is crucial. Hence, we designed a pipeline (Fig. 6) with three primary functions running in parallel: receiver, signal processing, and ML inference. This parallelization removes concurrent dependencies and reduces prediction delay. It also allows individual

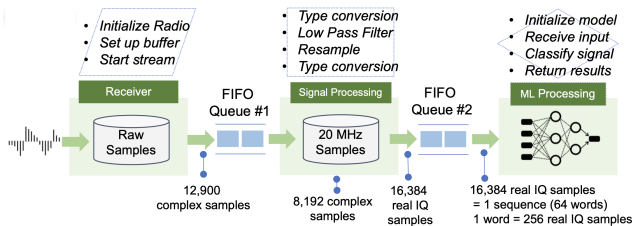


Fig. 6: T-PRIME pipeline overview showing three steps of the receiver, signal processing and ML processing modular blocks.

function modification or improvement without affecting others. For instance, we deployed the same Transformer model using two frameworks (PyTorch and TensorRT) transparently. Each thread includes timing functions to calculate the average time for function completion. Additionally, we established two thread-safe FIFO queues, q_1 between receiver and signal processing threads, and q_2 between signal processing and ML inference threads. Below, we briefly describe each of these three threads; see also our code [9] for further details.

1) *Receiver Thread*: The main function of the receiver thread is to capture the raw OTA signals. The receiver thread defines several parameters such as sample rate (31.25 MHz), gain mode, and frequency for the AIR-T receiver using SoapySDR APIs. After initializing the SDR, it creates a data buffer to receive samples from the SDR. After these one time initialization steps are complete, the receive thread enters a continuous loop that reads the buffer, ensures it received the expected number of samples, and writes the vector of samples to q_1 if there is space available.

2) *Signal Processing Thread*: The signal processing thread’s main function is to down-sample the received signal to 20 MHz and convert it to the format required by the ML inference thread. After reading the raw sample vector from q_1 , the thread converts the data type from interleaved shorts to `numpy.complex64`, normalized to the range of $[-1, 1]$. Then, we down-sample the signal from 31.25 MHz to 20 MHz, the standard channel bandwidth used in 802.11 WiFi protocols. Finally, the complex vector is converted into a real vector by interleaving the I and Q (real and imaginary) components, as shown in Fig. 3. If space is available, the thread writes this vector to q_2 .

3) *ML Inference Thread*: This thread is the core of our pipeline, performing the actual ML inference on the input samples. We deployed two different versions of this thread; a PyTorch version and a TensorRT version. TensorRT [24] uses techniques such as quantization, layer and tensor fusion, kernel tuning to optimize DL inference. Regardless of the optimization used, the thread follows the same process. After loading the model, it enters a continuous loop. If there is data available in q_2 it reads that into a tensor to use as the model input, classifies, and appends the classification label to a text file.

To handle varying signal power caused by propagation effects, we incorporate a power normalization block in the ML inference thread. This function normalizes the total power of

Dataset	Distance	Scatter	Interference	#Captures per protocol				
				ax	b	n	g	noise
RM_A_1	3m	Low	Low	200	206	200	200	0
RM_A_2	2m	Low	Medium	200	200	200	200	0
RM_B_1	1m	Med	Medium	73	200	200	200	0
RM_B_2	1m	Med	Medium	500	500	500	500	500
RM_C_1	1m	High	High	100	100	100	100	0
RM_C_2	1m	High	High	500	500	500	500	100

TABLE III: OTA single-protocol datasets comparison. Each capture file contains 198080 IQ samples - approximately 10 ms duration.

received signals to a nominal power of 1W. Given a generic discrete complex signal $s \in \mathbb{C}^N$ of N samples, we can obtain its normalized version \hat{s} as follows: $\hat{s} = s / \sqrt{\sum_{i=1}^N |s_i|^2}$ where the denominator is the Root-Mean-Square (RMS) value of the incoming signal. This block makes our model more robust to power and SNR variations, eliminating the need to collect data samples and train our models under multiple power profiles.

VI. T-PRIME AIR-T TESTBED EVALUATION

We evaluated the performance of our T-Prime implementation through: (a) the creation of an Over-the-Air (OTA) dataset, collected from the deployment of the testbed in five different locations, and use this dataset to conduct additional offline evaluation of T-Prime, (b) power and latency profile characterization of our implementation, (c) real-time, live tests of a deployment of our entire pipeline evaluating our models in a new environment. We describe each of these experiments below.

A. Experimental Setup

1) *Over-the-Air (OTA) Experiments*: In order to evaluate the performance of our proposed approach in a real-world setting, we collect OTA transmissions in a variety of conditions, including different indoor scenarios and transmission power levels. To collect this data, we re-use the same receiver and signal processing blocks. Instead of passing the resulting vectors to the ML inference block, we write them to a file for offline training and testing.

• **Dataset collection.** We collected a large OTA dataset, providing a robust collection of real-world wireless channel conditions. Table III summarizes the environments for each collection set. Room A is a large open classroom, while rooms B and C are small lab rooms. The level of interference was assessed based on visible WiFi access points and other known transmitters in the 2.4 GHz band. The level of scatter considers the ratio of room dimensions to the space used by equipment and furniture. Each room’s data was collected on two separate days (_1 vs _2). We also collected noise/background samples to measure the noise floor and test classifier performance when no active transmission occurred.

For each protocol, we generated standard compliant WiFi frames, including preamble and random data payload in base-band. We continuously transmitted these signals and captured them at the receiver side after traveling through the wireless medium. Each capture is approximately 10 ms in duration. The transmitter was an Ettus USRP X310 with 30 dBm transmission power, and both transmitter and receiver were

Configuration	Incumbent	Interferer
C1	802.11g	802.11n
C2	802.11b	802.11ax
C3	802.11b	802.11g
C4	802.11b	802.11n
C5	802.11g	802.11ax
C6	802.11n	802.11ax

TABLE IV: Protocols configuration for overlapping cases.

