# Binary Transformer Detectors for Automatic Modulation Detection Under Realistic Radio Frequency Impairments

AnuraagChandra Singh Thakur * and Masudul Imtiaz

Posted Date: 4 March 2026

doi: 10.20944/preprints202603.0346.v1

Keywords: automatic modulation detection; transformer neural networks; binary signal detection; radio frequency machine learning; transfer learning; spectrum monitoring

 doi:10.20944/preprints202603.0346.v1



*Article*

# Binary Transformer Detectors for Automatic Modulation Detection Under Realistic Radio Frequency Impairments

**AnuraagChandra Singh Thakur * and Masudul Imtiaz**

Department of Electrical and Computer Engineering, Clarkson University, Potsdam, NY 13699, USA
* Correspondence: anuthak@clarkson.edu

**Abstract**

Automatic modulation classification (AMC) is a core capability for spectrum monitoring, adaptive receivers, and electronic support. Most radio-frequency machine learning (RFML) studies train multi-class classifiers on benchmark datasets that contain a single modulation per recording at baseband. In operational settings, however, the objective is often to detect only a small set of signals of interest, making large multi-class models unnecessarily expensive to train and deploy. This paper investigates an alternative workflow based on targeted *binary* transformer detectors and evaluates their robustness under practical RF complications. Using the RadioML 2018.01A dataset, we construct binary detection tasks with BPSK as the signal of interest and introduce three increasingly realistic conditions: (i) center-frequency shifts away from baseband, (ii) sampling-rate mismatches via decimation and interpolation, and (iii) multi-signal mixtures where modulations co-occur either in frequency (simultaneous transmissions) or in time (temporal concatenation). The results show that baseband-trained detectors do not generalize to center-frequency-shifted signals, and multi-signal interference can cause complete detection failure unless explicitly modeled during training. We investigate early-exit transformer inference to reduce computation on high-confidence examples, showing it maintains (and occasionally improves) detection performance. We also evaluate inter-modulation transfer learning and intra-modulation adaptation from baseband to mixed- and multi-signal scenarios.

**Keywords:** automatic modulation detection; transformer neural networks; binary signal detection; radio frequency machine learning; transfer learning; spectrum monitoring

## 1. Introduction

Automatic signal detection and modulation recognition underpin many commercial, scientific, and defense applications, ranging from adaptive and cognitive radios to spectrum monitoring and electronic support [1,2]. In commercial wireless systems, modulation awareness can support adaptive links that change modulation and coding in response to channel conditions, spectrum access policies, and multi-user interference [3]. In public safety, industrial IoT, and satellite systems, fast recognition of the waveform in an occupied channel can provide situational awareness and facilitate coexistence in congested spectrum [4,5]. In defense and spectrum-surveillance settings, rapid detection of specific emitters enables downstream tasks such as interception, localization, jamming, and prioritization of operator attention [6]. These tasks are often time-critical in dynamic environments where an adversary may rapidly change waveform parameters or hop frequencies.

Classical automatic modulation classification (AMC) relies on expert-designed features extracted from the waveform or its statistics (e.g., higher-order moments and cyclic/cyclostationary features), followed by decision rules or conventional classifiers [7]. While many recent RFML models operate directly on raw IQ samples [8], alternative approaches rely on extracted representations such as constellation or eye diagrams [9,10].

 doi:10.20944/preprints202603.0346.v1



Convolutional and recurrent architectures are widely used in this setting [11–19]. Ensemble learning approaches combine multiple weak or moderately performing learners to form a stronger composite network. In RF modulation recognition, such ensembles can improve robustness to noise and channel impairments by aggregating complementary decision boundaries learned by individual models[20,21]. Transformer models have emerged as a strong alternative for modulation analysis [22,23]. Transformers are attractive because self-attention can represent global relationships across an observation window and can emphasize salient waveform regions without being restricted to a local receptive field [24]. This property is especially relevant for digitally modulated signals whose discriminative structure can span multiple symbol intervals.

Despite strong benchmark results, two gaps commonly arise when transitioning from benchmark-driven RFML to operational deployment. The first gap is *data realism*. Public benchmarks (including the RadioML family) typically provide short, single-modulation recordings centered at baseband and sampled at a nominal rate with labels provided per recording [25,26]. Field recordings often violate these assumptions. Center-frequency offsets and Doppler introduce linear phase ramps; sampling-rate mismatch (and any subsequent resampling) distorts time-domain structure [27]; and wideband sensing frequently yields mixtures where multiple transmissions co-occur within the same window [2]. Such co-occurrence may be spectral (simultaneous transmissions in adjacent or overlapping bands) or temporal (bursts of different modulations concatenated within a fixed-length capture). Models trained exclusively on idealized baseband single-signal examples can therefore fail unexpectedly when confronted with realistic RF complications.

The second gap is *task mismatch*. Many RFML studies train a single multi-class model over all available modulations because it is convenient for benchmarking. In practical monitoring systems, however, an operator is often interested in detecting only a small set of *signals of interest* (SOIs) within a particular band or mission context. Training and deploying a large multi-class model can be unnecessarily expensive, and it can increase false-alarm risk due to label confusion among non-essential classes. A practical alternative is a bank of lightweight binary detectors, one per SOI, each solving a simpler decision problem (SOI versus background) [28]. Such detectors can be tailored to the modulations that plausibly co-occur in the target band, can run in parallel across bands and hardware resources, and can be extended incrementally by adding new detectors without retraining an entire multi-class system.

This paper investigates this targeted workflow using transformer-based *binary* detectors and explicitly evaluates robustness under RF impairments that challenge common AMC assumptions. Using the RadioML 2018.01A dataset , we formulate binary detection tasks with BPSK as the SOI and introduce three increasingly realistic conditions: (i) center-frequency shifts away from baseband via frequency mixing, (ii) sampling-rate mismatch through decimation and interpolation, and (iii) multi-signal mixtures in which modulations co-occur either in frequency (simultaneous transmissions) or in time (temporal concatenation) [29,30]. We also study early-exit transformer inference [31] to reduce computation on high-confidence examples and assess transfer learning between related modulations as a potential pathway for scaling detector training [32]. Beyond aggregate confusion-matrix metrics, we analyze false alarms and missed detections as functions of SNR to identify the operating regimes that dominate the error budget.

The main contributions of this study are: a) A binary transformer detection formulation for AMC that targets a specific SOI (BPSK) against a background of other modulation types, enabling a modular "bank of detectors" deployment; b) A controlled robustness study under practical RF complications: frequency mixing (center-frequency shifts), sampling-rate changes (decimation/interpolation), and multi-signal mixtures (frequency- and time-domain co-occurrence); c) A time-domain perspective on why self-attention is a natural fit for IQ-based detection, including intuition for frequency-shifted and multi-signal settings; d) An early-exit transformer inference study, showing that high-confidence (high-SNR) samples can exit early to reduce computation while preserving detection performance; e) An evaluation of transfer learning across related modulation schemes, examining the limitations

 doi:10.20944/preprints202603.0346.v1



inherent in a detector-style formulation and extending the framework to intra-modulation tasks, such as adapting a baseband model from mixed-signal scenarios to multi-signal environments [29,32,33].

## 2. Motivation and Problem Formulation

In a typical wideband RF processing chain, captured IQ samples are often analyzed using time-frequency methods such as the short-time Fourier transform (STFT). The frequency resolution of an $N_{\mathrm{FFT}}$-point FFT is $\Delta f = f_s / N_{\mathrm{FFT}}$, where $f_s$ is the sampling rate. Increasing $N_{\mathrm{FFT}}$ improves frequency resolution but increases computational cost, which is critical for real-time monitoring.

Depending on the bandwidth per frequency bin and the spectral occupancy, a bin (or a small group of adjacent bins) may contain multiple bursts with different modulations. In some bands, regulations or operational knowledge may indicate that only a small set of modulations is plausible. In such settings, it can be more efficient to deploy a bank of lightweight binary detectors (one per SOI) instead of a single large multi-class model. Each detector focuses on discriminating the SOI from background signals likely to co-occur in that band.

In practice, wideband monitoring commonly combines signal processing and ML: an initial detector (e.g., energy detection [36] or image/object detection on spectrograms [37–39]) localizes candidate emissions; the isolated emission is then passed to a modulation classifier. The experiments in this paper focus on the classifier stage and study how realistic RF complications affect a transformer-based binary detector.

## 3. Mathematical Justification for Attention on Time-Domain IQ Signals

Before employing self-attention mechanisms for RF signal detection and modulation analysis, we first consider three increasingly realistic scenarios: (i) a single baseband signal, (ii) a single signal with a center-frequency offset, and (iii) the superposition of multiple signals at distinct center frequencies.

