Peer reviewed Journals
seron press

RESEARCH

Toward Robust Generalization in Automatic Modulation Recognition Under Channel, Parameter, and Distributional Shift Conditions

Rafael Mendes,[1] Thiago Ribeiro,[2] and Lucas Carvalho[3]

[1]Departamento de Engenharia Elétrica, Universidade Federal do Pampa, Avenida Pedro Anunciação 111, Vila Batista, Bagé, Brazil
[2]Departamento de Engenharia Elétrica e Eletrônica, Universidade Federal Rural do Semi-Árido, Avenida Francisco Mota 572, Costa e Silva, Mossoró, Brazil
[3]Departamento de Engenharia Elétrica, Universidade Estadual do Oeste do Paraná, Rua Universitária 2069, Jardim Universitário, Foz do Iguaçu, Brazil

Abstract

Background wireless receivers increasingly operate in settings where the statistical properties of signals vary across hardware platforms, propagation conditions, waveform parameterizations, and interference structures. Background constraints in these environments make automatic modulation recognition a generalization problem rather than a narrow classification exercise. A model that performs well only on a single curated dataset has limited operational value, because deployment commonly introduces carrier offsets, timing errors, pulse-shape mismatches, front-end distortion, unequal sampling assumptions, and previously unseen mixtures of nuisance factors. This paper develops a technical treatment of generalization in automatic modulation recognition by integrating statistical learning concepts with communication-theoretic signal modeling. The discussion formulates modulation recognition as supervised inference under latent environment variables and studies how inductive bias, representation geometry, optimization, and evaluation protocol shape the gap between in-domain accuracy and out-of-domain reliability. The analysis emphasizes invariance to nuisance transformations, sensitivity to hidden coupling between modulation identity and dataset construction, and the importance of environment-aware splitting strategies. It further examines how channel models, augmentation policies, transfer learning, calibration, uncertainty estimation, and deployment-time adaptation affect robustness. The paper also proposes mathematical formulations for invariant feature extraction, risk under environment shift, and representation regularization, while discussing their implications for complex baseband inputs. The central argument is that generalization in automatic modulation recognition emerges from joint design across data generation, architecture, training objective, and validation regime. Progress therefore depends less on maximizing a single benchmark score and more on constructing receivers whose internal representations remain stable when the physical realization of the same modulation class changes.

1. Introduction

Automatic modulation recognition has evolved from a classical signal processing task into a demanding representation learning problem because the modern receiver rarely observes clean textbook waveforms [1]. A practical radio front end encounters drifting oscillators, finite-length windows, residual synchronization error, device-specific nonlinearities, time-varying multipath, nonstationary interference, and data collection procedures that inadvertently leak side information into the training set. When these effects are aligned with the labels in a training corpus, a deep model can obtain impressive closed-world accuracy while learning fragile correlations that collapse outside the original sampling regime. The central scientific issue is therefore not only whether a classifier can separate modulation classes on a benchmark, but whether it can preserve that ability after the benchmark assumptions are perturbed in realistic ways.

This distinction is especially important because modulation identity is not directly observed in the raw in-phase and quadrature stream. The classifier must infer it through a finite noisy realization of a waveform generated by latent symbol sequences, pulse-shaping filters, carrier phase, channel response, oscillator mismatch, automatic gain control behavior, and hardware imperfections. The observed sample path is thus a composition of class-relevant structure and nuisance variation. A training procedure that fails to disentangle these components can exploit shortcuts that appear predictive in one dataset but are causally irrelevant to modulation type. For example, a model may overassociate a modulation class with a particular range of carrier offsets, an idiosyncratic pulse shape, or a

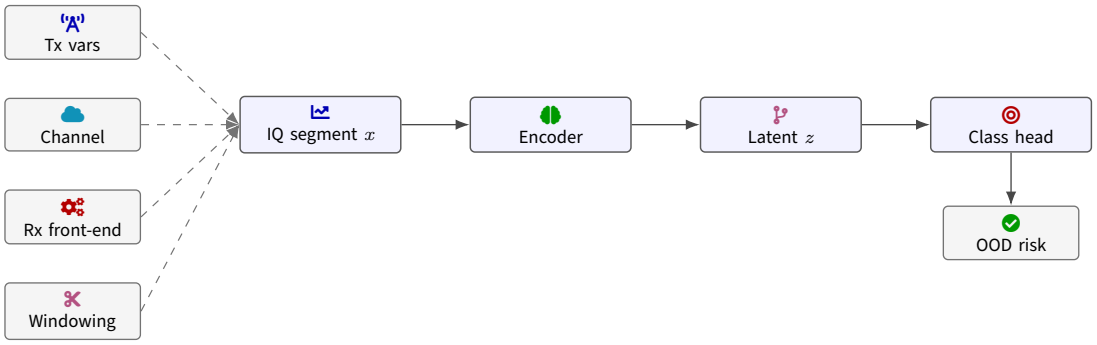


Figure 1. Environment-indexed automatic modulation recognition. A received IQ segment is shaped by transmitter, channel, receiver, and segmentation conditions before being mapped into a latent representation and classified. The figure emphasizes that robustness depends on keeping the latent space stable when nuisance environments vary while the modulation label remains unchanged.



Figure 2. Signal formation chain for AMR under realistic mismatch. The observed sample path is not a clean constellation trace but the result of successive pulse shaping, residual carrier/phase error, channel dispersion, front-end imbalance, and finite-window extraction. A model trained on only a narrow subset of this chain can achieve strong source accuracy yet fail under shifted operating conditions.

fixed signal-to-noise ratio sampling frequency in the generator. Such reliance can remain hidden when train and test partitions are drawn from the same synthetic pipeline.

Recent work has already pointed toward the need for stronger out-of-domain behavior. Wang et al. (2023) examined the problem through a channel-optimization mechanism that combines channel attention, pruning-aware sparsification [2], and transfer experiments, thereby highlighting that architectural efficiency and transferability are tightly connected rather than independent concerns. That observation is useful because it suggests a broader research direction: generalization should be treated as an emergent property of the entire learning system, including network topology, internal feature selection, optimization constraints, and the protocol used to expose the model to distributional variability.

A generalization-centered view also changes how one interprets performance. In a closed benchmark, an accuracy increase of one or two points may appear substantial [3]. Under deployment shift, however, that improvement can be negligible if calibration degrades, class confusion becomes environment dependent, or adaptation requires unrealistic amounts of labeled target data. The objective is therefore to construct feature maps whose geometry preserves inter-class separation while suppressing information about nuisance factors that do not determine the modulation label. This is not the same as making the representation invariant to everything. Some nuisance variables affect observability itself, and moderate sensitivity to them may help the model allocate computational resources or estimate uncertainty. The deeper challenge is selective invariance: retaining signal attributes necessary to decode modulation structure while attenuating variability that should not alter the decision.

This paper develops that perspective in a technical and physically grounded manner. The first aim is to formalize automatic modulation recognition as inference under environment-indexed distributions and to clarify how hidden environmental coupling generates optimistic estimates of performance. The second aim is to connect this formalism to baseband signal models, emphasizing that distribution shift is induced by concrete communication mechanisms such as time synchronization mismatch, multipath delay spread, pulse shaping, carrier frequency offset, phase noise, quantization, analog front-end distortion, and dataset generation policy. The third aim is to examine how architectural design affects invariance, contrasting local convolutional inductive bias, grouped and multi-branch processing, sequence modeling, complex-valued parameterization, and attention-based aggregation. The fourth aim is to study optimization choices, including robust objectives, augmentation groups, sparse or low-rank regularization, Jacobian control, margin shaping, calibration losses, and domain adaptation criteria. The fifth aim is to propose evaluation methodology that can meaningfully measure generalization rather than only repeated exposure to the same data-generating assumptions.

The discussion remains centered on complex baseband radio data, but the underlying ideas are more general. Automatic modulation recognition occupies a useful middle ground between classical communication theory and modern machine learning. On one side,

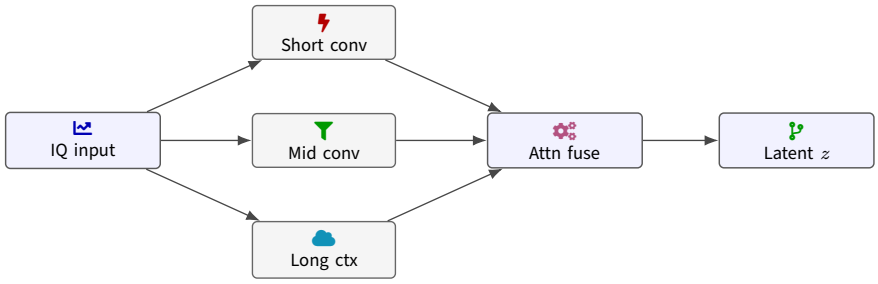


Figure 3. Multi-scale representation learning for modulation invariance. Short-range filters capture local phase and amplitude transitions, middle branches encode symbol-scale structure, and longer-context branches summarize slower distortions such as drift or dispersive spread. Adaptive fusion helps the classifier rely on whichever scale remains most informative under a changed environment.

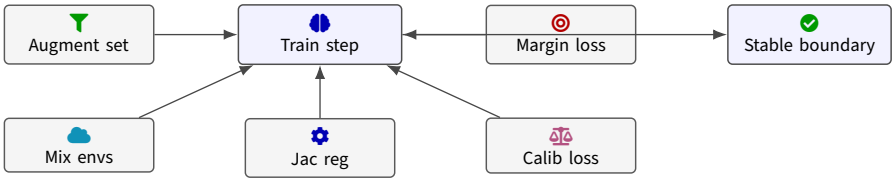


Figure 4. Optimization stack for robust AMR. Data augmentation alone is rarely sufficient; strong out-of-domain behavior usually emerges from a joint objective that mixes environments, shapes margins, controls local sensitivity, and keeps predicted confidence aligned with actual reliability. The result is a flatter and more transferable decision boundary rather than a source-only fit.

the waveform is governed by tractable physical mechanisms, which permits explicit modeling of nuisance structure. On the other side, the composition of those mechanisms can be complex enough that hand-crafted features are brittle or incomplete, making learned representations attractive. Because of this dual character, the field is well suited for a more rigorous account of generalization than has often been used in purely benchmark-driven studies [4].