Dataset	R	f_s^{Rx}	f_c^{Rx}	f_c^{Tx1}	f_c^{Tx2}	#Capt./Config	#IQ/Capt.
RM_C_O1	25%	20	2.442	2.442	2.457	200	198080
RM_C_O1	50%	20	2.442	2.442	2.452	200	198080
RM_C_O2	25%	62.5	2.45	2.442	2.457	200	309500
RM_C_O2	50%	62.5	2.45	2.442	2.452	200	309500

TABLE V: OTA overlapping dataset parameters. f_s^{Rx} (expressed in MHz) is receiver sampling rate. f_c^{Rx} , f_c^{Tx1} and f_c^{Tx2} (expressed in GHz) are receiver central frequency, central frequencies of first transmitter (incumbent) and central frequency of second transmitter, respectively.

set to a central frequency of $f_c = 2.442$ GHz, corresponding to WiFi Channel 7. In both training and testing phases, we applied power normalization, unless otherwise specified.

• **Single-protocol OTA dataset.** The above data collection results in an offline OTA dataset summarized in Table III. We split this into a training and test datasets in two ways. In our *time-split*, for each day of data collection, we consider the first 80% of samples collected per protocol as a training set, and the remaining 20% as a test set. In our *scenario-split*, we perform a cross-validation study across all the location-specific datasets. We train our classifier on all OTA collections except one, which we use in its entirety to test the resulting model. In both cases the test data channel conditions are unseen during the training phase.

• **Overlapping-protocol OTA datasets.** We stress-tested T-PRIME in the presence of spectral overlaps, a scenario where traditional preamble detection techniques struggle due to frame collisions corrupting the preamble structure. In RM_C, we collected two additional datasets with signals transmitted concurrently, overlapping at 25% and 50% spectral ratios. We considered six different overlapping protocol configurations (see Table IV). The datasets differed in the observed wireless spectrum resolution, achieved by varying the receiver sampling rate and central frequency.

Our goal was to test if our Transformer classifier could recognize both incumbent signals and the protocol of the interfering transmission using raw IQ time-domain samples. We defined two additional datasets with different receiver configurations: O1) the incumbent and receiver tuned to the same WiFi channel (WiFi Ch. 7) and sharing the same bandwidth, and O2) the receiver observing a wider bandwidth and tuned to an arbitrary frequency (2.45 GHz) to observe both overlapping signals in the frequency domain. Table V summarizes the parameters used for the overlapping data collections, and Fig. 7 visualizes each receiver configuration with C5 type (802.11g / 802.11ax) transmissions as an example.

We adapted our classifier from a single to multi-label output by replacing the last *LogSoftMax* activation function with *Sigmoid* at each output neuron and using *BCELoss* as

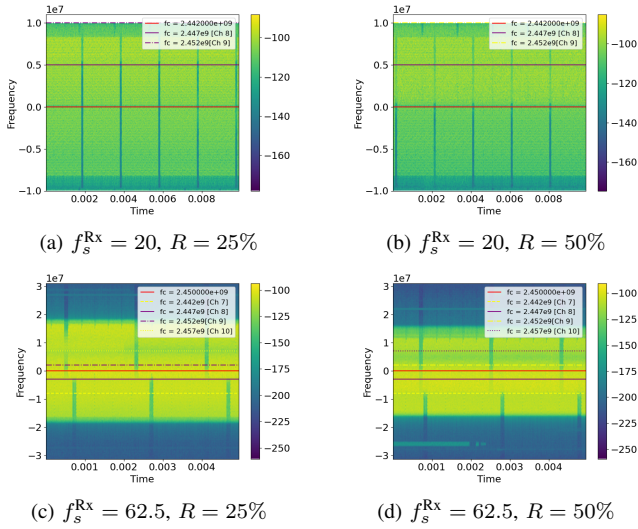


Fig. 7: Example spectrograms of recorded transmissions used in the interference dataset. Samples are shown for each configurations of overlapping rate R and sampling rate f_s^{Rx} expressed in MHz.

loss function. We trained our classifier on all collected OTA transmissions, including single protocol and overlapping ones, using the *time-split* method with the same 80/20 split for training/testing. Note that now we include both non-overlapping and overlapping cases, and signals with different sampling rates (e.g. *_O2* datasets) in the training set. This challenging dataset forces the Transformer to learn different patterns and ultimately classify the protocols irrespective of their sampling rate.

2) *Latency and Power Profile Experiments*: The AIR-T has 6 power profiles, providing several different combinations of CPUs enabled and various clock rates for the CPUs and GPU. We performed a thorough analysis of all 6 configurations to understand the constraints of our system along with the benefit and cost of each option. We used NVIDIA’s JTOP utility to monitor power consumed while T-PRIME performed real-time protocol classification.

3) *Real-Time Experiments*: We also evaluate the accuracy of T-PRIME LG in classifying OTA signals in real time. For this experiment we train T-PRIME using the entire single protocol OTA dataset (Table III). We deploy T-PRIME in two new locations, never seen during training. In both rooms, the transmitter is approximately 3m from the T-PRIME receiver. We sequentially generate and transmit each of the four 802.11 protocols in the same manner as described in Sec. VI-A1. While continuously transmitting a given protocol, we capture 1001 live predictions made by T-PRIME. Each experiment lasts approximately 10s.

B. Results

1) *OTA Dataset Evaluation*: First, we show the performance of our classifier on OTA signals.