### *3.1. Baseband Single-Signal Case*

Let the discrete-time complex baseband signal be

$$x[n] = I[n] + jQ[n] = a[n]e^{j\phi[n]}, \quad n = 0, 1, \ldots, T-1, \tag{1}$$

where $a[n]$ denotes the amplitude envelope and $\phi[n]$ denotes the instantaneous phase. For digitally modulated signals (e.g., PSK or QAM), the modulation structure is encoded in the temporal evolution of $\phi[n]$ and $a[n]$ across multiple symbol intervals.

A transformer-based detector operates on a sequence of embedded time samples

$$\mathbf{x}_n = g(I[n], Q[n]) \in \mathbb{R}^d, \tag{2}$$

where $g(\cdot)$ is a learned linear or nonlinear embedding.

Self-attention computes pairwise similarities between all time indices:

$$\alpha_{n,m} = \frac{\exp\Big(\langle \mathbf{q}_n, \mathbf{k}_m \rangle / \sqrt{d}\Big)}{\sum_{\ell=1}^{T} \exp\Big(\langle \mathbf{q}_n, \mathbf{k}_\ell \rangle / \sqrt{d}\Big)}, \tag{3}$$

where

$$\mathbf{q}_n = \mathbf{x}_n W_Q, \quad \mathbf{k}_m = \mathbf{x}_m W_K. \tag{4}$$

The resulting representation at time $n$ is

$$\mathbf{z}_n = \sum_{m=1}^{T} \alpha_{n,m} \mathbf{v}_m, \quad \mathbf{v}_m = \mathbf{x}_m W_V. \tag{5}$$

 doi:10.20944/preprints202603.0346.v1



This formulation allows the model to emphasize time indices that exhibit coherent phase and amplitude relationships. Unlike fixed local filters, self-attention captures long-range dependencies across symbol boundaries, which is critical for identifying modulation-specific temporal patterns in baseband signals.

### 3.2. Non-Baseband (Frequency-Shifted) Single-Signal Case

Next consider a signal that is not centered at baseband. The observed signal is

$$y[n] = x[n]e^{j2\pi\Delta f n/f_s}, \tag{6}$$

where $\Delta f$ is an unknown center-frequency offset and $f_s$ is the sampling rate.

The instantaneous phase of $y[n]$ is

$$\arg(y[n]) = \phi[n] + 2\pi\Delta f n/f_s. \tag{7}$$

Local operators (e.g., convolutions) are sensitive to this additive linear phase term, causing learned features at baseband to fail under frequency shifts. In contrast, self-attention compares pairs of time samples $(n, m)$ through inner products of learned projections:

$$\langle \mathbf{q}_n, \mathbf{k}_m \rangle \quad \propto \quad \cos(\phi[n] - \phi[m] + 2\pi\Delta f(n-m)/f_s). \tag{8}$$

The cosine term arises naturally because attention computes real inner products between complex baseband phasors, making similarity proportional to phase coherence under relative time shifts and frequency offsets. The key observation is that the frequency offset enters only through the *relative time difference* $(n - m)$ rather than the absolute time index. As a result, attention can learn to emphasize consistent phase-difference patterns across time, making it inherently more robust to unknown center-frequency offsets when trained on appropriately shifted data.

Thus, self-attention enables the model to learn invariances to frequency translation while operating purely on time-domain IQ samples.

### 3.3. Multi-Signal Superposition at Distinct Center Frequencies

Finally, consider a realistic scenario in which two signals coexist within the same observation window:

$$y[n] = x_1[n]e^{j2\pi f_1 n/f_s} + x_2[n]e^{j2\pi f_2 n/f_s}, \tag{9}$$

where $f_1 \neq f_2$.

Each time-domain sample embedding can be expressed as a superposition

$$\mathbf{x}_n = \mathbf{x}_n^{(1)} + \mathbf{x}_n^{(2)}. \tag{10}$$

The attention similarity between time indices $n$ and $m$ expands as

$$\langle \mathbf{q}_n, \mathbf{k}_m \rangle = \sum_{i=1}^{2} \langle \mathbf{q}_n^{(i)}, \mathbf{k}_m^{(i)} \rangle + \sum_{i \neq j} \langle \mathbf{q}_n^{(i)}, \mathbf{k}_m^{(j)} \rangle. \tag{11}$$

The cross-terms $(i \neq j)$ contain oscillatory factors of the form

$$e^{j2\pi(f_i - f_j)(n-m)/f_s}, \tag{12}$$

which decorrelate rapidly when $f_i \neq f_j$. Consequently, these terms contribute minimally after normalization by the softmax operation.

 doi:10.20944/preprints202603.0346.v1



As a result, attention weights decompose approximately into signal-specific components:

$$\alpha_{n,m} \approx \alpha^{(1)}_{n,m} + \alpha^{(2)}_{n,m}. \tag{13}$$

Multi-head attention further reinforces this behavior by allowing different heads to specialize in distinct phase-coherent structures. Each head effectively acts as a learned, soft matched filter that selectively attends to one signal while suppressing interference from others, all without explicit frequency-domain separation.

In summary, self-attention enables effective RF signal analysis on time-domain IQ data by modeling long-range phase and amplitude dependencies, adapting to unknown frequency offsets through relative-time correlations, and softly separating overlapping signals in realistic, non-isolated RF environments.

## 4. Study Data Preparation

### *4.1. RadioML 2018.01A Dataset*

RadioML 2018.01A contains 24 modulation classes (details provided in Table 1) with 106,496 samples per class, totaling 2,555,904 IQ recordings ($24 \times 106{,}496$) [30]. Each recording contains $T = 1{,}024$ complex samples represented by two real-valued streams (in-phase and quadrature). The dataset provides labeled SNR values ranging from $-20$ dB to 30 dB in 2 dB increments.

**Table 1.** Original class distribution versus binary prediction percentages. Red entries correspond to highlighted values in the plots. All values are reported as percentages relative to the original class size.

| Class | Early Exit (BB→Mixed) | | 2-Layer (BB→Mixed) | | Early Exit (Mixed→Multi) | | 2-Layer (Mixed→Multi) | |
|---|---|---|---|---|---|---|---|---|
| | P0 (%) | P1 (%) | P0 (%) | P1 (%) | P0 (%) | P1 (%) | P0 (%) | P1 (%) |
| 0 (OOK) | 39.38 | 60.62 | 100.00 | 0.00 | 45.42 | 54.58 | 99.73 | 0.27 |
| 1 (4ASK) | 58.35 | 41.65 | 100.00 | 0.00 | 64.23 | 35.77 | 98.04 | 1.96 |
| 2 (8ASK) | 58.81 | 41.19 | 100.00 | 0.00 | 74.31 | 25.69 | 98.85 | 1.15 |
| 3 (BPSK) | 15.69 | 84.31 | 55.38 | 44.62 | 48.31 | 51.69 | 45.77 | 54.23 |
| 4 (QPSK) | 51.77 | 48.23 | 99.00 | 1.00 | 99.88 | 0.12 | 99.81 | 0.19 |
| 5 (8PSK) | 51.88 | 48.12 | 100.00 | 0.00 | 99.92 | 0.08 | 99.85 | 0.15 |
| 6 (16PSK) | 51.42 | 48.58 | 100.00 | 0.00 | 99.96 | 0.04 | 99.85 | 0.15 |
| 7 (32PSK) | 52.31 | 47.69 | 99.96 | 0.04 | 99.88 | 0.12 | 99.88 | 0.12 |
| 8 (16APSK) | 34.62 | 65.38 | 99.73 | 0.27 | 99.65 | 0.35 | 99.54 | 0.46 |
| 9 (32APSK) | 8.31 | 91.69 | 97.65 | 2.35 | 98.15 | 1.85 | 95.23 | 4.77 |
| 10 (64APSK) | 12.35 | 87.65 | 99.27 | 0.73 | 99.42 | 0.58 | 99.54 | 0.46 |
| 11 (128APSK) | 8.92 | 91.08 | 99.42 | 0.58 | 99.69 | 0.31 | 99.46 | 0.54 |
| 12 (16QAM) | 13.73 | 86.27 | 99.31 | 0.69 | 99.42 | 0.58 | 99.15 | 0.85 |
| 13 (32QAM) | 14.31 | 85.69 | 98.92 | 1.08 | 99.65 | 0.35 | 99.19 | 0.81 |
| 14 (64QAM) | 9.58 | 90.42 | 99.19 | 0.81 | 99.54 | 0.46 | 98.19 | 1.81 |
| 15 (128QAM) | 11.50 | 88.50 | 99.15 | 0.85 | 99.38 | 0.62 | 99.15 | 0.85 |
| 16 (256QAM) | 9.27 | 90.73 | 99.23 | 0.77 | 99.46 | 0.54 | 98.96 | 1.04 |
| 17 (AM-SSB-WC) | 61.08 | 38.92 | 100.00 | 0.00 | 94.42 | 5.58 | 98.00 | 2.00 |
| 18 (AM-SSB-SC) | 61.15 | 38.85 | 100.00 | 0.00 | 94.08 | 5.92 | 98.65 | 1.35 |
| 19 (AM-DSB-WC) | 19.00 | 81.00 | 99.19 | 0.81 | 99.27 | 0.73 | 99.92 | 0.08 |
| 20 (AM-DSB-SC) | 16.19 | 83.81 | 98.19 | 1.81 | 99.38 | 0.62 | 99.85 | 0.15 |
| 21 (FM) | 55.54 | 44.46 | 100.00 | 0.00 | 99.96 | 0.04 | 99.88 | 0.12 |
| 22 (GMSK) | 55.23 | 44.77 | 100.00 | 0.00 | 99.96 | 0.04 | 99.88 | 0.12 |
| 23 (OQPSK) | 38.85 | 61.15 | 97.35 | 2.65 | 99.42 | 0.58 | 98.27 | 1.73 |
| 24 (BPSK+QPSK) | – | – | – | – | 16.96 | 83.04 | 3.65 | 96.35 |