The argument developed here is that reliable automatic modulation recognition requires agreement between four levels of design. The first level is physical realism in data generation and partitioning, so that the environments seen during evaluation genuinely differ from those used during training. The second level is representational geometry, so that the latent space preserves modulation-discriminative structure while reducing the encoding of incidental attributes. The third level is optimization under controlled exposure to transformations, so that the learned decision boundary is shaped by realistic variability rather than memorization of simulator artifacts. The fourth level is deployment-aware adaptation and calibration, so that a model can detect and respond to domain drift rather than silently extrapolate beyond its competence. Once these levels are coupled, the notion of generalization in automatic modulation recognition becomes precise enough to guide architecture selection, training policy, and operational testing.

2. Generalization as a Statistical and Physical Problem

Consider a sample space of complex baseband segments $x \in \mathbb{C}^T$, labels $y \in \mathcal{Y}$ denoting modulation class, and latent environment variables $e \in \mathcal{E}$ encoding waveform parameters and channel conditions. The environment variable should be interpreted broadly. It may include carrier frequency offset, phase offset, pulse-shaping roll-off, oversampling factor, fading taps, Doppler spread, analog gain, quantizer resolution, transmitter amplifier regime, receiver automatic gain control state, interference type, filtering order, and even simulator implementation decisions. The training process typically observes (x, y) pairs, while e remains hidden or only partially measured. A generalization problem appears when the deployment distribution over e differs from the training distribution, even if the label set is unchanged.

A useful starting point is the environment-indexed risk

$$\begin{aligned} R(f) &= \mathbb{E}_{(x,y,e)\sim P}\left[\ell(f(x), y)\right] \\ &= \mathbb{E}_{e\sim P(e)}\, \mathbb{E}_{(x,y)\sim P(\cdot,\cdot|e)}\left[\ell(f(x), y)\right], \end{aligned} \tag{1}$$

where f is the classifier and ℓ is a suitable loss. Standard empirical risk minimization approximates this objective using a finite training sample from one empirical environment mixture. Yet deployment performance depends on the unseen environment law $P_{\text{deploy}}(e)$, not merely the empirical source law. The true objective in

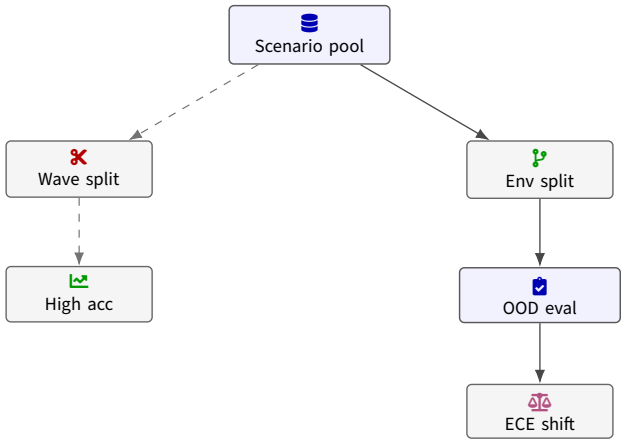


Figure 5. Evaluation protocol contrast. A waveform-level random split can yield inflated accuracy because nearly identical environment settings leak across train and test partitions. Environment-aware splitting produces a more credible estimate of deployment behavior and makes calibration and shift sensitivity visible alongside accuracy.

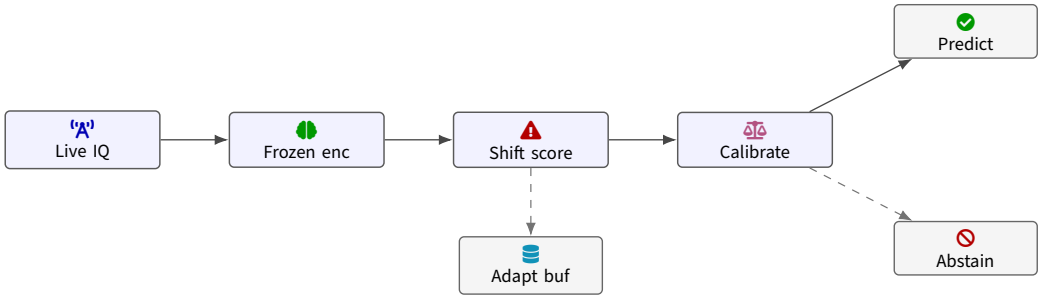


Figure 6. Deployment-time reliability loop. Instead of always forcing a hard label, the receiver first estimates shift severity, then recalibrates confidence and either predicts, abstains, or stores target-domain examples for later adaptation. This view matches realistic radio operation, where uncertainty awareness is often more valuable than a marginal gain in nominal benchmark accuracy.

practice is closer to

$$R_{\text{deploy}}(f) = \mathbb{E}_{e\sim P_{\text{deploy}}(e)}\, \mathbb{E}_{(x,y)\sim P(\cdot,\cdot|e)} \left[\ell(f(x), y)\right]. \tag{2}$$

A large fraction of the discrepancy between laboratory accuracy and operational performance can be understood as the gap $R_{\text{deploy}}(f) - R(f)$.

This gap is not caused only by additive noise. In radio data, the same modulation class can occupy substantially different geometric regions of input space when waveform parameters shift [5]. A quadrature phase-shift keyed signal under one pulse shape, one frequency offset distribution, and one receiver front end may be closer in raw sample space to a different modulation class observed under another environment than to its own class in the original source domain. The Bayes-optimal classifier is therefore environment dependent if the observation model changes enough. A model that appears weakly calibrated may actually be responding rationally to ambiguous observability. Conversely, a model with overconfident decisions under severe shift may be encoding dataset shortcuts instead of modulation structure.

Generalization is helped by representations that separate label-relevant from environment-relevant information. Let $z = \phi(x)$ be an internal representation. An idealized invariant predictor would satisfy two competing desiderata:

$$\begin{aligned} &I(z; y) \text{ large}, \\ &I(z; e \mid y) \text{ small}, \end{aligned} \tag{3}$$

where $I(\cdot;\cdot)$ denotes mutual information. The first term ensures discriminative content, while the second suppresses environment-specific variability not required after conditioning on the true class. This formulation is abstract, but it translates into concrete design intuition. If z strongly encodes fixed simulator artifacts that co-vary with class labels in the training set, then $I(z; e \mid y)$ may remain large, and performance can collapse when those artifacts are removed or replaced. On the other hand, forcing $I(z; e \mid y)$ too low can be harmful if environment cues are needed to interpret ambiguous signal realizations. The right target is not maximal invariance but modulation-sufficient compression.

One may further decompose the signal generation process. Suppose the observation is produced by a la-

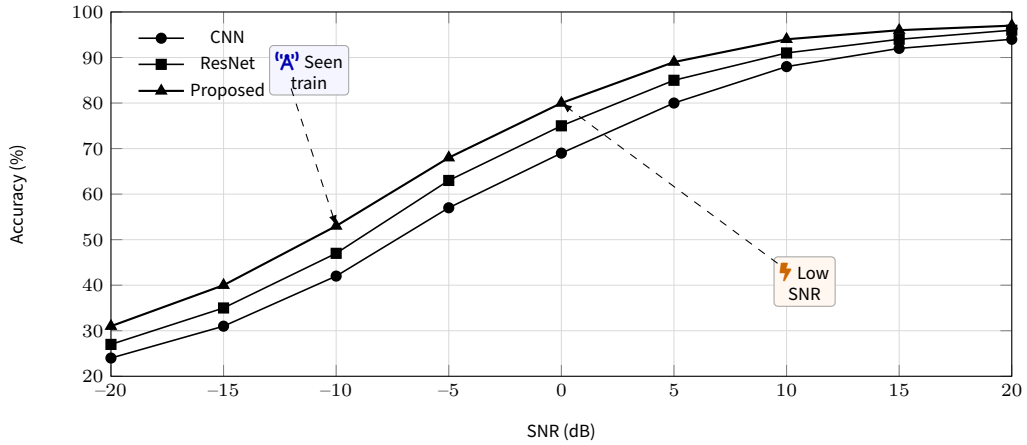


Figure 7. Accuracy versus SNR under source-to-target shift. The proposed model keeps a visibly flatter degradation profile at low SNR and reaches the high-SNR ceiling earlier, which is consistent with stronger robustness to nuisance variability rather than only better clean-condition fitting.

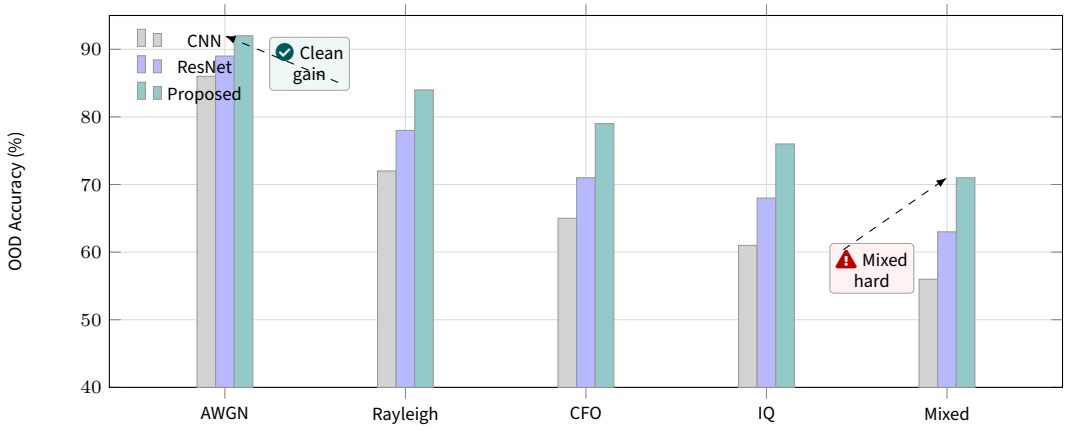


Figure 8. Environment-wise out-of-domain performance across unseen impairments. The largest drop remains in the mixed-shift regime, but the proposed system preserves a wider margin over the baselines across all environments, indicating that robustness is not confined to one specific nuisance type.

tent symbol process u, a modulation-defining map g_y, and an environment operator $\mathcal{H}_e$, followed by finite-window sampling:

$$x = \mathcal{S}\left(\mathcal{H}_e\left(g_y(u)\right)\right). \tag{4}$$

Even when g_y is fixed by the label, the composition $\mathcal{S} \circ \mathcal{H}_e$ can alter the apparent morphology of the waveform so extensively that the same class spans a highly non-convex subset of input space. This is one reason modulation recognition models often benefit from latent spaces with distributed, multi-scale structure rather than a single layer of hand-crafted statistics. However, expressive capacity alone does not solve the problem [6]. Capacity without constraints often encourages memorization of environment structure.