• **Single-protocol results.** Fig. 8 shows that both LG and SM models achieve an average classification accuracy $\geq 99\%$

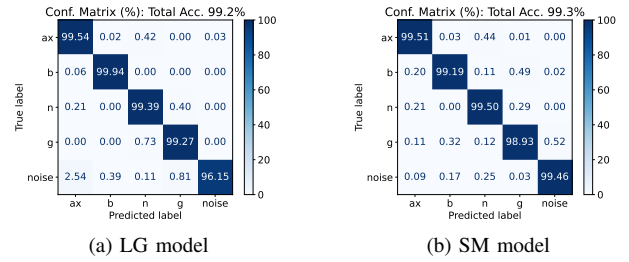


Fig. 8: WiFi protocol classification performance on the complete OTA test dataset for single transmissions. Comparison between Large (LG) and Small (SM) Transformer models.

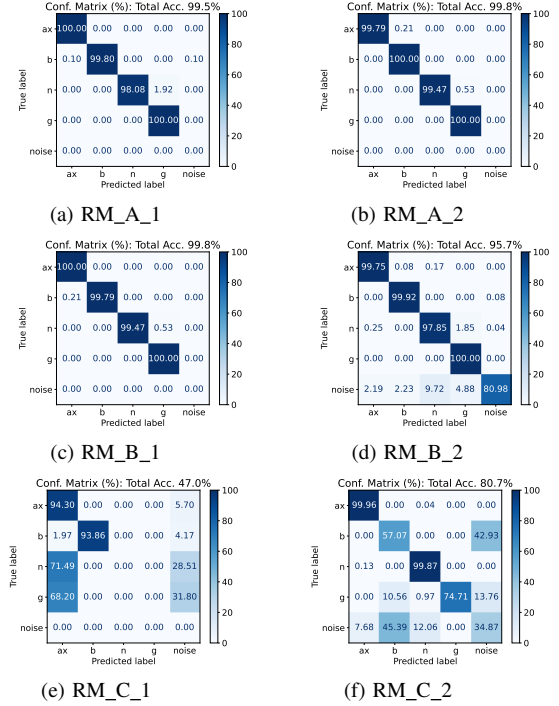


Fig. 9: Cross-validation using LG classifier over the different OTA data collection instances. Each plot indicates what dataset is missing from the training set and the resulting model performance on that specific dataset at inference time.

on *time-split* test data, with SM architecture slightly outperforming the LG one, possibly due to *overfitting* for the LG model. Intuitively, a smaller variation of input patterns may be easier to *learn* compared to randomly simulated scenarios in the synthetic dataset. Fig. 9 shows the effect of the cross-validation on *scenario-split* data. This highlights that RM_C constitutes a more unique and challenging environment for our protocol classifier compared to other locations.

• **Overlapping-protocol results.** Table VI shows the average test accuracy on the complete OTA test dataset and, separately, a detailed view of accuracy for each individual overlapping transmission dataset. We use three different metrics to determine the performance: 1) percent of “exact” classifications, i.e. correctly classifying both protocols present in the received signals, 2) percent of times we correctly recognize

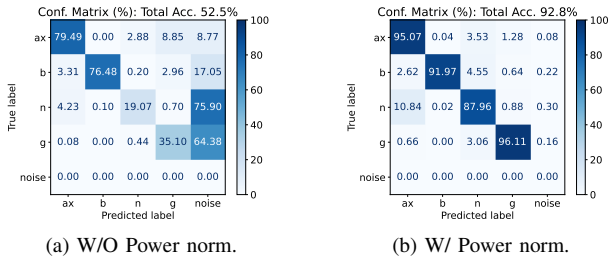


Fig. 10: Impact of power normalization for low SNR conditions, i.e. $P_{Tx} = 0$ dBm, on test data collected in RM_C.

Metric	Dataset				
	OTA (all)	RM_C_O1		RM_C_O2	
% Exact (2/2)	87.5	$R = 25\%$ 55.8	$R = 50\%$ 72	$R = 25\%$ 75.7	$R = 50\%$ 66
% Single (1/2)	98.2	99.9	99.9	94.4	93.1
AUC	0.98	0.878	0.954	0.963	0.929

TABLE VI: Average test accuracy and AUC. Classifier is trained and tested over data covering all OTA locations, including single and overlap cases.

a “single” protocol and 3) the Area-Under-the-Curve (AUC) of the trained classifier. These results show that our model tested on RM_C_O2 is still able to correctly detect at least one protocol 94.4% and 93.1% of times for 25% and 50% cases, respectively, and even able to detect both signals at the same time with 75.7% and 66% of accuracy for $R = 25\%$ and $R = 50\%$, respectively, which is not possible using legacy wireless hardware. Finally, we note that single-protocol accuracy is nearly perfect for RM_C_O1 (i.e. 99.9% in both overlap configurations), which suggests that incumbent signals that occupy most of the observed spectrum are correctly classified the majority of times.

Table VII presents Precision¹, Recall² and Support samples for each individual protocol in all receiver configurations. Precision in both RM_C_O1 cases shows higher scores for protocols 802.11b, 802.11g, and 802.11n, with a particular emphasis on 802.11b and 802.11g, as these protocols have more samples as incumbents than the others. Overall, the average Precision and Recall scores for all overlapping configurations and protocols are approximately 0.87 and 0.86, respectively. These results demonstrate the accuracy and completeness of positive predictions achieved by the proposed multi-protocol classifier architecture in this challenging interference scenario.

• **Impact of power normalization.** We also study the impact of power normalization on an additional set of data collected under different transmission powers. Specifically, for this experiment we train our model on data collected from all locations (see Table III) but we test it on a completely different set of samples collected in RM_C with transmissions performed at power $P_{Tx} = \{30, 20, 10, 0\}$ dBm respectively. Table VIII shows the average accuracy for each of the power profiles when power normalization is applied. While the average accuracy is consistently improved for all protocols,

¹Precision = $TP/(TP+FP)$; TP = True Positive; FP = False Positives.