For reliable evaluation across SNR, the dataset is not split using a purely random shuffle. Instead, training/validation/test splits are constructed to maintain a balanced representation across all SNR values. An 80–10–10 split is used for training, validation, and testing, respectively. Due to dataset size and the number of experiments in this paper, approximately half of the available data is used.

For validation and testing, there are 100 samples per SNR (26 distinct SNR values from $-20$ dB to 30 dB) per class, i.e., 2,600 samples per class. For the binary setting with one positive class (BPSK) and 23 negative classes, this yields 2,600 positive samples and 59,800 negative samples per split.

For training, there are 1,000 samples per SNR (26 distinct SNR values), i.e., 26,000 samples per class. In the same binary setting, this corresponds to 26,000 positive samples and 598,000 negative samples.

 doi:10.20944/preprints202603.0346.v1



*4.2. Mixed-Frequency (Center-Frequency Shifted) Data*

To model recordings where the SOI is not centered at baseband, each IQ sample is frequency shifted by multiplying it by a complex exponential. Given a discrete-time complex signal $x[n]$, a frequency shift of $\Delta f$ (Hz) is:

$$y[n] = x[n]\, e^{-j2\pi\Delta f \frac{n}{f_s}}, \tag{14}$$

where $f_s$ is the sampling frequency. If the offset is expressed in normalized form,

$$\epsilon = \frac{\Delta f}{f_s}. \tag{15}$$

The dataset does not provide $f_s$ explicitly; therefore, the experiments use a normalized relative frequency offset parameter $\epsilon$, uniformly sampled in the range $[-200, 200]$.

*4.3. Multi-Signal Mixture Data*

Operational RF captures frequently contain multiple simultaneous transmissions within the same observation window. To emulate this complexity, we generate two-signal mixtures by adding a discriminator modulation to the anchor signal.

BPSK is selected as the anchor (SOI), and QPSK is used as the discriminator. Let $x_{\text{BPSK}}[n] = I_{\text{BPSK}}[n] + jQ_{\text{BPSK}}[n]$ denote the anchor signal and $x_{\text{QPSK}}[n] = I_{\text{QPSK}}[n] + jQ_{\text{QPSK}}[n]$ denote the inserted signal. The mixed recording is:

$$x_{\text{mix}}[n] = x_{\text{BPSK}}[n] + x_{\text{QPSK}}[n]. \tag{16}$$

Equivalently, in terms of the I/Q components:

$$I_{\text{mix}}[n] = I_{\text{BPSK}}[n] + I_{\text{QPSK}}[n], \tag{17}$$

$$Q_{\text{mix}}[n] = Q_{\text{BPSK}}[n] + Q_{\text{QPSK}}[n]. \tag{18}$$

For interpretability, mixtures are generated by combining signals at the same nominal SNR value.

*4.4. Time-Domain Contamination Data*

To isolate the effect of temporal (rather than spectral) interference, a separate experiment concatenates two modulations in time at baseband. BPSK is placed at the beginning of the 1,024-sample window and QPSK fills the remainder. The proportion of the window occupied by BPSK is swept in 10% increments to study how detection metrics change as the SOI becomes more or less visible.

## 5. Research Methodology

*5.1. Binary Detection Formulation*

Each input is a time sequence with $T = 1{,}024$ samples and two channels (I and Q). The primary task is formulated as a binary detection problem with BPSK as the positive class ($y = 1$) and all remaining modulations as negatives ($y = 0$), unless otherwise specified for multi-signal experiments.

The data is then passed into a binary class transformer detector which either produces a scalar logit output $z = f_\theta(X)$. A probability is obtained using the sigmoid function $\sigma(z)$.

*5.2. Loss Function and Optimization*

Because the dataset is strongly imbalanced (one positive modulation versus many negatives in the first two experiments, and multiple positive mixture classes in the later experiments), training uses a weighted binary cross-entropy loss with logits (`BCEWithLogitsLoss`). The loss is:

 doi:10.20944/preprints202603.0346.v1



$$\mathcal{L}(y, z) = -\Big[w_+ \, y \log(\sigma(z)) + (1 - y) \log(1 - \sigma(z))\Big], \tag{19}$$

where $w_+$ is a positive-class weighting term (`pos_weight`). Training is performed using the Adam optimizer with mini-batch gradient descent. Each batch consists of tuples $(X, y, Z)$, where $X$ is the I/Q sequence, $y$ is the binary label, and $Z$ is the associated SNR value. The SNR term is carried for analysis but is not used directly by the loss. The batch size is 32.

### *5.3. Hyperparameter Tuning with Optuna*

Transformer training is computationally expensive, making exhaustive grid search impractical. Hyperparameters are tuned using Optuna, which prioritizes promising regions of the search space based on previous trials.

**Table 2.** Transformer hyperparameter search space (Optuna).

| **Hyperparameter** | **Search Range / Choices** |
|---|---|
| $d_{\text{model}}$ | {32, 64} |
| $n_{\text{head}}$ | {4, 8} |
| # Layers | {2, 4} |
| $\text{dim}_{\text{ff}}$ | {64, 128} |
| Learning rate ($lr$) | $[10^{-4}, 10^{-3}]$ (log-uniform) |

Each Optuna trial samples a transformer configuration, performs a short training run, and evaluates performance on the validation set. Due to class imbalance, trials are ranked using the area under the precision–recall curve (PR-AUC). After selecting the best configuration, a final model is trained for 10–15 epochs.

### *5.4. Decision Threshold and Performance Metrics*

After training, the model outputs probability scores $p$ on the validation set. A hard prediction $\hat{y}$ is obtained using a decision threshold $t$:

$$\hat{y} = \begin{cases} 1 & p \geq t, \\ 0 & p < t. \end{cases} \tag{20}$$

Rather than using a fixed threshold (e.g., $t = 0.5$), $t$ is selected from the validation precision–recall curve by maximizing the F1-score. Final performance is evaluated using the confusion matrix metrics—true negatives (TN), false positives (FP), false negatives (FN), and true positives (TP). From these, specificity, sensitivity, precision, recall, F1-score, and balanced accuracy are computed and reported. In addition, FP and FN rates are analyzed as a function of SNR.

## 6. Results and Analysis

### *6.1. Experiment 1: Baseband Training and Evaluation*

The first experiment trains and tests a BPSK detector on the original baseband dataset.

Figure 1 shows that most false negatives occur at low SNR (approximately $-20$ dB to 0 dB), where the SOI is heavily masked by noise. False positives also concentrate in this low-SNR regime.

 doi:10.20944/preprints202603.0346.v1



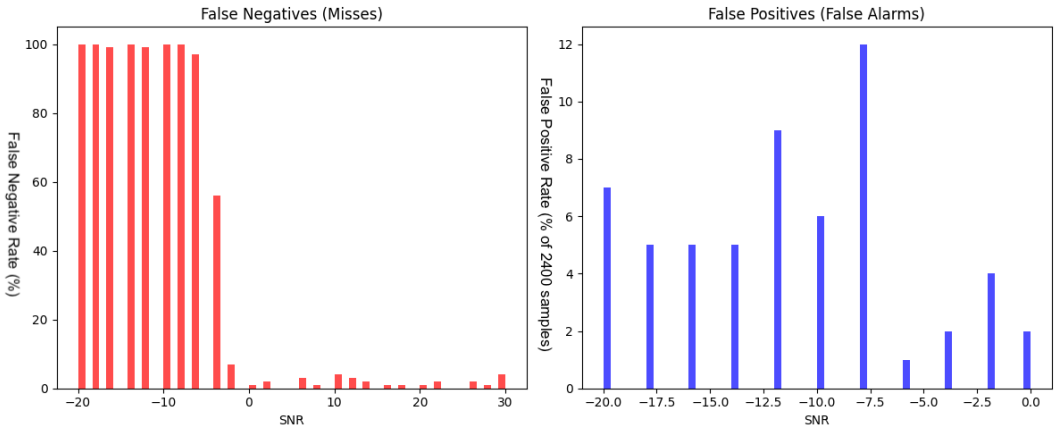


**Figure 1.** False negative rates (%) and false positive rates normalized by 2400 samples (%) as functions of SNR, where the model is trained and tested exclusively on baseband signals (Experiment 1).