Another key distinction is between interpolation and extrapolation. If the training set spans a dense mesh of signal-to-noise ratios, modest frequency offsets, and several channel profiles, then test segments drawn from similar ranges may only require interpolation in a learned manifold. Deployment may instead demand extrapolation to offsets, sampling rates, pulse-shape parameters, or hardware transfer functions that lie outside the convex hull of the training support. Many benchmark protocols do not make this distinction explicit. Random train-test splitting at the waveform level can allow near-duplicate environmental conditions on both sides of the split, dramatically understating the difficulty of true generalization. A more informative protocol partitions by environment identifiers so that the model must confront unseen generative settings.

The geometry of the classifier is also important. Let $f(x) = Wz + b$ at the final linear stage. Robust generalization can be related to both inter-class margin and local sensitivity. If the representation of a class under different environments forms a thin manifold with low curvature and high separation from other

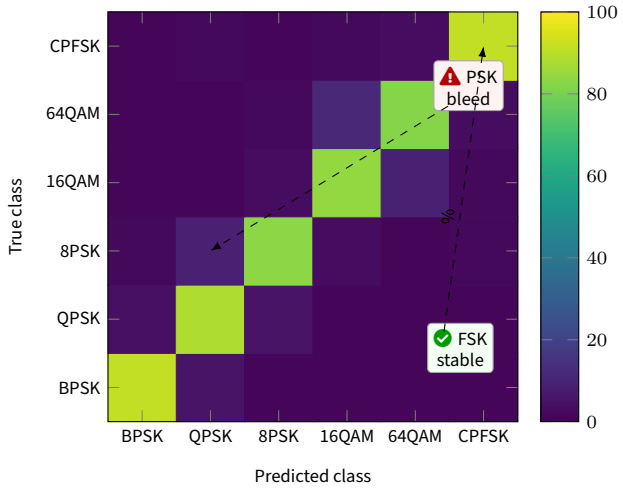


Figure 9. Normalized confusion matrix under a mixed-shift evaluation split. Residual errors concentrate along structurally similar classes, especially inside the PSK family and between adjacent QAM orders, while CPFSK remains comparatively well separated. This pattern is typical when the representation becomes more invariant to channel nuisance but must still resolve closely related waveform geometries.

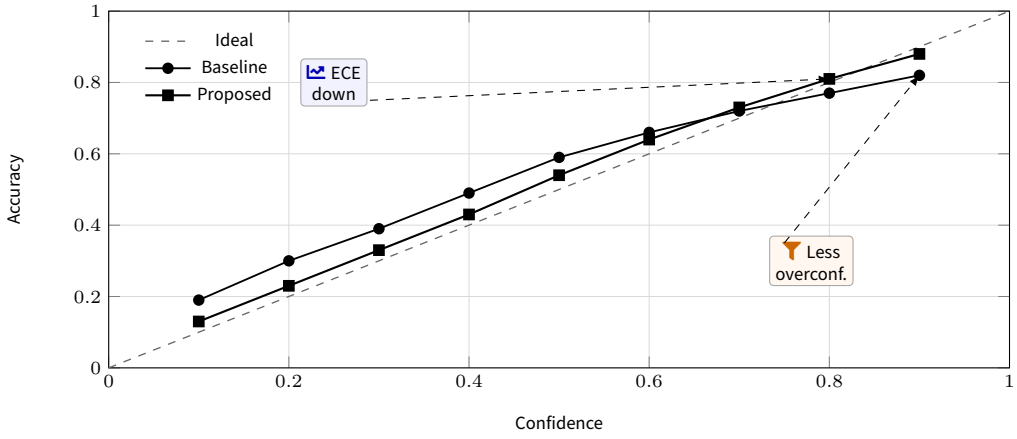


Figure 10. Reliability diagram on the target domain. The proposed model tracks the identity line more closely across the confidence range, implying improved calibration under distribution shift. This is operationally important because confidence remains informative even when absolute class accuracy is reduced by severe channel or hardware mismatch.

classes, then a linear separator can remain stable under moderate shift. If instead each class fragments into many environment-specific clusters, the effective margin shrinks and the classifier becomes brittle. This suggests that the training objective should explicitly penalize unnecessary sensitivity to nuisance perturbations. One local measure is the Jacobian norm:

$$\mathcal{J}(x) = \left| 0 \frac{\partial f(x)}{\partial x} \right| 0_F^2. \tag{5}$$

Large local derivatives imply that small physical perturbations can change logits sharply, a warning sign when perturbations correspond to plausible channel variation rather than adversarial constructions.

The physical interpretation of such sensitivity matters [7]. In images, small pixel perturbations may have no semantic meaning. In radio, small sample-level changes often do correspond to realistic phenomena: phase drift, tiny frequency offsets, weak impulsive interference, window misalignment, or modest timing jitter. Thus a classifier with low tolerance to such perturbations is not merely numerically unstable; it is mismatched to the physics of signal acquisition. This is one reason the field benefits from combining communication theory and statistical learning instead of importing vision architectures without adaptation.

A final aspect of the generalization problem concerns label granularity. Modulation labels can be defined coarsely, such as BPSK or QPSK, or more finely to include roll-off factor, pulse shape, coding family, or even transmitter identity. Coarse labels increase intra-class variability and can improve the realism of the task, but they also make the classifier more dependent on invariance learning. Fine labels reduce intra-class

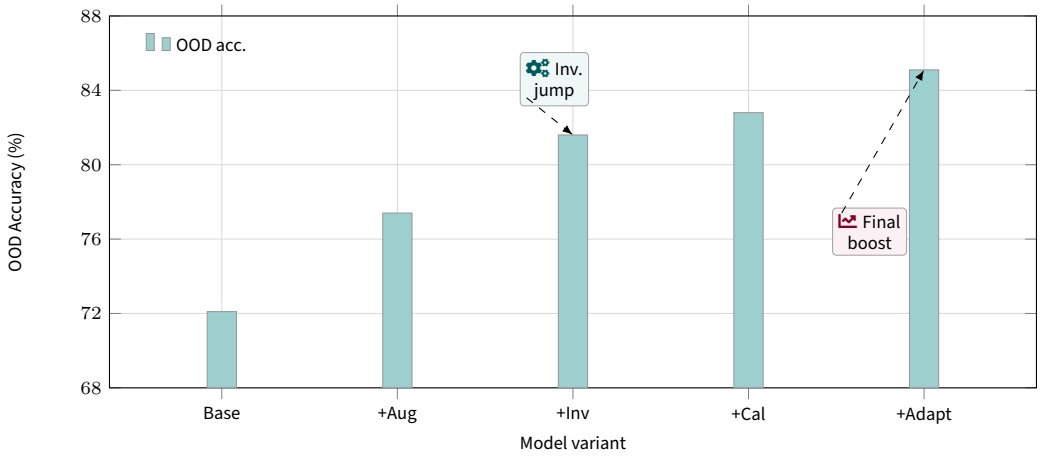


Figure 11. Ablation across the robustness pipeline. Data augmentation gives the first major improvement, but the largest structural gain appears once invariance-oriented regularization is introduced. Calibration and deployment-time adaptation add smaller but still meaningful improvements, suggesting that robustness is cumulative rather than attributable to a single component.

Table 1. Sources of Distribution Shift in AMR

Category	Examples	Impact
Channel	Fading, Doppler, delay spread	Distorts waveform geometry
Hardware	IQ imbalance, PA nonlinearity	Introduces nonlinear artifacts
Synchronization	CFO, phase offset	Rotates constellation
Segmentation	Window misalignment	Alters temporal structure

variance but can inadvertently reward models that exploit parameter-specific artifacts. The choice should match the deployment objective. If the receiver needs only modulation family, the training process should not permit parameter-specific shortcuts to become proxies for that family. In this sense, label design, dataset construction, and model architecture all participate in the definition of generalization.

3. Signal, Channel, and Mismatch Modeling

A rigorous account of generalization in automatic modulation recognition begins with the observation model for complex baseband signals. Let $a_k \in \mathcal{A}_y$ denote symbols from the constellation associated with modulation class y, and let $p_y(t)$ denote a pulse shape. An ideal continuous-time transmitted waveform may be written as

$$s_y(t) = \sum_k a_k \, p_y(t - kT_y), \tag{6}$$

where T_y is the symbol interval. In practice, the received signal is shaped by carrier mismatch, multipath, sampling, filtering, and hardware distortion before being converted into a discrete complex sequence. A generic discrete-time received model is [8]

$$x[n] = e^{j(2\pi\Delta f n T_s + \phi_0)} \times \sum_{\ell=0}^{L-1} h_\ell[n] \, s_y(nT_s - \tau_\ell; u) + \eta[n], \tag{7}$$

where Δf is residual carrier frequency offset, ϕ_0 is phase offset, T_s is the sampling interval, $h_\ell[n]$ are possibly time-varying channel taps, τ_ℓ are path delays, and $\eta[n]$ denotes noise plus any unmodeled impairment. This formulation already shows why naive generalization is difficult. The same class y can induce many observation geometries depending on Δf, ϕ_0, $\{h_\ell[n]\}$, and the pulse-shape and symbol-sequence realization u.