²Recall = $TP/(TP+FN)$; FN = False Negative.

Protocol	Precision		Recall		Support
OTA (all)					
802.11ax	0.94		0.93		25482
802.11b	0.94		0.98		26154
802.11n	0.9		0.88		26106
802.11g	0.91		0.91		26106
RM_C_O1	$R = 25\%$	$R = 50\%$	$R = 25\%$	$R = 50\%$	$R = 25\%$ $R = 50\%$
802.11ax	0.72	0.83	0.75	0.86	2808 2808
802.11b	1	1	1	0.99	2808 2808
802.11n	0.77	0.85	0.62	0.83	2808 2808
802.11g	0.86	0.96	0.78	0.83	2808 2808
RM_C_O2	$R = 25\%$	$R = 50\%$	$R = 25\%$	$R = 50\%$	$R = 25\%$ $R = 50\%$
802.11ax	0.98	0.97	0.97	0.9	4329 4329
802.11b	0.82	0.87	0.93	0.96	4329 4329
802.11n	0.91	0.78	0.8	0.86	4329 4329
802.11g	0.86	0.74	0.92	0.8	4329 4329

TABLE VII: Test accuracy for individual protocols in single and overlapping transmissions. Classifier is trained and tested over data covering all OTA locations, including single and overlap cases.

P_{Tx} (dBm)	W/O Power Norm.	W/ Power Norm.
0	52.5	92.8
10	93.8	95.9
20	93.9	95.3
30	97.3	97.6

TABLE VIII: Average test accuracy (%) on multiple power profiles with and without signal power normalization.

the most significant boost in performance is on the lower SNR end, reaching a staggering 40.3% performance improvement at $P_{Tx} = 0$ dBm without the need to include any additional power variation samples in the dataset.

2) **Real-Time Latency and Power Results:** Building a pipeline with modular processing blocks that can operate in parallel is one of the keys to deploying a real-time classification system on an edge device. For example, as seen in Table IX, when using TensorRT, the Signal processing and ML inference blocks take nearly the same amount of time to process a sample. However, our pipeline allows for these processes to run in parallel producing a prediction every 10.9 ms - equivalent to 92 predictions per second. Table IX shows the power consumption and time to complete each block for all power profiles except 4-MAXP CORE DENVER which failed to run our pipeline. The top performance and most efficient power profiles from our testing are in bold. There is a clear trade off between higher performance in terms of predictions per second and higher power usage. For all power profiles, TensorRT provides better performance and is more efficient.

• **Receiver thread:** The data buffer in the receiver thread presented a crucial I/O constraint that required careful engineering. A large buffer ensured enough samples for our model, but caused frequent buffer overflows, adding over 1s to reset the connection. On the other hand, a small buffer could lead to insufficient data for the ML model, requiring multiple reads of the buffer. Thus, we designed the buffer to hold a specific number of complex samples based on the ML model used. For the LG Transformer, we configured the buffer to capture 12,900 complex samples.

• **Signal processing thread:** Understanding the effects of the low pass filter and re-sampling is crucial for this block. To down-sample from 31.25MHz to 20MHz, we perform rational re-sampling with a factor of $\frac{16}{25}$. For our LG Transformer trained with a 20MHz bandwidth, we need 8,192 complex

	0-MAXN ALL	1-MAXQ	2-MAXP ALL	3-MAXP ARM	5-MAXN ARM
PyTorch					
Average Power Consumption (mW)					
CPU	1551	704	807	1457	1472
DDR	1595	842	900	999	1007
GPU	1074	486	366	434	435
SOC	1000	625	625	627	628
ALL	6933	4283	4338	5195	5242
Average Latency Time (ms)					
Reciever	0.4105	0.7252	0.44307	0.4103	0.4105
Signal Pro.	9.843	15.39	14.68	10.49	10.46
ML Thread	18.98	29.18	29.58	22.28	22.27
TensorRT					
Average Power Consumption (mW)					
CPU	1579	672	825	1515	1540
DDR	904	780	958	1131	997
GPU	404	264	436	629	578
SOC	610	598	617	668	628
ALL	5170	3923	4505	5668	5426
Average Latency Time (ms)					
Reciever	0.5095	0.6162	0.524	0.4797	0.5054
Signal Pro.	10.05	15.45	13.79	9.308	9.62
ML Thread	10.91	15.91	14.82	12.46	10.86
ML inference*	1.96	2.51	2.09	2.11	2.12

TABLE IX: T-PRIME LG pipeline performance for each block showing the **lowest latency** and **lowest power** using different power profiles and ML inference optimization. The last line reports average inference times of T-PRIME LG for individual inference operations within the ML Thread.

samples, so the input size should be $8,192 \times \frac{25}{16} = 12,800$. However, in practice, the output of this thread is slightly less than the required 8,192 samples. After empirical testing, we found that providing 12,900 complex samples as input gives the correct output size. This highlights both the complexity of deployed systems and the necessity for variable length inputs to the ML model.

- **ML inference thread:** The key optimization for this thread was moving from PyTorch to TensorRT. The vast performance gains in the ML thread for TensorRT model shown in Table IX, without any loss of accuracy, highlight the importance of using well designed functional blocks which enable optimization of one function without making any changes to the others.