#### 6.1.1. Generalization to Mixed-Frequency Data

A key question is whether a detector trained only on baseband signals generalizes to signals with a nonzero center-frequency offset. Table 3 (row "Baseband Train, Mixed Test") shows a large degradation: FN increases to 4.16% of all samples and TP collapses to 0.01%. The sensitivity drops from 65.9% to 0.27%. In other words, a baseband-trained detector is not robust to center-frequency shifts and must be retrained (or otherwise adapted) for mixed-frequency operation.

 doi:10.20944/preprints202603.0346.v1



**Table 3.** Binary confusion matrix values and derived performance metrics across training and testing scenarios. All values, including derived metrics, are reported as percentages (%). Sensitivity and Recall are equivalent but both are shown for completeness.

| Scenario | TN (%) | FP (%) | FN (%) | TP (%) | Spec. (%) | Sens. (%) | Prec. (%) | Recall (%) | F1 (%) | Bal. Acc. (%) |
|---|---|---|---|---|---|---|---|---|---|---|
| Baseband Train/Test | 95.74 | 0.09 | 1.42 | 2.75 | 99.90 | 65.92 | 96.73 | 65.92 | 78.56 | 82.91 |
| Baseband Train, Mixed Test | 95.76 | 0.08 | 4.16 | 0.01 | 99.92 | 0.27 | 12.96 | 0.27 | 0.53 | 50.10 |
| Baseband Train/Test, Upsampled (×2) | 91.41 | 4.42 | 1.57 | 2.60 | 95.39 | 62.32 | 36.98 | 62.32 | 46.18 | 78.86 |
| Baseband Train/Test, Downsampled (×1/3) | 90.28 | 5.56 | 1.59 | 2.57 | 94.19 | 61.72 | 31.65 | 61.72 | 41.62 | 77.96 |
| Baseband Train/Test, Downsampled (×1/2) | 90.63 | 5.21 | 1.71 | 2.46 | 94.57 | 58.92 | 32.06 | 58.92 | 41.39 | 76.74 |
| Baseband Train/Test, Downsampled (×1/4) | 91.25 | 4.58 | 2.11 | 2.06 | 95.22 | 49.35 | 30.98 | 49.35 | 38.05 | 72.28 |
| Baseband Train/Test, Early Exit | 95.80 | 0.03 | 1.35 | 2.82 | 99.97 | 67.61 | 98.93 | 67.61 | 80.70 | 83.79 |
| Mixed Train/Test | 95.68 | 0.16 | 2.08 | 2.09 | 99.84 | 50.15 | 93.08 | 50.15 | 64.93 | 74.99 |
| Mixed Train/Test, Upsampled (×2) | 95.17 | 0.66 | 4.12 | 0.05 | 99.31 | 1.19 | 7.01 | 1.19 | 1.99 | 50.25 |
| Mixed Train/Test, Decimated (×1/2) | 92.43 | 3.40 | 2.84 | 1.33 | 96.45 | 31.83 | 28.04 | 31.83 | 29.80 | 64.14 |
| Mixed Train/Test, Upsampled (×3) | 95.83 | 0.00 | 4.17 | 0.00 | 100.00 | 0.00 | 0.00 | 0.00 | 0.00 | 50.00 |
| Mixed Train/Test, Decimated (×1/4) | 86.47 | 9.37 | 3.61 | 0.55 | 90.22 | 13.31 | 5.59 | 13.31 | 7.92 | 51.77 |
| Mixed Train/Test, Early Exit | 95.79 | 0.05 | 1.43 | 2.74 | 99.95 | 65.71 | 98.27 | 65.71 | 78.80 | 82.83 |
| Multi-Signal (2-Class) Train/Test | 90.92 | 1.08 | 2.15 | 5.85 | 98.83 | 73.12 | 84.41 | 73.12 | 78.36 | 85.98 |
| Multi-Signal (3-Class) Train/Test | 91.18 | 1.13 | 3.41 | 4.28 | 98.78 | 55.62 | 79.15 | 55.62 | 65.33 | 77.20 |
| Transfer Learning, Early Exit (BPSK→QPSK) | 14.15 | 81.68 | 0.05 | 4.11 | 14.76 | 98.70 | 4.80 | 98.70 | 9.15 | 56.73 |
| Transfer Learning, Last Two Layers (BPSK→QPSK) | 1.62 | 23.98 | 71.85 | 2.55 | 6.31 | 3.43 | 9.62 | 3.43 | 5.19 | 4.87 |
| Transfer Learning, Early Exit (Baseband→Mixed BPSK) | 38.12 | 57.72 | 0.65 | 3.51 | 39.76 | 84.30 | 5.74 | 84.30 | 10.74 | 62.03 |
| Transfer Learning, Last Two Layers (Baseband→Mixed BPSK) | 95.00 | 0.80 | 2.33 | 1.88 | 99.17 | 44.61 | 70.20 | 44.61 | 54.83 | 71.89 |
| Transfer Learning, Early Exit (Mixed BPSK→Multi-Signal) | 86.58 | 5.42 | 2.61 | 5.39 | 94.11 | 67.37 | 49.86 | 67.37 | 57.32 | 80.74 |
| Transfer Learning, Last Two Layers (Mixed BPSK→Multi-Signal) | 91.16 | 0.84 | 1.98 | 6.02 | 99.08 | 75.27 | 87.70 | 75.27 | 81.01 | 87.18 |

### 6.1.2. Sampling-Rate Mismatch

Sampling-rate mismatch is modeled via resampling. Table 3 shows that upsampling (×2) reduces TP from 2.75% to 2.60% and increases FN from 1.42% to 1.57%. Sensitivity degrades from 65.92% at baseband to 62.32% with upsampling and to 61.72%, 58.92%, and 49.35% under downsampling by 1/3, 1/2, and 1/4. Downsampling also degrades performance, with larger penalties as the resampling ratio deviates further from 1.

 doi:10.20944/preprints202603.0346.v1



### 6.2. Experiment 2: Mixed-Frequency Training and Robustness

Because Experiment 1 indicates poor generalization to frequency-shifted signals, Experiment 2 trains and tests the detector on mixed-frequency data.

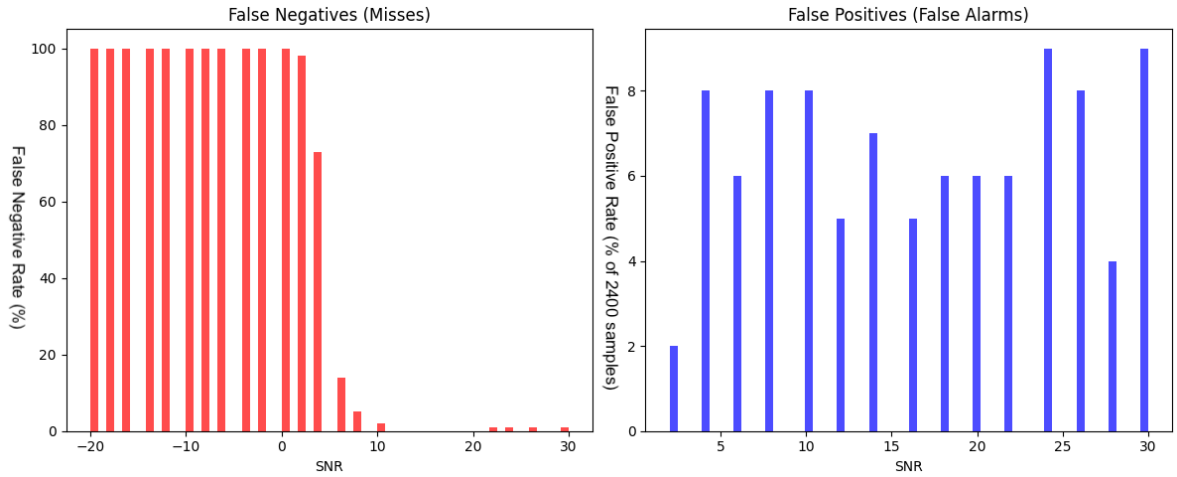


**Figure 2.** False negative rates and false positive rates normalized by 2400 samples (%) as functions of SNR, where the model is trained and tested on mixed-frequency signals (Experiment 2).

Compared to baseband training, mixed-frequency training slightly reduces the overall TP count (2.09% versus 2.75% in Table 3, which is consistent with the added complexity introduced by frequency offsets.