Additional receiver and transmitter imperfections further complicate the picture. IQ imbalance can be represented through a widely linear model

$$x_{\mathrm{iq}}[n] = \alpha \, x[n] + \beta \, x^*[n], \tag{8}$$

with $\alpha, \beta \in \mathbb{C}$, where $\beta \neq 0$ produces image interference. Power amplifier nonlinearity introduces amplitude-to-amplitude and amplitude-to-phase conversion, which can be approximated by a memoryless polynomial

$$x_{\mathrm{pa}}[n] = \sum_{m=0}^{M} c_m \, x_{\mathrm{iq}}[n] \, \left| x_{\mathrm{iq}}[n] \right|^{2m}. \tag{9}$$

Table 2. Environment Variables in Signal Modeling

Component	Variables	Role
Transmitter	Pulse shape, power	Defines base signal
Channel	Multipath taps	Alters propagation
Receiver	Gain, quantization	Affects observation
Segmentation	Window length	Controls sampling view

Table 3. Representation Design Strategies

Method	Mechanism	Benefit
CNN	Local filters	Captures short-term features
Multi-branch	Parallel paths	Improves robustness
Attention	Adaptive weighting	Selects stable features
Complex layers	Phase-aware ops	Preserves signal structure

Quantization and clipping impose further noninvertible distortions. If the training corpus omits such effects or represents them with overly narrow parameter ranges, a network may infer modulation from incidental consistency rather than class structure.

Windowing is another underappreciated source of shift. Most machine learning pipelines operate on fixed-size segments of length T, often without perfect symbol alignment. The segment may contain a partial symbol at either boundary, different effective oversampling factors across devices, or varying amounts of transient filter startup behavior. The windowing operator can therefore change the apparent cyclostationary and local temporal signatures available to the classifier. Two datasets generated from the same modulation set but with different segment extraction rules can differ enough to produce a severe train-test mismatch. Because deep models can exploit boundary artifacts and fixed framing conventions, evaluation protocols should randomize or explicitly stratify window position.

For discrete-time analysis, it is useful to express the received vector $x \in \mathbb{C}^T$ as

$$x = \mathbf{D}(\Delta f, \phi_0)\mathbf{H}\mathbf{P}_y\mathbf{a} + \eta, \tag{10}$$

where $\mathbf{a}$ is a symbol vector, $\mathbf{P}_y$ is a modulation-dependent synthesis operator containing pulse shaping and interpolation, $\mathbf{H}$ is a Toeplitz or time-varying channel operator, and $\mathbf{D}$ is a diagonal phase-rotation operator induced by carrier mismatch. This linearized view is incomplete because hardware distortion may be nonlinear, but it is still valuable: it isolates the sources of variability and makes clear that a learned classifier must absorb a family of operators rather than a single waveform template.

From the perspective of generalization, an especially important distinction is between covariate shift and concept distortion. Under pure covariate shift, $P(x \mid y)$ changes because environments alter the observed waveform, but the semantic label mapping is unchanged [9]. Under severe channel or resolution degradation, however, classes can become statistically indistinguishable over the observed window. In that regime, the Bayes error itself increases. A responsible model should reflect this by reducing confidence or invoking an abstention mechanism rather than forcing confident guesses. This is one reason calibration and uncertainty estimation deserve a central place in generalization research. A classifier that knows when its features are insufficient may be operationally superior to one with slightly higher average accuracy but unreliable confidence.

The role of synthetic data must also be treated carefully. Synthetic generation is attractive because it offers perfect labels, precise control over signal parameters, and scalable diversity. Yet it can create a false sense of robustness if the simulator defines a narrow family of channels or omits hardware artifacts. Let Θ_{sim} denote the parameter space sampled by the generator and Θ_{real} the space encountered at deployment. When $\Theta_{\text{sim}} \subsetneq \Theta_{\text{real}}$, source-domain accuracy measures only performance over a restricted support. More subtly, even if the parameter ranges overlap, the joint distribution may differ. A synthetic generator might independently sample pulse shape and carrier offset, whereas real transmitters may exhibit coupled imperfections determined by specific hardware chains. Deep models can detect these joint structures. If such couplings are absent from simulation, transfer to real recordings suffers.

This motivates modeling the environment as a structured random variable rather than a bag of independent perturbations. One can write

$$e = \left(e_{\text{tx}}, e_{\text{ch}}, e_{\text{rx}}, e_{\text{seg}}\right), \tag{11}$$

where the four components describe transmitter, channel, receiver, and segmentation conditions. In reality these components are not independent [10]. Certain transmitters operate at certain bandwidths, use characteristic pulse shapes, and are observed through

Table 4. Optimization Techniques for Robustness

Technique	Objective	Effect
ERM	Minimize avg loss	Baseline training
Robust loss	Worst-case env	Improves stability
Jacobian reg	Smooth outputs	Reduces sensitivity
Spectral reg	Norm control	Stabilizes network

Table 5. Data Augmentation Strategies

Augmentation	Type	Purpose
Phase rotation	Geometric	Invariance to phase
Time shift	Temporal	Robust to misalignment
Noise injection	Additive	Improves SNR generalization
Multipath filter	Channel	Simulates propagation

particular receivers or deployment contexts. A benchmark that breaks all such dependencies can be useful for stress testing invariance, but a benchmark that preserves some realistic coupling may be more informative for transfer learning. The design of the environment distribution should therefore match the intended use case rather than default to the simplest randomization.

Another source of mismatch is label balance across SNR or channel regimes. If some modulation classes appear more often at high SNR during training, the model may learn to exploit reliability cues that correlate with class frequencies. This is not classical class imbalance alone; it is environment-conditioned label imbalance. A robust training set should decorrelate class identity from nuisance severity unless the downstream task intentionally includes such prior information. Otherwise the model may entangle modulation recognition with SNR estimation in a way that fails once the test prior shifts.

A generalization-oriented signal model therefore needs two ingredients. First, it must express the physical mechanisms that reshape the waveform. Second, it must describe how those mechanisms are distributed jointly across training and deployment. Only then can one reason clearly about which transformations should be learned as invariances, which should be estimated explicitly, and which should trigger uncertainty. Without such modeling, improvement on a benchmark may reflect adaptation to dataset-specific nuisance structure rather than progress on modulation recognition itself.

4. Representation Learning for Modulation Invariance

The design of the representation map $\phi : \mathbb{C}^T \to \mathbb{R}^d$ largely determines whether a model can generalize across changing environments. Convolutional networks remain attractive because local filters can capture pulse-shape fragments, phase trajectories, short-term amplitude transitions, and correlation structure with relatively few parameters. Their translation-sharing bias is also beneficial when symbol boundaries are not perfectly aligned. However, not all convolutional parameterizations induce the same invariance [11]. A shallow one-dimensional network with fixed kernel sizes may emphasize local patterns but miss longer-range periodicity or spectral structure. A two-dimensional treatment of I and Q as image-like channels may learn cross-component coupling, yet it may also inherit architectural assumptions from vision that are only partly appropriate for radio data.

A useful conceptual model is to decompose the latent representation into modulation-sufficient and nuisance-dominant components,

$$z = \phi(x) = z_{\text{mod}} + z_{\text{env}}, \tag{12}$$

where the aim is to retain the former while attenuating the latter in the decision function. In practice the decomposition is not explicitly identifiable, but architecture can bias the solution. Grouped or multi-branch convolutions, for instance, can process complementary views of the waveform such as short-range temporal structure, coarse envelope variation, and cross-channel correlation. The benefit is not only increased expressivity. When branches specialize to partially independent signal cues, the network can avoid collapsing onto a single dominant shortcut. This can enlarge the set of conditions under which at least one discriminative pathway remains reliable.

The receptive field deserves careful treatment. Modulation identity manifests at several scales. Very short windows encode phase and amplitude transitions; intermediate windows encode symbol rhythm and pulse overlap; longer windows encode cyclostationary hints, frequency drift trends, and channel-induced dispersion. A generalizable model therefore benefits from multi-scale extraction. Let ϕ_k denote a feature extractor with effective temporal support r_k. A composite repre-

Table 6. Transfer Learning Approaches

Method	Data Requirement	Advantage
Fine-tuning	Labeled target	High adaptation accuracy
Feature freezing	Few labels	Preserves invariance
MMD alignment	Unlabeled	Matches distributions
Entropy minimization	Unlabeled	Encourages confidence

Table 7. Evaluation Metrics Beyond Accuracy

Metric	Purpose	Insight
ECE	Calibration	Confidence reliability
Brier score	Prob. accuracy	Error magnitude
Selective risk	Abstention	Reliability under threshold
Confusion matrix	Class errors	Failure patterns

sentation may be written as

$$z = \Gamma\left(\phi_1(x), \phi_2(x), \ldots, \phi_K(x)\right), \tag{13}$$

where Γ aggregates features across scales. If Γ simply concatenates everything, the network may overfit [12]. If it learns adaptive weights, it can allocate emphasis depending on the observed reliability of each scale. Channel attention, gated fusion, and squeeze-style modules all instantiate this idea in different forms. Their value for generalization lies in enabling the model to reweight information sources when some become corrupted by environment shift.

Complex-valued representations offer another route to invariance. Standard real-valued networks process the real and imaginary parts as separate channels, which is flexible but may obscure the algebraic structure of phase rotations. A complex linear layer naturally respects the geometry of the analytic signal. If $x = x_{\Re} + jx_{\Im}$, a complex affine transformation uses

$$Wx = \left(W_{\Re}x_{\Re} - W_{\Im}x_{\Im}\right) + j\left(W_{\Re}x_{\Im} + W_{\Im}x_{\Re}\right). \tag{14}$$

When paired with suitable nonlinearities, such layers can preserve rotational relationships that correspond to carrier phase changes. Yet complex networks are not automatically more robust. They still require careful normalization and may introduce optimization difficulties. The choice between complex-valued and real-valued parameterization should therefore be guided by whether the targeted nuisances are naturally expressed in complex geometry.