- **Queue length management for real-time inference:** The above pipeline offers efficient use of computing resources, flexibility in optimizing each function, and reduced latency. However, it also presents potential bottlenecks with I/O and buffer management. To maintain low end-to-end latency, optimizing the queue length between blocks is crucial. Long queues guarantee input availability for each function, maximizing parallelization gains and prediction rate. However, it significantly increases latency from signal reception to prediction. On the other hand, very short queues may lead to idle time, reducing parallelization efficiency and prediction rate. Our pipeline’s progressive processing times allow us to set short queues without starving any processes. However, jitters in processing time, IO blocks, and hardware limitations occasionally impact processing times dramatically. We found that setting both q_1 and q_2 to a length of 2 minimized end-to-end delay while ensuring no thread was starved.

3) *Real-Time Deployment Results:* Fig. 11 shows the confusion matrix for our Real-Time evaluation where we achieve an

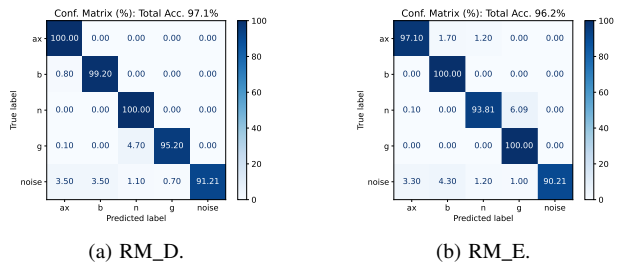


Fig. 11: T-PRIME real-time prediction results in two new environments, RM_D and RM_E, never seen during training.

incredible 96.7% overall accuracy. From the confusion matrix, we see lowest accuracy is for the noise. However, given that other WiFi transmitters are operating in the area, some of the incorrect classifications for noise may in fact be correctly identifying other transmissions. Regardless, it is clear that T-PRIME has excellent accuracy in real-time predictions, even in environments it has never been exposed to during training.

4) Discussion and Insights from Real-time Inference:

Our implementation and testing of T-PRIME on the AIR-T device for real-time protocol classification have yielded valuable insights applicable to other AI/ML-enabled edge device classification tasks. Utilizing integrated SOM edge devices significantly reduces latencies and complexities often encountered in SDR-based systems. Parallelization plays a crucial role in increasing the prediction rate. Designing key functions as separate threads allows for more efficient resource utilization and ease of optimization for each function. Understanding I/O points and buffer management is essential for real-time systems. The ML Inference thread is the most resource-intensive, and optimizing this task offers significant performance gains.

VII. CONCLUSION AND FUTURE WORK

T-PRIME achieves near-perfect classification accuracy for 4 different WiFi protocols, outperforming legacy methods in challenging conditions. T-PRIME proves superior to existing state of the art ML models at protocol detection, especially with low SNR and overlapping signals. While our current architecture employs fixed-length inputs, Transformers excel with variable length inputs, as shown by successful NLP models like ChatGPT. In the future, we will adapt our ML block to accommodate variable length inputs. As part of our real-time implementation, we identified different bottlenecks and offered insights for solutions, which can benefit other AI/ML-based classification systems running on resource-constrained edge devices. Finally, we provide public access to our code and datasets at [9].

VIII. ACKNOWLEDGMENT

This work was funded by the Defence Science and Technology Laboratory (DSTL), a part of the UK Ministry of Defence.

REFERENCES

- [1] W. Zhang and M. Krunz, "Machine learning based protocol classification in unlicensed 5 GHz bands," in *2022 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2022, pp. 752–757.
- [2] M. Schmidt, D. Block, and U. Meier, "Wireless interference identification with convolutional neural networks," in *2017 IEEE 15th international conference on industrial informatics (INDIN)*. IEEE, 2017, pp. 180–185.
- [3] A. Jagannath and J. Jagannath, "Multi-task learning approach for automatic modulation and wireless signal classification," in *ICC 2021-IEEE International Conference on Communications*. IEEE, 2021, pp. 1–7.
- [4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [5] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.
- [6] J. Cai, F. Gan, X. Cao, and W. Liu, "Signal modulation classification based on the transformer network," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 3, pp. 1348–1357, 2022.
- [7] Q. Zheng, P. Zhao, H. Wang, A. Elhanashi, and S. Saponara, "Fine-grained modulation classification using multi-scale radio transformer with dual-channel representation," *IEEE Communications Letters*, vol. 26, no. 6, pp. 1298–1302, 2022.
- [8] S. Hamidi-Rad and S. Jain, "Mcformer: A transformer based deep neural network for automatic modulation classification," in *2021 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2021, pp. 1–6.
- [9] Genesys Lab, "T-PRIME Repository," <https://github.com/genesys-neu/t-prime>, 2023, [Accessed 2024-01-05].
- [10] "IEEE Standard for Information Technology–Telecommunications and Information Exchange between systems - local and metropolitan area networks–specific requirements - part 11: Wireless lan medium access control (MAC) and physical layer (PHY) specifications," *IEEE Std 802.11-2020 (Revision of IEEE Std 802.11-2016)*, pp. 1–4379, 2021.
- [11] J. Terry and J. Heiskala, *OFDM wireless LANs: A theoretical and practical guide*. Sams publishing, 2002.
- [12] T. J. O'Shea, J. Corgan, and T. C. Clancy, "Convolutional radio modulation recognition networks," in *Engineering Applications of Neural Networks: 17th International Conference, EANN 2016, Aberdeen, UK, September 2-5, 2016, Proceedings 17*. Springer, 2016, pp. 213–226.
- [13] Y. Shi, K. Davaslioglu, Y. E. Sagduyu, W. C. Headley, M. Fowler, and G. Green, "Deep learning for RF signal classification in unknown and dynamic spectrum environments," in *2019 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*. IEEE, 2019, pp. 1–10.
- [14] H. Elyoussef and M. L. Altamimi, "Deep learning radio frequency signal classification with hybrid images," in *2021 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*. IEEE, 2021, pp. 7–11.
- [15] T. J. O'Shea, T. Roy, and T. C. Clancy, "Over-the-air deep learning based radio signal classification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 1, pp. 168–179, 2018.
- [16] C. Gravelle and R. Zhou, "SDR demonstration of signal classification in real-time using deep learning," in *2019 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2019, pp. 1–5.
- [17] T. Huynh-The, C.-H. Hua, Q.-V. Pham, and D.-S. Kim, "MCNet: An efficient CNN architecture for robust automatic modulation classification," *IEEE Communications Letters*, vol. 24, no. 4, pp. 811–815, 2020.
- [18] J. Zhang, T. Wang, Z. Feng, and S. Yang, "AMC-Net: An Effective Network for Automatic Modulation Classification," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [19] S. Rajendran, W. Meert, D. Giustiniano, V. Lenders, and S. Pollin, "Deep learning models for wireless signal classification with distributed low-cost spectrum sensors," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 3, pp. 433–445, 2018.
- [20] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.
- [21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [22] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler *et al.*, "Emergent abilities of large language models," *arXiv preprint arXiv:2206.07682*, 2022.
- [23] "AIR-T Overview - Deepwave Digital Docs — docs.deepwavedigital.com," May 2022, [Accessed 30-07-2023]. [Online]. Available: <https://docs.deepwavedigital.com/AIR-T/>
- [24] "Quick Start Guide : NVIDIA Deep Learning TensorRT Documentation — docs.nvidia.com," <https://docs.nvidia.com/deeplearning/tensorrt/quick-start-guide/index.html>, [Accessed 30-07-2023].