#### 6.2.1. Sampling-Rate Mismatch in Mixed-Frequency Data

Table 3 shows a notable asymmetry for mixed-frequency resampling: upsampling ($\times 2$) yields a TP of only 0.05%, whereas decimation ($\times 1/2$) yields 1.33%. Increasing the upsampling ratio to 3 completely reduces the true positives to 0%, making the model unusable. Although ideal interpolation should preserve information content, the observed behavior suggests that resampling may introduce artifacts that are more disruptive when the signal is not centered at baseband.

#### 6.2.2. Failure Under Two-Signal Mixtures Without Targeted Training

The mixed-frequency detector is evaluated on two-signal mixtures (BPSK+QPSK) to test robustness to co-occurring transmissions at different center frequencies. Table 1 (class 24 under the "Mixed->Multi-Signal(2)" column) shows that none of the mixture samples are predicted as containing BPSK (P1$=$ 0). This indicates complete failure to detect the SOI in the presence of a simultaneous in-band interferer unless such interference is explicitly modeled during training. This introduces a new challenge in real-life spectrum with overlapping modulation types.

### 6.3. Experiment 3: Multi-Signal Training

Experiment 3 trains a detector directly on multi-signal data. Any sample containing a BPSK component (pure BPSK or BPSK+QPSK mixture) is labeled positive, while all other signals are labeled negative. This experiment evaluates whether explicit exposure to interference can restore detection capability.

 doi:10.20944/preprints202603.0346.v1



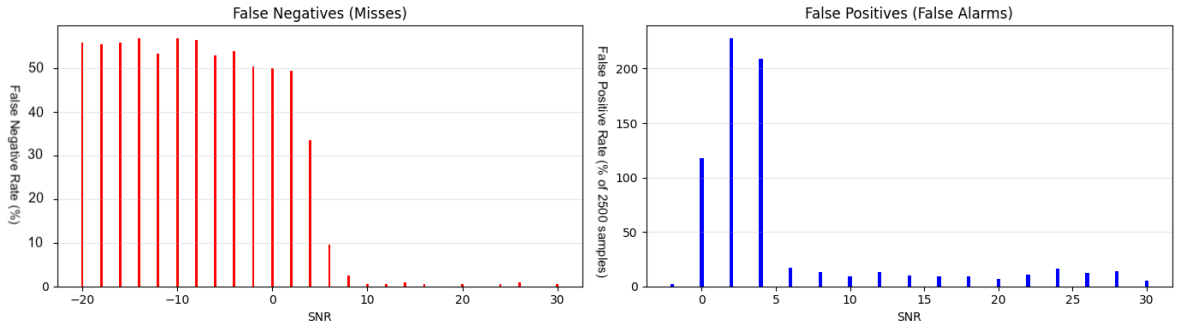


**Figure 3.** False negatives rates (left) and false positives rates normalized by 2500 samples (right) as functions of SNR, where the model is trained and tested on multi-signal scenario (BPSK+QPSK). In these graphics, an additional composite class is added to the classification set, increasing the number of classes and therefore the number of samples per SNR level used for normalization (2500 samples).

The multi-signal detector recovers detection on mixtures (Table 1, class 24 under "Multi-Signal(2)") while maintaining comparable performance on pure BPSK. However, Table 3 also shows that multi-signal training can induce false positives on certain AM sideband classes (classes 17 and 18). This is a significant observation, as it suggests that the underlying issue may not originate from the BPSK or QPSK modulation schemes themselves, but rather from another part of the system (a different modulation that the superposition resembles).

To mitigate these confusions, an extended discriminator set is introduced by including closely related modulation types in training (Table 1, classes 25 and 26 under "Multi-Signal"). This modification increases the number of true positives for the target signal from 50.04% in the two-signal case to 53.15%. It also reduces false positives for classes 17 and 18 from 8.88% and 9.08% to 0.04% and 0.81%, respectively.

### *6.4. Summary of Evaluated Scenarios*

Table 3 summarizes confusion matrix values for all evaluated scenarios, including baseband versus mixed-frequency conditions, sampling-rate mismatches, early-exit inference, multi-signal training, and transfer learning.

### *6.5. Early-Exit Inference*

Transformer inference is expensive because all encoder layers must be executed for every input. Many samples are easy to classify (especially at high SNR), motivating an early-exit strategy [31]: intermediate classifiers are attached to early layers and inference stops once confidence exceeds a threshold.

At high SNR, approximately 80% of samples exit early, substantially reducing average computation. Table 3 shows that early-exit inference preserves detection performance and, in some cases, slightly improves TN/TP (e.g., "Baseband Train/Test, Early Exit" and "Mixed Train/Test, Early Exit").

 doi:10.20944/preprints202603.0346.v1



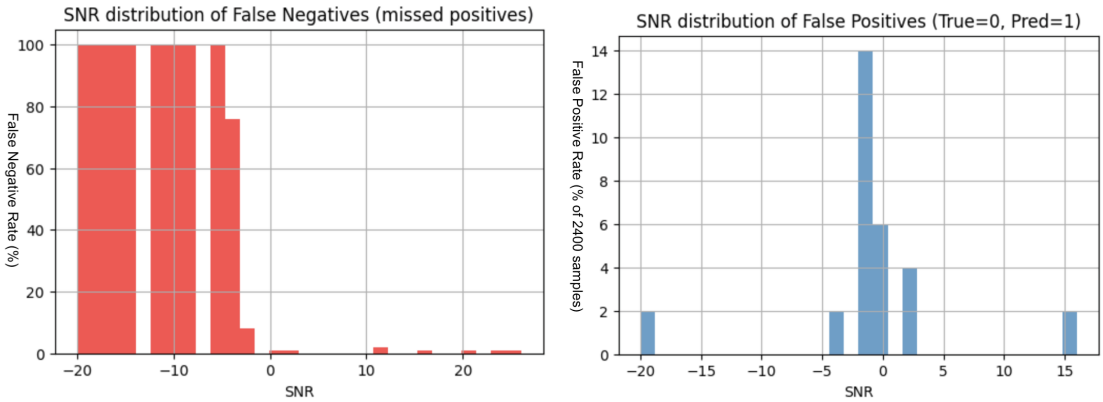


**Figure 4.** Early-exit model : false negatives (left) and false positives normalized by 2400 samples (right) as functions of SNR, where the model is trained and tested on baseband scenario.


**Figure 5.** Early-exit model: false negatives (left) and false positives normalized by 2400 samples (right) as functions of SNR, where the model is trained and tested on mixed-frequency scenario.

As with the standard transformer, most early-exit errors are concentrated at low SNR values.

### *6.6. Comparison with a 24-Class Transformer Model*

For reference, a conventional 24-class transformer model is trained to predict one of the 24 modulation classes. The same approach as earlier was used, except that the problem was formulated as a 24-class output layer instead of the earlier binary class classifier.