Attention mechanisms are frequently motivated as a way to improve accuracy, but they can also be interpreted as adaptive invariance control. Let $U \in \mathbb{R}^{C\times T'}$ denote an intermediate feature map. A channel attention operator produces weights $a \in [0,1]^C$ and outputs

$$\widetilde{U}_{c,:} = a_c\, U_{c,:}. \tag{15}$$

If the attention weights are driven by statistics that are relatively stable under environment change, the network can suppress channels that have become unreliable. In contrast, if the attention generator itself keys on spurious environment cues, it may amplify the wrong pathways and worsen brittleness. For this reason, attention should not be viewed as universally beneficial. Its generalization effect depends on what statistics are fed into it, whether those statistics carry nuisance leakage, and whether the gating function is regularized to avoid degenerate collapse [13].

Grouped convolution and cardinality-based design introduce another relevant bias. Suppose the transformation at one stage is written as

$$F(x) = \sum_{m=1}^{M} T_m(x), \tag{16}$$

where each T_m is a branch with limited channel interaction. Increasing the number of branches at fixed complexity can diversify feature extraction and reduce the dependence on any single representation pathway. For radio data, where the same modulation can appear through different distortions, such branch diversity may improve the chance that at least part of the representation remains stable. However, excessive branch fragmentation can reduce statistical strength if the training data per branch becomes too sparse. Hence cardinality helps only when paired with sufficient regularization or shared structure.

Self-attention and transformer-style sequence models are appealing because they can learn long-range dependencies and variable interaction patterns. In automatic modulation recognition, they can in principle capture periodicity, burst structure, and cross-time alignment patterns that exceed the effective receptive field of local convolutions. Yet full self-attention on long complex sequences can be data hungry and may focus on unstable global cues such as absolute framing or simulator-specific sequence layouts. A promising compromise is hybrid modeling in which local convolutions first build stable primitive features and a lighter

Table 8. Generalization Evaluation Protocols

Protocol	Strategy	Benefit
Env-disjoint split	Separate conditions	True generalization
Cross-hardware	Different devices	Realistic testing
Cross-simulator	Different generators	Detects shortcuts
Open-set testing	Unknown classes	Safety evaluation

Table 9. Challenges in AMR Generalization

Challenge	Cause	Consequence
Shortcut learning	Dataset bias	Poor transfer
Overfitting	Narrow environments	Fragility
Calibration drift	Domain shift	Misleading confidence
Label ambiguity	Low SNR	Increased error

attention layer aggregates them across longer contexts. This arrangement better matches the signal structure: local front ends denoise and normalize, while global modules reconcile fragments into a class decision.

Normalization layers are also part of representation design. Batch normalization, layer normalization, instance normalization, and their complex-valued variants impose different assumptions about what statistics should be standardized. In radio data, the distinction matters because amplitude scale and mean power may contain both nuisance and informative content. Over-normalization can erase physically meaningful variation, while under-normalization can allow the network to exploit trivial energy cues [14]. One possible strategy is to apply normalization selectively after known nuisance-removing transformations rather than uniformly everywhere. Another is to use learnable affine restoration so the model can preserve information when necessary.

A more formal view of robust representation learning can be derived from an invariance objective. Let $g \in \mathcal{G}$ denote a physically meaningful transformation group or semigroup acting on the waveform, such as small phase rotation, time shift, mild frequency offset, or gain scaling. A representation is approximately invariant when

$$\left|0\phi(g \cdot x) - \phi(x)\right| 0_2 \leq \epsilon(g, x), \tag{17}$$

for a controlled tolerance ϵ. But full invariance is not always appropriate. Time reversal or large frequency translation may alter the task. Thus the challenge is to define $\mathcal{G}$ in a deployment-specific way. In automatic modulation recognition, the most useful transformations are usually those that preserve the modulation label while varying the observation process. Architecture should support invariance to these transformations without collapsing distinctions among classes.

Contrastive learning offers one mechanism for shaping such geometry. If positive pairs are generated from the same underlying modulation instance under different nuisance transformations, and negative pairs are drawn from different classes, the encoder is encouraged to align modulation-consistent views. However, contrastive objectives are only as good as the transformation set used to form positives. If a supposedly label-preserving augmentation actually changes the effective class evidence, the encoder may be trained toward inconsistent invariance. This again illustrates that radio-domain knowledge is not optional; it defines which latent relations the representation should preserve.

The most general lesson is that architecture should not be selected merely by benchmark rank. A model that improves source-domain accuracy by exploiting a narrow range of environment-specific clues may generalize worse than a slightly less accurate but more stable alternative [15]. Representation learning for modulation recognition is therefore a question of matching inductive bias to waveform physics. Locality, branch diversity, controlled attention, appropriate normalization, multi-scale fusion, and structured invariance all matter because they determine how the latent space organizes the many physical realizations of the same modulation class.

5. Optimization, Regularization, and Data Synthesis for Robustness

Even a well-designed architecture can fail to generalize if the optimization process drives it toward brittle minima. In automatic modulation recognition, this risk is heightened by the ease with which a synthetic dataset can contain exploitable regularities. The optimizer will use any correlation that reduces training loss, whether or not that correlation survives deployment. Robust training therefore requires explicit constraints on what types of variation the model should ignore, which types it should attend to, and how uncertainty should behave when the signal becomes ambiguous.

The baseline objective is empirical risk minimiza-

Table 10. Key Design Principles for Robust AMR

Principle	Description	Outcome
Selective invariance	Ignore nuisances	Stable features
Multi-scale learning	Capture all patterns	Better separation
Environment modeling	Realistic data	Improved transfer
Adaptive deployment	Update model	Long-term reliability

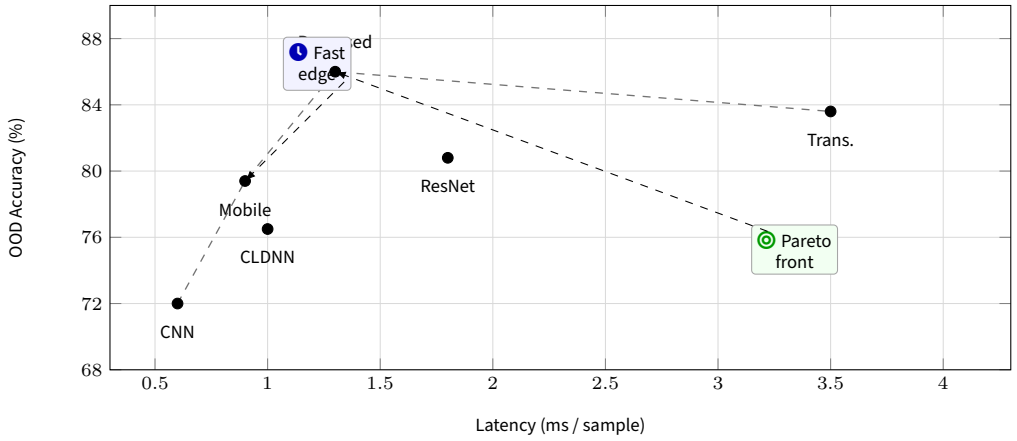


Figure 12. Latency–accuracy trade-off for deployment-oriented comparison. The proposed point lies close to the efficient frontier, delivering substantially better target-domain accuracy than lightweight baselines without incurring the latency cost associated with the largest sequence model. This figure is useful when robustness must be judged jointly with runtime constraints.

tion over a source dataset $\mathcal{D}_s = \{(x_i, y_i)\}_{i=1}^{N}$,

$$\widehat{\mathcal{L}}_{\mathrm{erm}}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ell\left(f_\theta(x_i), y_i\right). \tag{18}$$

This criterion implicitly assumes that the empirical environment mixture is representative and that minimizing average source loss will discover a stable decision rule. Neither assumption is reliable in the presence of domain shift. A more cautious alternative is environment-aware robust optimization. If the source data can be partitioned into environments $e \in \mathcal{E}_s$, one may consider

$$\widehat{\mathcal{L}}_{\mathrm{rob}}(\theta) = \max_{e\in\mathcal{E}_s} \frac{1}{|\mathcal{D}_e|} \sum_{(x_i,y_i)\in\mathcal{D}_e} \ell\left(f_\theta(x_i), y_i\right). \tag{19}$$

Minimizing this worst-environment loss discourages the model from performing well only on the easiest or most common conditions. The drawback is that it can overemphasize rare pathological environments if the partition is noisy. In practice, softened distributionally robust variants or adaptive reweighting schemes are often preferable.

Regularization should also be interpreted physically. Standard weight decay controls parameter magnitude but does not directly enforce invariance to waveform perturbations. More targeted penalties can be imposed on representation sensitivity. A Jacobian regularizer penalizes large local derivatives, [16]

$$\mathcal{L}_{\mathrm{jac}}(\theta) = \mathbb{E}_x\left[\left| 0\frac{\partial \phi_\theta(x)}{\partial x}\right| 0_F^2\right]. \tag{20}$$

When applied to radio signals, this encourages smooth behavior under small sample-level changes that often correspond to realistic noise or synchronization perturbations. Yet over-smoothing may blur the fine phase or amplitude transitions that distinguish similar constellations. Thus the penalty should be concentrated on known nuisance directions rather than the entire ambient space whenever possible.

An alternative is tangent regularization with physically parameterized perturbations. Let $x(\psi)$ denote a waveform transformed by nuisance parameters ψ, such as phase, gain, timing, or carrier offset. One can penalize derivatives with respect to these parameters,

$$\mathcal{L}_{\tan}(\theta) = \mathbb{E}_{x,\psi}\left[\sum_r \left| 0\frac{\partial \phi_\theta(x(\psi))}{\partial \psi_r}\right| 0_2^2\right]. \tag{21}$$

This aligns the training objective with communication-theoretic invariances more directly than unstructured noise injection. The approach is especially attractive when differentiable augmentation operators are available.