APPENDIX

A. Hyperparameter Exploration on Baseline Models

In this section, baseline models are further explored to find optimal hyperparameters in order to compare with T-PRIME. In all tables, metrics are accuracy and cross-entropy loss respectively. In Tables X, XI, XII, and XIII, we experimented with different values of slice length, batch size, and learning rate to find the optimal configuration over the synthetic dataset. In Table XIV, optimal hyperparameters for each baseline model is shown. Except ResNet, increasing slice length provided a better validation for the models. Additionally, decreasing the batch size allow the models train better, and smaller learning rates provide a better convergence. Afterwards, including T-PRIME LG and T-PRIME SM, all models are run for 50 epochs. Overall training and validation results including the fine-tuned baseline models and T-PRIME models are shown in Table XV, where M is the sequence length, and S is the slice length. The only difference between default and fine-tuned results over T-PRIME models are the number of epochs changing from 5 to 50, while the other (already optimized) hyperparameters remain same. We put the finalized comparisons that also include the runtimes in Table XVI.

Slice Length	Batch Size	Learning Rate	Acc. (%)	Loss
128	512	1e-3	68.8	0.672
512	512	1e-3	78.8	0.464
1024	512	1e-3	80.4	0.438
2048	512	1e-3	80.0	0.432
4096	512	1e-3	75.0	0.706
8192	32	1e-3	70.2	0.713
8192	64	1e-3	75.3	0.568
8192	128	1e-3	79.9	0.450
8192	256	1e-3	81.0	0.433
8192	512	1e-4	85.9	0.324
8192	512	2e-4	86.2	0.311
8192	512	5e-4	86.4	0.315
8192	512	8e-4	84.0	0.367
8192	512	1e-3	82.6	0.430
8192	512	5e-3	51.4	1.103
8192	512	1e-2	30.0	1.357

TABLE X: AMCNet Hyperparameter Exploration

Slice Length	Batch Size	Learning Rate	Acc. (%)	Loss
512	512	1e-3	66.3	0.770
1024	32	1e-3	74.5	0.561
1024	64	1e-3	77.1	0.517
1024	128	1e-4	65.3	0.786
1024	128	2e-4	73.3	0.580
1024	128	5e-4	76.6	0.516
1024	128	1e-3	78.3	0.486
1024	128	5e-3	29.2	1.362
1024	128	1e-2	27.8	1.361
1024	256	1e-3	74.6	0.550
1024	512	1e-3	71.2	0.620
1024	1024	1e-3	68.2	0.771
1024	2048	1e-3	64.2	0.814
2048	512	1e-3	65.1	0.760
4096	512	1e-3	62.8	0.814
8192	512	1e-3	67.4	0.713
12288	512	1e-3	48.2	1.060
16384	512	1e-3	34.9	1.290

TABLE XI: ResNet Hyperparameter Exploration

Slice Length	Batch Size	Learning Rate	Acc. (%)	Loss
512	512	1e-3	45.6	1.094
1024	512	1e-3	52.3	0.997
2048	512	1e-3	39.3	1.254
4096	512	1e-3	44.6	1.152
8192	4	1e-3	55.8	0.863
8192	8	1e-3	57.2	0.850
8192	16	1e-3	57.6	0.837
8192	32	1e-3	57.6	0.839
8192	64	1e-3	57.8	0.835
8192	128	1e-4	60.8	0.828
8192	128	2e-4	60.8	0.815
8192	128	5e-4	59.8	0.832
8192	128	1e-3	58.1	0.835
8192	128	5e-3	29.8	1.345
8192	128	1e-2	29.8	1.345
8192	256	1e-3	57.6	0.864
8192	512	1e-3	55.0	0.891
8192	1024	1e-3	57.0	0.942
8192	2048	1e-3	54.5	1.057
12288	512	1e-3	33.4	1.287
16384	512	1e-3	39.4	1.305

TABLE XII: 1D CNN Hyperparameter Exploration

Slice Length	Batch Size	Learning Rate	Acc. (%)	Loss
1024	16	1e-3	52.6	0.964
1024	32	1e-3	51.7	0.968
1024	64	1e-3	51.6	0.970
1024	128	1e-3	51.4	0.974
1024	256	1e-3	51.9	0.959
2048	16	1e-3	54.2	0.903
2048	32	1e-3	53.5	0.903
2048	64	1e-4	54.3	0.909
2048	64	2e-4	55.2	0.889
2048	64	5e-4	54.7	0.899
2048	64	1e-3	54.9	0.889
2048	64	5e-3	28.3	1.358
2048	64	1e-2	28.3	1.358
4096	16	1e-3	53.7	0.895