 doi:10.20944/preprints202603.0346.v1



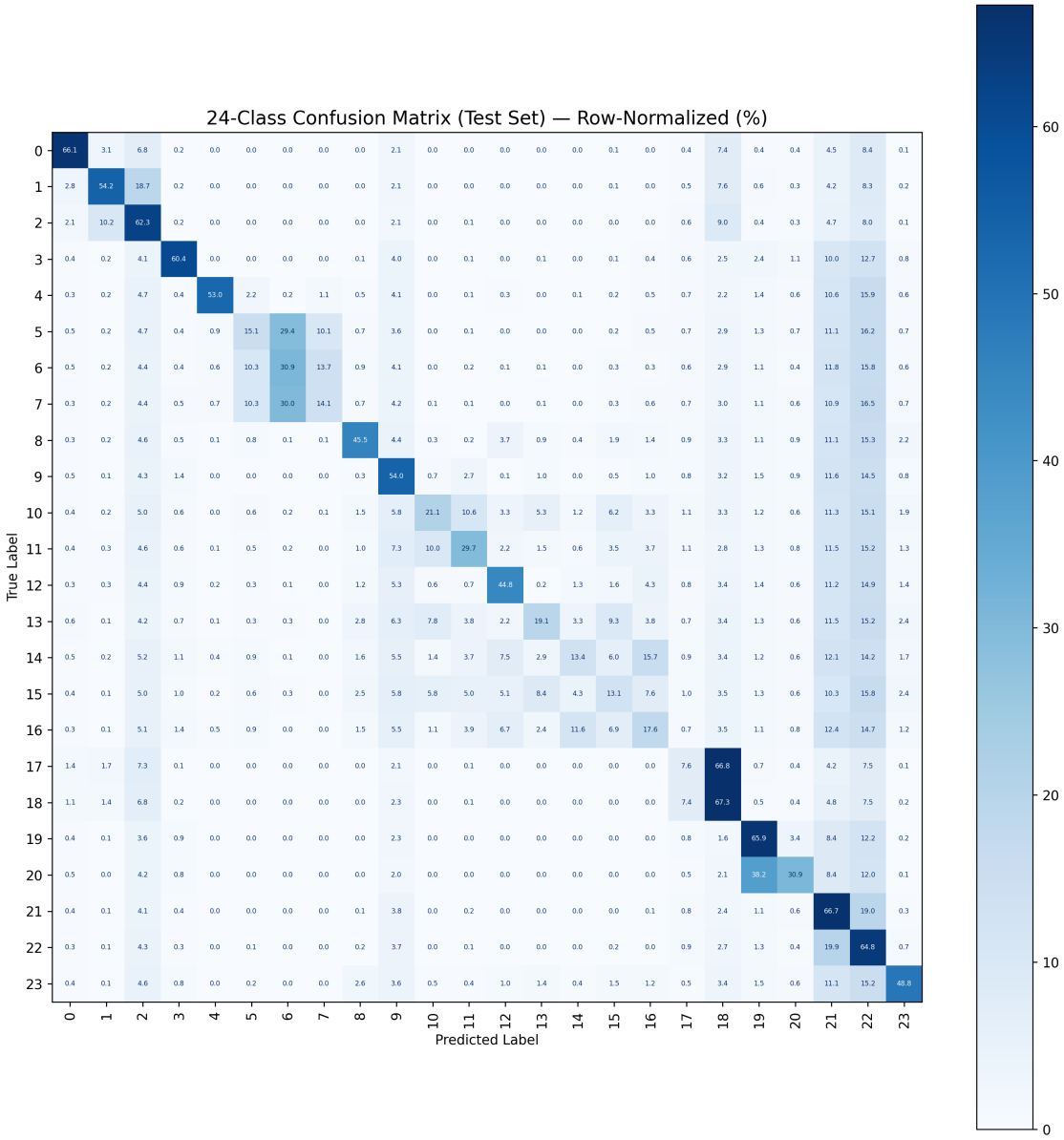


**Figure 6.** Confusion matrix of a 24-class transformer classifier trained on the RadioML 2018.01A modulations.

In this study, the binary detectors generally provide stronger SOI-versus-background discrimination than the 24-class model. A likely contributing factor is that the same hyperparameter scale is used, whereas multi-class classification typically requires higher-capacity models to learn separations across many classes.

Class 3 achieved a false positive rate of 14% and a false negative rate of 38%. For the binary classifier, the detection rate has a false positive rate of 0.09% and a false negative rate of 1.42%. Binary detectors are generally better suited for identifying features associated with a single class, rather than distinguishing across multiple classes, particularly when operating under the same architectural constraints. This result reflects a capacity-allocation effect: if model size is fixed, binary detectors can devote capacity exclusively to the SOI-Specific discrimination, whereas multi-class models must distribute capacity across all the 24 decision boundaries.

### *6.7. Time-Domain Contamination Study*

The previous multi-signal experiments considered simultaneous transmissions (overlap in time) at different center frequencies. This study instead examines temporal co-occurrence at baseband: the SOI (BPSK) occupies the beginning of the window and a second modulation (QPSK) occupies

 doi:10.20944/preprints202603.0346.v1



the remainder. By varying the fraction of the window containing BPSK, we evaluate how much uncontaminated SOI content is required for reliable detection.

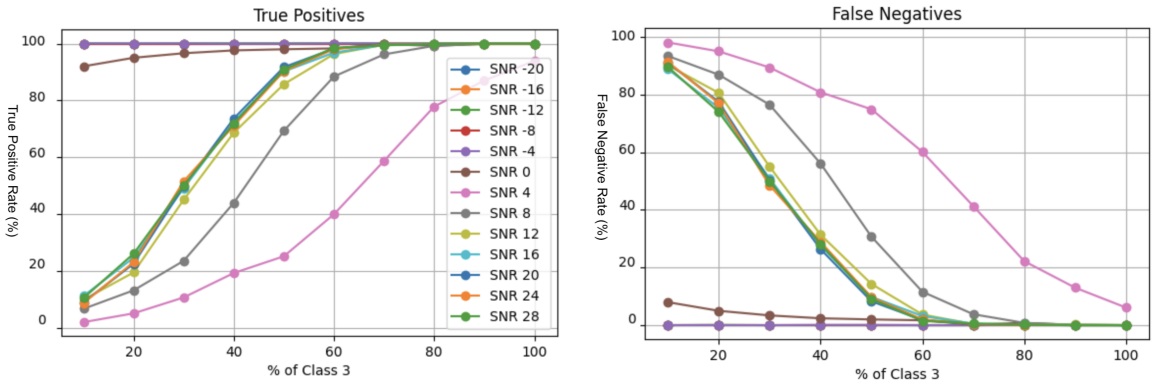


**Figure 7.** Time-domain contamination experiment: true positives (left) and false negatives (right) versus the fraction of the recording occupied by the SOI (BPSK).

As expected, robustness improves as the proportion of SOI content increases. Importantly, temporal co-occurrence has a smaller impact than spectral co-occurrence: when modulations are separated in time, the SOI segment remains uncontaminated and its phase/amplitude structure is more readily detected.

### *6.8. Transfer Learning*

Transfer learning is investigated as a means of efficiently training detectors for additional modulations. Given the structural similarity between BPSK and QPSK, a detector pretrained on BPSK might be expected to adapt to QPSK with limited retraining. If successful, this approach would reduce the need to train every modulation from scratch, enabling scalable deployment by training only a subset of representative signals and adapting them to others.

Two transfer learning strategies are evaluated: (i) partial fine-tuning of the backbone network by retraining a limited subset of layers (specifically, the final two layers) while keeping the remaining layers fixed, and (ii) retraining only the early-exit classifiers while preserving the backbone network parameters.

For the first strategy, the results indicate a substantially elevated false positive count of 81.68%, which is the highest observed across all evaluated experiments. A plausible explanation is that the model has been strongly optimized for BPSK detection; introducing QPSK as a new target while keeping early feature representations fixed may bias the network toward patterns previously associated with positive detections. Consequently, the decision boundary shifts in a manner that significantly increases false alarms across all classes. This reflects a common issue in signal classification: when two modulation schemes are superimposed, they can introduce composite features that resemble those of a third or fourth modulation type. As a result, the root cause of the misclassification may stem from signal superposition effects rather than from deficiencies in the individual modulation schemes themselves.

Partial retraining of the backbone by fine-tuning the final two layers yields a notable reduction in false positives to 23.98%. While this represents an improvement, the false positive rate remains unacceptably high relative to other evaluated configurations.

This outcome is problematic for signal detection applications, where false alarms are highly undesirable. Each false positive typically requires additional verification, increasing operational burden and reducing system efficiency. As such, the current transfer learning approaches examined here require further refinement before they can be considered viable for accelerated training of related modulations. Overall, these results highlight a key limitation: transfer learning does not generalize effectively across distinct modulation classes in this setting.

 doi:10.20944/preprints202603.0346.v1



Transfer learning may still be effective when applied within the same modulation family or task domain. To evaluate this hypothesis, the baseband BPSK detector was employed as a pretrained model and subsequently adapted to more complex signal environments. Specifically, it was fine-tuned on mixed-signal data derived from baseband conditions, as presented in Table 3 (Transfer Learning, Early Exit (Baseband→Mixed BPSK)). The early-exit retraining approach yielded 3.51% true positives but an excessive 57.72% false positives, indicating no meaningful improvement. In contrast, retraining the final two layers reduced false positives to 0.8%, while achieving 1.88% true positives and 2.33% false negatives, representing a substantial improvement in false-alarm control, as presented in Table 3 (Transfer Learning, Last Two Layers (Baseband→Mixed BPSK)).

In this experiment, the Mixed Signal Model was trained and evaluated exclusively on baseband-shifted data, without exposure to multi-signal interference, and then applied directly to a multi-signal detection scenario. Despite this training–deployment mismatch, the model achieved a 67.37% sensitivity and 49.86% precision, indicating reasonably strong baseline performance given that the model had never been exposed to multi-signal data during training. In Experiment 2, a transfer-learning strategy was employed in which the final two layers of the same model were retrained, enabling targeted adaptation to the multi-signal environment. This approach yielded 75.27% sensitivity and 87.7% precision, demonstrating a substantial improvement in detection capability and a pronounced reduction in false alarms achieved through limited fine-tuning rather than architectural modification. The results are presented in Table 3 under the scenario "Transfer Learning, Early Exit (Mixed BPSK→Multi-Signal)" and "Transfer Learning, Last Two Layers (Mixed BPSK→Multi-Signal)."

## 7. Discussion

### *7.1. Key Empirical Findings*

The experiments demonstrate that realistic RF conditions substantially alter detector behavior compared to the idealized baseband, single-signal setting typically assumed in public benchmarks. Three findings are important for practitioners.