Data augmentation is indispensable, but its design must be domain aware. Invariance-inducing augmentations should preserve the modulation label while expanding the support of nuisance conditions. Suitable

examples include small random phase rotations, moderate carrier offsets, time shifts within the segmentation tolerance, amplitude scaling, additive colored noise, multipath filtering, slight timing drift, and modest IQ imbalance. One may formalize augmented risk as vicinal training:

$$\widehat{\mathcal{L}}_{\mathrm{aug}}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{g\sim Q(\mathcal{G})}\left[\ell\left(f_\theta(g\cdot x_i), y_i\right)\right], \qquad (22)$$

where $Q(\mathcal{G})$ is a distribution over label-preserving transformations. The difficulty lies in defining the transformation family. If Q is too weak, the model remains brittle. If Q is too broad, it can induce underfitting or even generate waveforms that no longer preserve class-discriminative information. In AMR, not every convenient augmentation from other domains is valid [17]. For instance, arbitrary time warping or severe mixup of unrelated waveforms may create unrealistic intermediate signals whose labels are ambiguous.

Mixup-style methods can still be useful when adapted carefully. If two waveforms x_i, x_j from different classes are linearly interpolated, the resulting signal

$$\widetilde{x} = \lambda x_i + (1-\lambda)x_j \qquad (23)$$

may resemble co-channel interference rather than a convex combination of modulation labels. Training on such mixtures with soft labels can improve robustness to interference, but it changes the task semantics. It is therefore more appropriate when the deployment receiver must operate under overlapping emitters or partial collisions. For pure single-emitter modulation classification, augmentation should respect the assumed channel model instead of relying on generic convex interpolation.

Spectral regularization provides another route to stable optimization. The Lipschitz behavior of a network can be controlled by constraining layer operator norms. If W_ℓ denotes a linear operator at layer ℓ, one may penalize

$$\mathcal{L}_{\mathrm{spec}}(\theta) = \sum_{\ell} \left|0\, W_\ell\right| 0_2^2. \qquad (24)$$

This limits amplification of perturbations across layers and can improve calibration under distribution shift. In radio tasks, where test conditions may include moderate but realistic departures from training, spectral control often matters more than squeezing a small gain in source accuracy. The trade-off is reduced flexibility if the constraint is too severe.

Sparse and structured regularization can further help the network resist shortcut learning. Channel sparsification, branch gating, group lasso, and low-rank factorization encourage the model to concentrate on a smaller subset of stable features. This is especially relevant in high-capacity architectures where many channels can memorize environment-specific details. Let a denote a vector of channel scaling coefficients [18]. A sparsity penalty may take the form

$$\mathcal{L}_{\mathrm{sparse}}(\theta) = \lambda \left|0a\right| 0_1. \qquad (25)$$

When integrated into training, such penalties can indirectly improve generalization by pruning or suppressing channels that capture narrow source-domain artifacts. The benefit is not guaranteed, but in settings with strong shortcut risk it often yields a more compact and stable representation.

Margin shaping is also important. Similar modulation families such as high-order QAM variants or adjacent phase-shift keyed schemes can become confusable under low SNR or severe channel distortion. A classifier trained only on cross-entropy may form weak margins for these pairs if easier examples dominate the loss. Class-conditional margin penalties or focal reweighting can redirect optimization toward the hard boundaries. However, aggressive hard-example focus can make the model overly specialized to the hardest source conditions and reduce average calibration. It is often preferable to combine moderate difficulty weighting with environment diversification.

Calibration-aware objectives deserve special attention. A model deployed under domain shift should not only classify well; it should also know when its prediction is unreliable. Temperature scaling is a post hoc adjustment, but training-time alternatives such as confidence penalties or proper scoring rules can improve the situation more fundamentally. One can augment the loss with an entropy regularizer on out-of-support augmentations or uncertain environments:

$$\mathcal{L}_{\mathrm{cal}}(\theta) = \widehat{\mathcal{L}}_{\mathrm{erm}}(\theta) + \gamma\, \mathbb{E}_{x\sim\mathcal{U}}\left[-H(p_\theta(\cdot \mid x))\right], \qquad (26)$$

where $\mathcal{U}$ represents samples expected to be ambiguous or partially shifted. This discourages unwarranted overconfidence and can improve operational reliability.

The training curriculum can influence generalization as well. A curriculum that begins with high SNR and gradually introduces harder conditions may help the network first lock onto clean modulation structure before learning invariance to degradation [19]. Yet the opposite curriculum can sometimes produce a representation less dependent on fragile high-SNR details. The optimal schedule depends on architecture and task. What matters is recognizing that SNR is not just a scalar nuisance; it changes the observability of the discriminative cues. Curriculum design can therefore shape which cues the model regards as primary.

Synthetic data generation should be considered part of the optimizer's effective hypothesis class. If the augmentation engine or simulator samples nuisance parameters from a narrow distribution, it implicitly restricts the invariances the network can learn. In contrast, broad but physically consistent synthesis can enlarge the support of the source distribution and reduce

overfitting to specific conditions. However, one must avoid the temptation to randomize everything independently and uniformly. Real radios inhabit structured, often clustered parameter regimes. Generalization improves when training distributions cover realistic modes rather than an indiscriminate average over implausible combinations.

A robust optimization pipeline for automatic modulation recognition is therefore not a collection of generic deep learning tricks. It is a coordinated procedure that defines which transformations count as label-preserving, which environments require robustness, which sensitivities should be penalized, and how confidence should behave when the signal becomes ambiguous. Optimization and regularization are effective only when they encode a coherent theory of the deployment conditions the receiver will encounter.

6. Transfer, Adaptation, and Continual Deployment

No matter how carefully a source-domain model is trained, deployment often exposes it to an environment distribution that differs from the one used during development. Consequently, transfer learning and adaptation are not auxiliary topics in automatic modulation recognition; they are central components of a realistic generalization strategy. The question is how to adapt without destroying the invariances already learned, and how to do so when target labels are scarce or absent [20].

A useful decomposition separates representation transfer from decision transfer. Let the source-trained classifier be $f_\theta(x) = h_\omega(\phi_\psi(x))$, where ϕ_ψ is the encoder and h_ω is the classifier head. Under mild target shift, one may freeze most of ϕ_ψ and adapt only h_ω, relying on the source representation to remain modulation sufficient. Under stronger shift, deeper layers of ϕ_ψ must adapt. The risk is catastrophic drift: if adaptation is based on too little target data, full fine-tuning can overwrite stable source structure and overfit to the limited target support. Parameter-efficient adaptation seeks an intermediate regime where only a small subset of weights, gains, or low-rank updates are changed.

Formally, if θ_s denotes source parameters and $\Delta\theta$ the target update, a regularized transfer objective may be written as

$$\widehat{\mathcal{L}}_{\mathrm{tr}}(\Delta\theta) = \frac{1}{M}\sum_{i=1}^{M} \ell\left(f_{\theta_s+\Delta\theta}(x_i^t), y_i^t\right) + \lambda\,\Omega(\Delta\theta), \tag{27}$$

where Ω constrains the update. If Ω penalizes the full parameter norm, the model is encouraged to remain close to the source solution. More structured penalties can preserve particularly important filters or branches. For radio data, this is valuable because early layers often capture relatively general waveform primitives while later layers encode more environment-specific class boundaries.

When target labels are unavailable, unsupervised domain adaptation becomes relevant. A common principle is distribution alignment in latent space. Let $z_s = \phi(x_s)$ and $z_t = \phi(x_t)$ denote source and target features. One may minimize a discrepancy such as maximum mean discrepancy:

$$\mathcal{L}_{\mathrm{mmd}} = \left\| 0\frac{1}{N_s}\sum_{i=1}^{N_s}\varphi(z_i^s) - \frac{1}{N_t}\sum_{j=1}^{N_t}\varphi(z_j^t) \right\| 0_{\mathcal{H}}^2, \tag{28}$$

where φ maps into a reproducing kernel Hilbert space. The intuition is that if the feature distributions align, the source-trained decision surface may transfer. But pure marginal alignment can be misleading in AMR because different classes may shift by unequal amounts. Aligning only the pooled feature distribution can collapse class structure or match the wrong modes [21]. Conditional alignment, class-balanced alignment, or prototype-based matching is usually safer.

Entropy minimization on target predictions is another widely used adaptation mechanism,

$$\mathcal{L}_{\mathrm{ent}} = \frac{1}{N_t}\sum_{j=1}^{N_t} H\left(p_\theta(\cdot \mid x_j^t)\right). \tag{29}$$

Minimizing this encourages confident target predictions, effectively pushing features away from uncertain regions. Yet under strong shift or open-set conditions, the model may become confidently wrong. For automatic modulation recognition, where unseen waveform parameterizations or previously absent impairments are common, entropy minimization should be combined with confidence thresholding, prototype consistency, or uncertainty-aware pseudo-labeling rather than applied blindly.

A more physically grounded adaptation strategy is self-supervised synchronization-aware pretext learning. Because many target-domain differences arise from synchronization mismatch and front-end variability, the encoder can be adapted to auxiliary tasks such as phase offset estimation, coarse frequency offset regression, time-shift discrimination, or reconstruction under known perturbations. These tasks do not require modulation labels but can force the representation to attend to the structures that govern domain shift. Once the encoder has reoriented to the target signal statistics, a smaller amount of supervised fine-tuning may suffice.

Continual deployment raises further issues. In a live system, the environment can drift gradually as oscillators age, traffic patterns change, or hardware is replaced. A static model may slowly degrade even if it transferred well initially. Continual adaptation must then balance plasticity and stability. One formalization is online risk minimization with a memory of rep-

resentative source or past target samples:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \left[\mathcal{L}_{\text{new}}(\theta_t) + \beta \mathcal{L}_{\text{mem}}(\theta_t) \right]. \tag{30}$$

The memory term preserves previously learned structure. For AMR, a replay buffer stratified by environment and class is often more useful than a purely random memory, because forgetting can be highly condition specific.