TABLE XIII: MCformer Hyperparameter Exploration

Models	Slice Length		Batch Size		Learning Rate	
	Default	Fine-tuned	Default	Fine-tuned	Default	Fine-tuned
AMCNet	128	8192	512	512	1e-3	2e-4
ResNet	1024	1024	512	128	1e-3	1e-3
1D CNN	512	8192	512	128	1e-3	2e-4
MCformer	128	2048	512	64	1e-3	2e-4

TABLE XIV: Default and Fine-tuned Hyperparameters of Baseline Models

Models	Training				Validation			
	Default		Fine-tuned		Default		Fine-tuned	
	Acc. (%)	Loss	Acc. (%)	Loss	Acc. (%)	Loss	Acc. (%)	Loss
T-PRIME LG	92.4	0.213	96.4	0.179	93.5	0.166	97.5	0.068
T-PRIME SM	80.6	0.448	84.0	0.405	81.4	0.421	85.3	0.334
AMCNet	68.6	0.653	93.4	0.155	68.8	0.672	93.5	0.161
ResNet	64.7	0.667	79.9	0.403	66.3	0.770	80.3	0.439
1D CNN	53.3	0.886	60.7	0.700	45.6	1.094	61.8	0.757
MCformer	48.7	1.034	56.6	0.771	50.1	0.992	57.1	0.838

TABLE XV: Overall Training and Validation Results of T-PRIME Models, and Default and Fine-tuned Baseline Models

Models	Input Size (M × S)		# Parameters		Runtime	
	Default	Fine-tuned	Default	Fine-tuned	Default	Fine-tuned
T-PRIME LG	64 × 128	64 × 128	6.8M	6.8M	0d 0h 28m	0d 3h 43m
T-PRIME SM	24 × 64	24 × 64	1.6M	1.6M	0d 2h 26m	0d 19h 2m
AMCNet	1 × 128	1 × 8192	462K	269M	3d 20h 37m	0d 12h 34m
ResNet	1 × 1024	1 × 1024	162K	162K	0d 7h 27m	2d 2h 48m
1D CNN	1 × 512	1 × 8192	4.1M	67M	0d 7h 25m	0d 5h 37m
MCFormer	1 × 128	1 × 2048	78K	72K	2d 9h 13m	1d 11h 3m

TABLE XVI: Comparisons between T-PRIME Models, and Default and Fine-tuned Baseline Models

B. Test Results on Channel Models

In this section, T-PRIME models and optimized baseline models are compared on each individual channel model and all channel models over different values of SNR. Fig. 12 illustrates that even after the hyperparameter exploration, proposed T-PRIME LG model with fewer parameters in comparison, still consistently outperforms other optimized baseline models, achieving 96.7% accuracy at SNR = -15 dB. Despite being outperformed by AMCNet and 1D CNN on several SNR conditions, T-PRIME SM still performs well with very few parameters compared to huge amount of parameters of these baseline models.

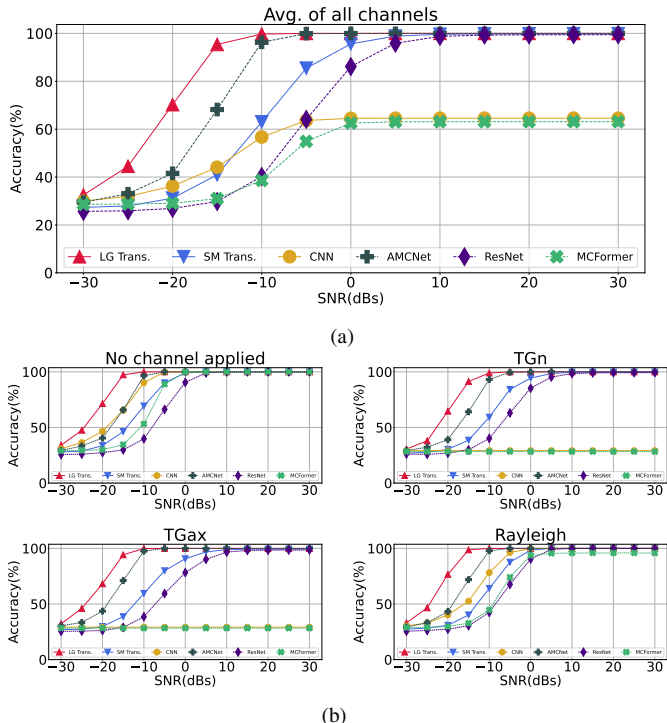


Fig. 12: (a) Comparison between SM, LG Transformer-based architectures and fine tuned baseline models tested on all channel models and different SNR conditions. The LG Transformer achieves the best overall accuracy and it retains high accuracy (i.e. > 96.7%) for SNR as low as -15 dB. (b) Performance comparison for each individual channel model tested during simulation.

C. Detailed accuracy metrics for multi-class classification on overlapping datasets

In this section, we present a detailed analysis of the classification accuracy of T-PRIME LG across various overlapping

transmission configurations outlined in Table IV. We introduce three distinct metrics for these results, one of which is innovative compared to the metrics presented in Table VI.

We emphasize two specific accuracy metrics: Exact Accuracy and Single Accuracy. Exact Accuracy pertains to precisely identifying all and only the protocols being transmitted, while single accuracy denotes the percentage of transmissions where we successfully detect at least one of the transmitted protocols, regardless of any incorrect detection of protocols not present in the signal.