First, *center-frequency shifts are a major failure mode* when they are not represented during training. A detector trained only on baseband recordings does not generalize to frequency-shifted signals: missed detections increase sharply and true positives collapse. This indicates that a "train at baseband, deploy anywhere" assumption is not valid for the evaluated transformer configuration, and that either explicit data augmentation (frequency mixing) or an operational downconversion stage is required.

Second, *sampling-rate mismatch consistently degrades performance*, with larger penalties as the resampling ratio deviates further from one. While moderate mismatch may be tolerable, the results suggest that resampling artifacts (especially interpolation in the mixed-frequency setting) can be disruptive. This emphasizes that practical RFML deployments should treat sampling-rate assumptions as part of the model specification: the front-end resampling strategy and the classifier should be designed and validated together.

Third, *simultaneous multi-signal mixtures can cause complete detection failure* unless interference structure is explicitly modeled during training. When a second modulation co-occurs at a different center frequency, a detector trained only on single-signal examples can fail to identify the SOI altogether. Training directly on multi-signal mixtures restores detection capability, but it can introduce new confusions (e.g., elevated false positives for certain AM sideband classes). Extending the discriminator set with "hard negative" mixture classes reduces these false alarms, illustrating that discriminator selection is central to robust binary detection.

Two additional observations refine the above points. The time-domain contamination study shows that co-occurrence in *time* is less damaging than co-occurrence in *frequency*: when the SOI occupies a sufficient fraction of the window, detection remains reliable because a contiguous, uncontaminated segment retains the phase/amplitude structure required for recognition. Early-exit inference provides a meaningful efficiency gain: high-SNR samples often reach high confidence at early layers (Figure 8),

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0346.v1



reducing average computation while preserving (and in some cases slightly improving) overall detection metrics.

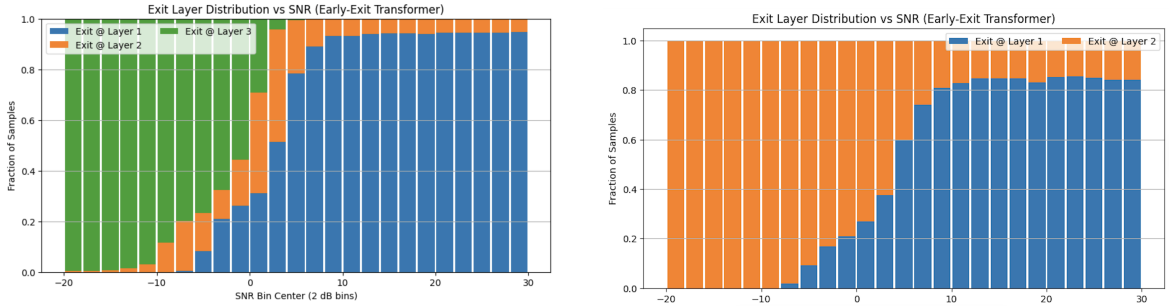


**Figure 8.** Exit behavior as a function of SNR for baseband (left) and mixed-frequency (right) data. At high SNR, a large fraction of samples exit early.

Finally, transfer learning does not generalize effectively across inter-modulation tasks, suggesting that distinct modulation types may require independent training. In contrast, transfer learning performs well for intra-modulation scenarios, such as adapting a baseband model to operate in a mixed-signal environment or extending a mixed-signal model to a multi-signal setting.

### *7.2. Implications for End-to-End Spectrum Monitoring*

These findings support a systems-oriented view in which modulation detection/classification is preceded by signal detection and isolation. In wideband spectrum monitoring, candidate signal emissions are commonly localized in the time–frequency domain using classical energy detection techniques [36] or spectrogram-based methods. Recent work has extended object-detection frameworks from computer vision to RF spectrograms, where emissions are treated as visual objects and localized using bounding-box detectors such as Faster R-CNN, YOLO, and transformer-based architectures [37–39]. In parallel, traditional image-processing and unsupervised approaches, including thresholding and clustering methods such as $k$-means, remain widely used for grouping energy concentrations into candidate signal regions. Once a contiguous region is isolated, a modulation detector can be applied on a narrower bandwidth segment where the number of co-occurring signals is reduced. Vision Transformer approaches have also been applied to RF spectrograms by treating signals as images [40].

After isolation, two practical processing paths exist. The first is to mix the isolated band to baseband and then apply a baseband-trained detector. This approach leverages the simplicity of baseband representations but depends on accurate downconversion and adequate filtering. The second path is to classify the filtered signal directly at its observed center frequency using a detector trained with frequency offsets. This avoids additional mixing and can be advantageous when downconversion is imperfect or when center-frequency alignment varies across captures.

Filtering and band isolation play an important role in both paths. In practice, filter cutoffs may not be perfectly centered around the desired emission, and nearby signals or leakage can remain in-band. FIR filters can provide linear phase (often desirable for preserving waveform structure) but may require more taps and higher computational cost; IIR filters can be more efficient but introduce nonlinear phase distortion. The experiments in this paper highlight that even modest residual interference can change detector behavior, motivating an integrated design philosophy: filtering, downconversion (if used), and detector training should be co-designed so that the detector is exposed to the residual artifacts expected at deployment.

### *7.3. Designing and Deploying Binary Detector Banks*

The binary-detector formulation is well-suited for modular deployment. Because each detector focuses on a single SOI against a background set, model capacity and training effort can be concentrated where it matters operationally. Detectors can be executed in parallel across bands and hardware

 doi:10.20944/preprints202603.0346.v1



resources, and thresholds can be tuned per SOI using validation precision–recall curves (as done in this work) rather than relying on a universal probability cutoff.

However, the results also show that binary detection is not "set and forget." The discriminator set must reflect the RF environment the detector will see. In multi-signal scenarios, including realistic hard negatives (e.g., likely co-occurring modulations and mixture classes) can substantially reduce false alarms. This mirrors classical detection practice, where the most important confusers are explicitly modeled. From an engineering standpoint, it suggests that deploying detector banks should be accompanied by a process for updating discriminators as spectrum occupancy evolves.

### *7.4. Limitations and Future Work*

Several limitations frame the scope of these conclusions. First, the study uses synthetically generated frequency shifts and mixtures derived from the RadioML dataset. While this enables controlled experiments, it does not capture the full diversity of over-the-air impairments such as hardware nonlinearities, phase noise, carrier frequency drift dynamics, nonstationary noise, and multipath fading. Second, the experiments anchor on a single SOI (BPSK) to enable a focused analysis; extending the study across additional modulations would further clarify which trends are universal versus modulation-specific. Third, the early-exit mechanism provides limited benefit in low-SNR regimes where confidence is inherently low; complementary approaches such as uncertainty estimation, calibration, or joint detection/denoising may be needed when low-SNR performance is critical.

Future work can therefore proceed in three directions: (i) evaluation on over-the-air captures and hardware-in-the-loop datasets; (ii) richer augmentation strategies that jointly model frequency offset, sampling-rate offset, fading, and interference statistics; and (iii) scalable procedures for building detector banks, including principled discriminator selection, continual learning for evolving bands, and lightweight calibration to control false alarms under distribution shift.

## 8. Conclusions

This paper reformulates AMC as targeted binary detection and evaluates transformer-based detectors under realistic RF complications. The results show that center-frequency shifts and multi-signal interference can strongly degrade performance unless explicitly modeled, whereas time-domain co-occurrence is less disruptive when the SOI remains uncontaminated. Transfer learning is effective for intra-modulation training but does not generalize well to inter-modulation tasks. Early-exit inference provides a promising pathway to reduce computation in high-SNR regimes while maintaining accuracy. Overall, the most practical deployment strategy is to combine traditional RF detection/isolation with banks of SOI-specific binary detectors.

**Author Contributions:** Please complete during submission. Suggested roles: Conceptualization, A.C.S.T. and M.I.; methodology, A.C.S.T.; software, A.C.S.T.; writing—original draft, A.C.S.T.; writing—review and editing, A.C.S.T. and M.I.; supervision, M.I.