Open-set and partial-set conditions must also be considered [22]. Deployment may contain modulation types absent from the source label space or class priors drastically different from those seen in training. A model that assumes closed-world classification can misclassify unknown waveforms with high confidence. Energy-based scores, distance-to-prototype metrics, and Bayesian posterior concentration can help detect such events. Suppose the penultimate features for class k are modeled by a prototype μ_k. A simple reject score is

$$s_{\text{rej}}(x) = \min_k \left| 0\phi(x) - \mu_k \right| 0_2^2. \tag{31}$$

Large values indicate poor fit to known classes. Although simplistic, prototype distance often reveals when the representation itself no longer supports the label set. More refined versions can model class covariance or use density estimators on latent space.

Another transfer issue is calibration shift. A classifier may retain ranking quality under target shift while its posterior probabilities become miscalibrated. This matters if downstream decisions depend on confidence thresholds, sensor fusion, or selective abstention. One can adapt calibration using a small labeled target set or unsupervised temperature heuristics based on entropy statistics. More principled approaches treat calibration as a separate transfer problem, learning a mapping from logits and environment indicators to probabilities. In communication systems where signal processing modules exchange uncertainty information, this can materially affect system-level performance.

Few-shot transfer is particularly relevant for specialized emitters or newly observed hardware platforms. Meta-learning methods attempt to prepare a source model that adapts quickly with very few labeled target examples [23]. In principle this suits AMR because new devices may arrive with scarce annotations. In practice, success depends on the meta-training task distribution. If episodes do not reflect realistic target shifts, the learned adaptation rule may again overfit to synthetic patterns. A more reliable approach may combine broad source-domain invariance training with modest parameter-efficient fine-tuning and a physically meaningful augmentation prior.

The wider lesson is that transfer should not be evaluated merely by comparing full retraining and naive fine-tuning. What matters is how the source representation interacts with target-domain structure, how much labeled target data is needed to recover performance, how confidence behaves during adaptation, and whether the model can remain reliable as environments drift over time. Generalization in automatic modulation recognition is therefore inseparable from adaptation policy. A system that cannot detect and respond to environment change remains fragile no matter how strong its source benchmark accuracy appears.

7. Evaluation Protocols for Generalization-Centric Automatic Modulation Recognition

The validity of any claim about generalization depends on the evaluation protocol. In automatic modulation recognition, many reported accuracies are measured under random train-test splits drawn from the same data generator, using the same waveform parameter ranges, sometimes even the same transmitter or channel identities. Such protocols quantify interpolation within a narrow source domain, not robustness to distribution shift. A more informative evaluation strategy must treat environment as a first-class variable and deliberately separate source and target conditions.

The simplest improvement is environment-disjoint partitioning. If each example is associated with an environment descriptor e, then the split should ensure

$$\mathcal{E}_{\text{train}} \cap \mathcal{E}_{\text{test}} = \varnothing \tag{32}$$

for at least one meaningful component of the environment. This could mean holding out carrier offset ranges, channel models, pulse-shape roll-off values, transmitter identities, receiver front ends, bandwidths, or segmentation rules [24]. The specific held-out variable should be chosen based on the intended deployment scenario. If the receiver must operate across unseen hardware platforms, transmitter or receiver identity should define the split. If it must withstand changing propagation, channel families should be held out. Random segmentation alone is insufficient.

Signal-to-noise ratio deserves nuanced treatment. Evaluating accuracy versus SNR is useful, but it can be deceptive if all other variables are shared across train and test. A model may appear robust at low SNR because it has seen the exact same generator configurations with only different noise realizations. Stronger protocols include leave-one-SNR-band-out training, mixed-SNR training with unseen channel combinations, and cross-SNR calibration analysis. Since SNR also changes the intrinsic Bayes difficulty of the task, one should report not only accuracy but confidence behavior and rejection performance as SNR varies.

A generalization-oriented benchmark should measure more than top-1 accuracy. Confusion matrices stratified by environment reveal whether the model fails uniformly or collapses on specific class pairs under certain conditions. For example, two modulations

may be separable at one pulse shape and nearly indistinguishable at another. Such structure is invisible in a single aggregated score. Latent-space diagnostics are also valuable. If class clusters fragment by environment in the representation space, the model is unlikely to transfer well even if the current benchmark accuracy is high.

Calibration metrics are indispensable [25]. Expected calibration error, negative log-likelihood, Brier score, and selective risk all complement accuracy. A receiver that occasionally abstains on genuinely ambiguous signals may be preferable to one that always predicts. Selective classification can be formalized through a confidence threshold τ. Let $c(x)$ be a confidence score and define the accepted set $\mathcal{A}_\tau = \{x : c(x) \geq \tau\}$. The selective risk is

$$R_{\mathrm{sel}}(\tau) = \frac{\mathbb{E}\left[\ell(f(x), y)\, \mathbf{1}\{x \in \mathcal{A}_\tau\}\right]}{\mathbb{P}(x \in \mathcal{A}_\tau)}. \tag{33}$$

Reporting this as a function of coverage is more operationally meaningful than raw accuracy alone. In radio systems, abstention can trigger additional sensing time, higher-layer fusion, or specialized demodulators.

Open-set evaluation should also be standard. The test set ought to include waveforms from modulation families not seen during training, as well as severe distortions that move known classes off the source manifold. Metrics such as area under the receiver operating characteristic for unknown detection, false acceptance of unknowns, and calibration under open-set exposure help determine whether the classifier can recognize its limits. Without such tests, a model may look strong while remaining hazardous in deployment.

Cross-simulator and cross-hardware evaluation are especially revealing. If a model is trained on one synthetic generator and tested on another with nominally equivalent modulation definitions but different implementation details, performance often drops sharply. This indicates sensitivity not to modulation itself but to generator artifacts. Similarly, cross-hardware tests can expose hidden dependence on front-end transfer functions, sampling clocks, and device-specific distortions. These evaluations should be regarded as core rather than optional, because real deployment almost always mixes acquisition pipelines.

Sample-efficiency curves offer another perspective on generalization [26]. Instead of reporting only the final performance after full target adaptation, one should measure accuracy as a function of labeled target samples. A transferable model is one whose performance rises quickly with limited supervision. Let m denote the number of labeled target examples per class or per environment. The adaptation profile

$$A(m) = \mathrm{Acc}\left(f_{\mathrm{adapt}}^{(m)}\right) \tag{34}$$

characterizes how useful the source representation is. Two models with similar zero-shot performance can differ greatly in $A(m)$, reflecting very different latent geometries.

Evaluation should also distinguish between class-conditional robustness and prior robustness. A model might remain accurate if target class priors match the source priors but fail under class imbalance. This matters in spectrum monitoring, where some modulations dominate certain bands or time periods. Reporting per-class recall, macro-averaged measures, and confusion stability under prior shift provides a fuller picture than micro-averaged accuracy.

Statistical significance matters as well. Radio benchmarks often involve stochastic training, synthetic random seeds, and limited target adaptation samples. Single-run reporting can exaggerate gains or hide instability. Confidence intervals over multiple seeds, environment resamplings, and adaptation trials are therefore essential. In transfer experiments especially, the variance across runs may be as important as the mean, because high variance indicates unreliable adaptation behavior.

Finally, evaluation should include computational and memory constraints [27]. A generalization technique that requires expensive adaptation, large ensembles, or impractical buffering may be unsuitable for embedded receivers even if it improves average accuracy. The relevant metrics include parameter count, inference latency, energy per sample, and adaptation overhead. Robustness should be interpreted jointly with these costs. A compact model with slightly lower source accuracy but stronger cross-environment stability may offer superior system-level value.

In summary, a protocol capable of supporting claims about generalization in automatic modulation recognition must separate training and testing by meaningful environment variables, include calibration and open-set analysis, quantify adaptation efficiency, expose cross-generator and cross-hardware transfer, and report statistical variability alongside computational cost. Without these elements, high benchmark accuracy can be a misleading proxy for the reliability that operational systems actually require.

8. Implications for System Design and Open Research Directions

The preceding analysis implies that generalization in automatic modulation recognition is best approached as a co-design problem between signal modeling, representation learning, adaptation policy, and evaluation. This perspective has several consequences for system design. First, the receiver should not treat the classifier as an isolated black box downstream of fixed preprocessing. Synchronization quality, front-end normalization, segmentation policy, and uncertainty thresholds all influence the statistics presented to the network. A modest improvement in upstream alignment can reduce the invariance burden imposed on the model. Con-

versely, a representation explicitly trained to tolerate segmentation and phase uncertainty can reduce the need for expensive synchronization. The design choice depends on where computational budget and reliability constraints lie.

Second, the most useful inductive biases are likely to be those that express partial rather than absolute invariance. A modulation classifier should be stable to mild carrier phase drift, bounded frequency offset, gain changes, and timing jitter, but not necessarily to arbitrary waveform deformation. This suggests parameterized invariance modules linked to physically interpretable nuisance variables [28]. Instead of relying only on generic augmentation, future architectures may benefit from learned latent synchronization layers, differentiable resampling blocks, or branch-specific operators targeted to known distortion families. Such components would not replace deep representation learning; they would shape it toward the structure of radio observations.

Third, target-domain adaptation should be built into the deployment loop. An operational receiver often has access to unlabeled target streams, occasional labels from external metadata, or higher-layer consistency checks. These signals can be exploited to maintain calibration and detect domain drift. A static model frozen after source training is rarely the optimal design. The challenge is to build adaptation mechanisms that are conservative enough to avoid catastrophic forgetting yet flexible enough to track changes in hardware and propagation. Lightweight parameter updates, environment-conditioned normalization, and prototype maintenance in latent space are promising directions because they can be implemented under practical resource limits.

Fourth, representation diagnostics should become standard engineering tools. Visualizing latent clusters by environment, measuring within-class dispersion across held-out conditions, and tracking prototype drift across devices can reveal failure modes before deployment. Such analyses are not mere academic embellishments. They identify whether the model is organizing signals by modulation or by simulator artifact, by transmitter identity or by actual waveform family. In a field where synthetic data is common, such checks are essential to prevent misleading conclusions.