The new metric, referred to as Single Exact Accuracy, quantifies instances in which we detect at least one of the transmitted protocols while simultaneously avoiding the detection of any non-transmitted protocols. Specifically, this metric recognizes as correct only those samples containing either one or both of the transmitted protocols, counting a sample as incorrect if classifier outputs any other protocol combination.

From Table XVII we observe an average of 86.88% correct detection rates for protocol classes that are actually present within the overlapping signals in the test set, while detection rate for protocols not present is on average 7.16%. This demonstrates how our classifier can accurately identify the protocols even in presence of active interference, and it is able to correctly determine both incumbent and interferer transmissions in the majority of cases.

Upon a more detailed analysis of the results, looking at Tables XVIII, XIX, XX and XXI it becomes evident that 802.11b is the protocol most easily identified by the model. This behavior may arise from a number of factors, including distinctiveness of its waveforms compared to others and over-sampling applied to signals in this category in order to match other protocols' configurations in the rest of the dataset.

Another noteworthy observation is the distinct behavior of the model across the two datasets. In dataset RM_C_02, we generally observe superior detection accuracy for the incumbent signal compared to the interferer. However, in dataset RM_C_01, the model deviates from this pattern, exhibiting numerous instances where higher accuracy is achieved for the interferer.

Finally, it is noteworthy to observe the model's difficulty in identifying the 802.11g protocol particularly when mixed as an interferer with 802.11b. In this specific configuration, the former protocol presents the most significant challenge for the model in terms of identification. Nevertheless, the model overall continues to yield satisfactory results, motivating further research efforts in these overlapping scenarios.

	Overlapping configuration (incumbent - interferer)					
Metrics (%)	b - ax	b - g	b - n	n - ax	g - ax	g - n
Exact Accuracy	69.9	63.8	69.4	62.5	76.9	79.5
Single Exact Accuracy	80.2	76.5	79.4	70.7	82.2	82.6
Single Accuracy	99.8	100.0	99.9	99.9	100.0	100.0
ax detected	80.8	6.4	9.2	92.9	90.6	16.5
b detected	92.9	100.0	96.4	7.0	2.7	0.8
n detected	13.8	17.3	77.7	75.1	15.2	84.5
g detected	6.2	65.7	11.6	22.3	88.9	97.1
noise detected	0.0	0.0	0.0	0.0	0.0	0.0

TABLE XVII: Global results

	Overlapping configuration (incumbent - interferer)					
Metrics (%)	b - ax	b - g	b - n	n - ax	g - ax	g - n
Exact Accuracy	72.6	97.9	90.9	38.9	84.8	89.1
Single Exact Accuracy	79.6	98.8	94.3	53.1	90.9	94.3
Single Accuracy	99.8	100.0	100.0	99.8	99.9	100.0
ax detected	92.4	0.2	1.5	97.8	99.8	4.0
b detected	84.7	100.0	95.4	19.7	7.6	1.7
n detected	16.5	1.0	96.2	45.9	1.5	98.2
g detected	4.0	98.0	4.2	27.2	86.2	92.7
noise detected	0.0	0.0	0.0	0.1	0.0	0.0

TABLE XVIII: Dataset RM_C_O1 (overlap ratio $R = 25\%$)

	Overlapping configuration (incumbent - interferer)					
Metrics (%)	b - ax	b - g	b - n	n - ax	g - ax	g - n
Exact Accuracy	82.8	58.8	70.5	40.3	69.6	96.3
Single Exact Accuracy	87.5	64.0	75.6	51.8	76.3	97.3
Single Accuracy	99.7	100.0	99.9	100.0	100.0	99.9
ax detected	93.1	1.4	5.1	79.8	95.9	1.6
b detected	94.0	100.0	93.3	2.3	1.2	1.1
n detected	12.3	34.7	85.9	74.2	22.5	98.8
g detected	0.3	62.6	19.3	45.9	78.4	98.4
noise detected	0.0	0.0	0.0	0.0	0.0	0.0

TABLE XIX: Dataset RM_C_O1 (overlap ratio $R = 50\%$)

	Overlapping configuration (incumbent - interferer)					
Metrics (%)	b - ax	b - g	b - n	n - ax	g - ax	g - n
Exact Accuracy	49.5	33.4	28.1	97.6	71.4	55.0
Single Exact Accuracy	70.7	58.3	57.1	98.6	76.8	58.7
Single Accuracy	100.0	100.0	99.8	99.9	100.0	100.0
ax detected	53.3	20.8	26.7	99.1	72.6	41.3
b detected	99.7	100.0	99.7	1.2	0.0	0.0
n detected	10.4	21.8	31.3	98.5	23.2	57.7
g detected	19.2	35.0	17.2	0.3	99.7	99.5
noise detected	0.0	0.0	0.0	0.0	0.0	0.0

TABLE XX: Dataset RM_C_O2 (overlap ratio $R = 25\%$)

	Overlapping configuration (incumbent - interferer)					
Metrics (%)	b - ax	b - g	b - n	n - ax	g - ax	g - n
Exact Accuracy	64.6	49.4	75.6	97.6	81.5	63.2
Single Exact Accuracy	79.5	79.4	84.5	99.1	83.2	65.9
Single Accuracy	99.8	100.0	100.0	100.0	99.9	100.0
ax detected	71.4	9.3	10.1	99.5	86.4	34.1
b detected	96.9	100.0	99.5	0.4	0.0	0.0
n detected	15.2	11.3	83.0	98.2	16.8	68.3
g detected	5.6	51.4	5.3	0.4	98.7	99.7
noise detected	0.0	0.0	0.0	0.0	0.0	0.0

TABLE XXI: Dataset RM_C_O2 (overlap ratio $R = 50\%$)