**Funding:** This research received no external funding.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Jafarigol, E.; Alaghband, B.; Gilanpour, A.; Hosseinipoor, S.; Mirmozafari, M. AI/ML-Based Automatic Modulation Recognition: Recent Trends and Future Possibilities. *arXiv* **2025**, arXiv:2502.05315. Available online: https://arxiv.org/abs/2502.05315 (accessed on 17 February 2026).
2. Yucek, T.; Arslan, H. A Survey of Spectrum Sensing Algorithms for Cognitive Radio Applications. *IEEE Communications Surveys & Tutorials* **2009**, *11*, 116–130. https://doi.org/10.1109/SURV.2009.090109.
3. Goldsmith, A. *Wireless Communications*; Cambridge University Press: Cambridge, UK, 2005.
4. Davey, C.P.; Shakeel, I.; Deo, R.C.; Sharma, E.; Salcedo-Sanz, S.; Soar, J. End-to-End Learning of Adaptive Coded Modulation Schemes for Resilient Wireless Communications. *Applied Soft Computing* **2024**, *159*, 111672. https://doi.org/10.1016/j.asoc.2024.111672.

 doi:10.20944/preprints202603.0346.v1




5. Akyildiz, I.F.; Lee, W.-Y.; Vuran, M.C.; Mohanty, S. Next Generation/Dynamic Spectrum Access/Cognitive Radio Wireless Networks: A Survey. *Computer Networks* **2006**, *50*, 2127–2159. https://doi.org/10.1016/j.comnet.2006.05.001.
6. Zhu, M.; Li, Y.; Pan, Z.; Yang, J. Automatic Modulation Recognition of Compound Signals Using a Deep Multi-Label Classifier: A Case Study with Radar Jamming Signals. *Signal Processing* **2023**, *205*, 108874. https://doi.org/10.1016/j.sigpro.2022.108874.
7. Dobre, O.A.; Abdi, A.; Bar-Ness, Y.; Su, W. Survey of Automatic Modulation Classification Techniques: Classical Approaches and New Trends. *IET Communications* **2007**, *1*, 137–156. https://doi.org/10.1049/iet-com:20050176.
8. Abd-Elaziz, O.F.; Abdalla, M.; Elsayed, R.A. Deep Learning-Based Automatic Modulation Classification Using Robust CNN Architecture for Cognitive Radio Networks. *Sensors* **2023**, *23*, 9467. https://doi.org/10.3390/s23239467.
9. Peng, S.; Jiang, H.; Wang, H.; Alwageed, H.; Zhou, Y.; Sebdani, M.M.; Yao, Y.-D. Modulation Classification Based on Signal Constellation Diagrams and Deep Learning. *IEEE Transactions on Neural Networks and Learning Systems* **2018**, *30*, 718–727.
10. Zha, X.; Peng, H.; Qin, X.; Li, G.; Yang, S. A Deep Learning Framework for Signal Detection and Modulation Classification. *Sensors*, vol. 19, no. 18, p. 4042, Sep. 2019. doi: 10.3390/s19184042.
11. Hong, D.; Zhang, Z.; Xu, X. Automatic Modulation Classification Using Recurrent Neural Networks. In *Proceedings of the IEEE International Conference on Communications in China (ICCC)*, Chengdu, China, Dec. 2017, pp. 695–700.
12. Peng, S.; Jiang, H.; Wang, H.; Alwageed, H.; Yao, Y.-D. Modulation Classification Using Convolutional Neural Network Based Deep Learning Model. In *Proceedings of the Wireless and Optical Communications Conference (WOCC)*, 2017; pp. 1–5. https://doi.org/10.1109/WOCC.2017.7929000.
13. Wang, T.; Yang, G.; Chen, P.; Xu, Z.; Jiang, M.; Ye, Q. A Survey of Applications of Deep Learning in Radio Signal Modulation Recognition. *Applied Sciences* **2022**, *12*, 12052. https://doi.org/10.3390/app122312052.
14. Zhang, M.; Zeng, Y.; Han, Z.; Gong, Y. Automatic Modulation Recognition Using Deep Learning Architectures. In *Proceedings of the IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Kalamata, Greece, June 2018; pp. 1–5.
15. Sang, Y.; Li, L. Application of Novel Architectures for Modulation Recognition. In *Proceedings of the IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*, Chengdu, China, October 2018; pp. 159–162.
16. Daldal, N.; Yıldırım, Ö.; Polat, K. Deep Long Short-Term Memory Networks-Based Automatic Recognition of Six Different Digital Modulation Types under Varying Noise Conditions. *Neural Computing and Applications* **2019**, *31*, 1967–1981.
17. Wang, Z.; Sun, D.; Gong, K.; Wang, W.; Sun, P. A Lightweight CNN Architecture for Automatic Modulation Classification. *Electronics* **2021**, *10*, 2679.
18. Ghanem, H.S.; Al-Makhlasawy, R.M.; El-Shafai, W.; Elsabrouty, M.; Hamed, H.F.; Salama, G.M.; El-Samie, F.E.A. Wireless Modulation Classification Based on Radon Transform and Convolutional Neural Networks. *Journal of Ambient Intelligence and Humanized Computing* **2022**.
19. Du, R.; Liu, F.; Xu, J.; Gao, F.; Hu, Z.; Zhang, A. D-GF-CNN Algorithm for Modulation Recognition. *Wireless Personal Communications* **2022**, *124*, 989–1010.
20. Le, H.K.; Doan, V.S.; Hoang, V.-P. Ensemble of Convolutional Neural Networks for Improving Automatic Modulation Classification Performance. *Journal of Science and Technology* **2022**, *20*, 25–32.
21. Shi, F.; Hu, Z.; Yue, C.; Shen, Z. Combining Neural Networks for Modulation Recognition. *Digital Signal Processing* **2022**, *120*, 103264.
22. Cai, J.; Gan, F.; Cao, X.; et al. Signal Modulation Classification Based on the Transformer Network. *IEEE Transactions on Cognitive Communications and Networking* **2022**, *8*, 1348–1357.
23. Rashvand, N.; Witham, K.; Maldonado, G.; Katariya, V.; Marer Prabhu, N.; Schirner, G.; Tabkhi, H. Enhancing Automatic Modulation Recognition for IoT Applications Using Transformers. *IoT* **2024**, *5*, 212–226. https://doi.org/10.3390/iot5020011.
24. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, 2017; pp. 6000–6010.
25. O'Shea, T.J.; Roy, T.; Clancy, T.C. Over-the-Air Deep Learning Based Radio Signal Classification. *IEEE Journal of Selected Topics in Signal Processing* **2018**, *12*, 168–179. https://doi.org/10.1109/JSTSP.2018.2797022.
26. O'Shea, T.J.; Corgan, J.; Clancy, T.C. Convolutional Radio Modulation Recognition Networks. In *Engineering Applications of Neural Networks*, 2016; pp. 213–226.

 doi:10.20944/preprints202603.0346.v1




27. Proakis, J.G.; Salehi, M. *Digital Communications*, 5th ed.; McGraw-Hill: New York, NY, USA, 2008.
28. Shah, A.; Yao, Y.; Reed, J.H.; Zhang, J. Transformer-Based Spectrum Sensing for Cognitive Radio Networks. *IEEE Transactions on Cognitive Communications and Networking* **2023**, *9*, 1453–1465. https://doi.org/10.1109/TCCN.2023.3271984.
29. Jagannath, J.; Polosky, N.; O'Shea, T.; Drozd, A.; Furman, S. Transfer Learning for Wireless Spectrum Classification. In *Proceedings of the IEEE Military Communications Conference (MILCOM)*, 2019; pp. 1–6. https://doi.org/10.1109/MILCOM47813.2019.9020794.
30. Pinxau1000. RadioML 2018 Dataset. Available online: https://www.kaggle.com/datasets/pinxau1000/radioml2018 (accessed on 17 February 2026).
31. Teerapittayanon, S.; McDanel, B.; Kung, H.T. BranchyNet: Fast Inference via Early Exiting from Deep Neural Networks. In *Proceedings of the International Conference on Pattern Recognition*, 2016; pp. 2464–2469.
32. Zhang, Y.; Liu, X.; Wang, H.; Li, Z. Early-Exit Transformer Networks for Wireless Signal Recognition. In *Proceedings of the IEEE Global Communications Conference (GLOBECOM)*, 2023; pp. 1–6. https://doi.org/10.1109/GLOBECOM54140.2023.10436852.
33. Wong, L.J.; Michaels, A.J. Transfer Learning for Radio Frequency Machine Learning: A Taxonomy and Survey. *Sensors* **2022**, *22*, 1416. https://doi.org/10.3390/s22041416.
34. Li, X.; Jiang, Z.; Ting, K.; Zhu, Y. An Online Automatic Modulation Classification Scheme Based on Isolation Distributional Kernel. *arXiv* **2024**, arXiv:2410.02750. Available online: https://arxiv.org/abs/2410.02750 (accessed on 17 February 2026).
35. Zhang, F.; Luo, C.; Xu, J.; Luo, Y. An Efficient Deep Learning Model for Automatic Modulation Recognition Based on Parameter Estimation and Transformation. *IEEE Communications Letters* **2021**, *25*, 3522–3526. https://doi.org/10.1109/LCOMM.2021.3102656.
36. Urkowitz, H. Energy Detection of Unknown Deterministic Signals. *Proceedings of the IEEE* **1967**, *55*, 523–531.
37. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.
38. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*, 2015.
39. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *Proceedings of the European Conference on Computer Vision*, 2020.
40. Kim, S.; Park, J.; Kim, J. Vision Transformer for RF Signal Classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023; pp. 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096213.