Several open research directions follow naturally. One concerns the definition of physically meaningful transformation groups for AMR. Current augmentation practice is often heuristic, yet the success of invariance learning depends on choosing transformations that preserve class semantics while spanning realistic deployment variation [29]. A formal catalog of such transformations, stratified by task and receiver capability, would clarify both training and evaluation. Another open direction is uncertainty modeling under observability limits. When low SNR, heavy interference, or severe channel distortion makes classes intrinsically hard to distinguish, confidence should reflect that fact. Bayesian approximations, evidential methods, and conformal prediction adapted to radio data remain underexplored.

Another promising area is the interaction between self-supervised learning and modulation recognition. Large unlabeled radio corpora are easier to collect than exhaustively labeled datasets. Pretext tasks based on synchronization, spectrum occupancy transitions, channel consistency, or contrastive views of the same burst could produce encoders with stronger cross-environment stability than purely supervised training. The key challenge is selecting pretext objectives whose invariances align with downstream modulation discrimination rather than with irrelevant acquisition details.

Open-set recognition and hierarchical labeling also deserve more attention. Realistic monitoring systems often need to distinguish known modulation families, unknown but structured signals, and noncommunication interference. A flat closed-set classifier is too restrictive for this purpose. Hierarchical outputs that separate broad waveform families before finer modulation distinctions may yield better calibration and easier adaptation. They may also reduce the penalty of label mismatch when new signal types appear.

Cross-modal and multi-sensor integration is another underdeveloped opportunity. Automatic modulation recognition is usually framed as a single-stream task, but operational systems may have access to coarse metadata such as band allocation, geolocation priors, antenna array measurements, or temporal persistence across bursts. Integrating such cues can reduce uncertainty under shift, provided the model avoids overreliance on context that may itself drift [30]. Designing fusion architectures that remain robust when some modalities are missing or corrupted is therefore a substantial research problem.

Finally, there is a need for benchmarks that genuinely reflect deployment diversity. Many existing datasets are invaluable for method development, but future progress will require corpora that span multiple generators, multiple hardware chains, richer impairment models, and standardized environment-disjoint splits. Such benchmarks should encourage reporting of calibration, open-set behavior, and adaptation efficiency, not only accuracy. Once evaluation changes, architectural progress is likely to become more meaningful because it will be optimized for the problem that practitioners actually face.

The deeper message is that generalization is not an add-on metric attached after model training. It is the property that determines whether automatic modulation recognition systems remain useful outside the narrow conditions under which they were created. For that reason, the field benefits from a shift in emphasis: from maximizing performance on homogeneous

datasets toward constructing representations and protocols whose stability can be justified in terms of waveform physics, environment variation, and operational adaptation.

9. Conclusion

Generalization in automatic modulation recognition is fundamentally a problem of learning class-discriminative structure while controlling sensitivity to the many environment variables that reshape observed waveforms. The relevant shifts arise from concrete physical mechanisms, including synchronization error, channel variability, pulse-shape differences, front-end distortion, segmentation policy, and simulator or hardware mismatch. Because of this, benchmark accuracy under random splits is not a sufficient indicator of reliability.

A more faithful treatment views each signal as the product of a modulation process composed with an environment-dependent observation operator. Under this view, robust recognition requires agreement between several design choices: realistic source and target environment modeling, architectures that support selective invariance, optimization objectives that penalize shortcut learning and excessive sensitivity, adaptation procedures that preserve stable features while responding to drift, and evaluation protocols that partition data by meaningful environments rather than by arbitrary sample indices. Calibration, abstention, open-set detection, and adaptation efficiency are then as important as raw closed-set accuracy.

The technical analysis developed here suggests that improvements in generalization are unlikely to come from any single modification in isolation. Instead, progress depends on integrating communication-theoretic knowledge with statistical learning principles so that invariance, confidence, and transfer behavior are shaped by the actual physics of waveform acquisition. As automatic modulation recognition moves toward realistic deployment, models will be judged less by their ability to memorize a generator and more by their capacity to remain stable when the same modulation class is realized through new channels, devices, and operating conditions [31].

References

[1] E. Baena, S. Fortes, O. Alay, M. Xie, H. Lonsethagen, and R. Barco, "Cellular network radio monitoring and management through virtual ue probes: A study case based on crowded events.," Sensors (Basel, Switzerland), vol. 21, pp. 3404–, 5 2021.

[2] F. Wang, Y. Zhou, H. Yan, and R. Luo, "Enhancing the generalization ability of deep learning model for radio signal modulation recognition," Applied Intelligence, vol. 53, no. 15, pp. 18758–18774, 2023.

[3] J. P. Garcia-Martin and A. Torralba, "Model of a device-level combined wireless network based on nb-iot and ieee 802.15.4 standards for low-power applications in a diverse iot framework.," Sensors (Basel, Switzerland), vol. 21, pp. 3718–, 5 2021.

[4] C. Paoloni, R. Basu, L. R. Billa, J. M. Rao, R. Letizia, Q. Ni, E. Wasige, A. Al-Khalidi, J. Wang, and R. Morariu, "Long-range millimetre wave wireless links enabled by travelling wave tubes and resonant tunnelling diodes," IET Microwaves, Antennas & Propagation, vol. 14, pp. 2110–2114, 11 2020.

[5] S. Jeon, L. Youjin, D. Ryu, Y. K. Cho, Y. Lee, S. B. Jun, and C. H. Ji, "Implantable optrode array for optogenetic modulation and electrical neural recording.," Micromachines, vol. 12, pp. 725–, 6 2021.

[6] P. Kurdel, M. Češkovič, N. Gecejová, J. Labun, and J. Gamec, "The method of evaluation of radio altimeter methodological error in laboratory environment.," Sensors (Basel, Switzerland), vol. 22, pp. 5394–5394, 7 2022.

[7] A. Genner, P. Martín-Mateos, H. Moser, and B. Lendl, "A quantum cascade laser-based multi-gas sensor for ambient air monitoring," Sensors (Basel, Switzerland), vol. 20, pp. 1850–, 3 2020.

[8] A. R. Uppala, V. Narasimhulu, and S. Prasad, "Improved convolutional neural network based cooperative spectrum sensing for cognitive radio," KSII Transactions on Internet and Information Systems, vol. 15, 6 2021.

[9] A. Masiukiewicz, "Throughput comparison between the new hew 802.11ax standard and 802.11n/ac standards in selected distance windows," International Journal of Electronics and Telecommunications, pp. 79–84, 2 2019.

[10] A. Lipovac, V. Lipovac, and B. Modlic, "Phy, mac, and rlc layer based estimation of optimal cyclic prefix length," Sensors (Basel, Switzerland), vol. 21, pp. 4796–, 7 2021.

[11] L. Ruiz-Ortega, G. Schitter, and P. Mesquida, "Electrostatic read out for label-free assays based on kelvin force principle," Sensing and Imaging, vol. 20, pp. 23–, 5 2019.

[12] Y. Du, G. Zhu, J. Zhang, and K. Huang, "Automatic recognition of space-time constellations by learning on the grassmann manifold," IEEE Transactions on Signal Processing, vol. 66, pp. 6031–6046, 11 2018.

[13] R. R. Gujjula, C. Perumal, P. Kodali, and B. V. Rajanna, "Design and analysis of dual-mode numerically controlled oscillators based controlled oscillator frequency modulation," International Journal of Electrical and Computer Engineering (IJECE), vol. 12, pp. 4935–4935, 10 2022.

[14] Y. Sun and E. A. Ball, "Automatic modulation classification using techniques from image classification," IET Communications, vol. 16, pp. 1303–1314, 1 2022.

[15] S. K. Pattnaik, S. R. Samal, S. Bandopadhaya, K. Swain, S. Choudhury, J. K. Das, A. Mihovska, and V. Poulkov, "Future wireless communication technology towards 6g iot: An application-based analysis of iot in real-time location monitoring of employees inside underground mines by using ble.," Sensors (Basel, Switzerland), vol. 22, pp. 3438–3438, 4 2022.

[16] J. P. Lousado and S. Antunes, "Monitoring and support for elderly people using lora communication technologies: Iot concepts and applications," Future Internet, vol. 12, pp. 206–, 11 2020.

[17] Z. Yigit and E. Basar, "Space-time media-based modulation," IEEE Transactions on Signal Processing, vol. 67, pp. 2389–2398, 5 2019.

[18] M. Krondorf, S. Bittner, D. Plettemeier, A. Knopp, and M. Wikelski, "Icarus-very low power satellite-based iot.," Sensors (Basel, Switzerland), vol. 22, pp. 6329–6329, 8 2022.

[19] V. B. Vales, O. C. Fernández, T. Domínguez-Bolaño, C. J. Escudero, and J. A. García-Naya, "Fine time measurement for the internet of things: A practical approach using esp32," IEEE Internet of Things Journal, vol. 9, pp. 18305–18318, 10 2022.

[20] L. Sahraoui, D. Messadeg, S. e Hariz, and N. Doghmane, "Multiuser detection with decision-feedback detectors and

pic in mc-cdma system," TELKOMNIKA (Telecommunication Computing Electronics and Control), vol. 16, pp. 2578–2587, 12 2018.

[21] O. Kursu, M. E. Leinonen, G. Destino, N. Tervo, M. Sonkki, T. Rahkonen, A. Parssinen, S. Tammelin, M. Pettissalo, and A. Korvala, "Design and measurement of a 5g mmw mobile backhaul transceiver at 28 ghz," EURASIP Journal on Wireless Communications and Networking, vol. 2018, pp. 1–11, 8 2018.

[22] V. B. Raskar and S. L. Lahudkar, "Performance comparison of tr and fsruwb system using particle filter: Effects of frequency, data rate, multi-path and multi-channel communication," International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, pp. 269–280, 12 2022.

[23] R. Sabella, P. Iovanna, G. Bottari, and F. Cavaliere, "Optical transport for industry 4.0 [invited]," Journal of Optical Communications and Networking, vol. 12, pp. 264–276, 7 2020.

[24] N. Polyakov and A. Platonova, "Assessing latency of packet delivery in the 5g 3gpp integrated access and backhaul architecture with half-duplex constraints," Future Internet, vol. 14, pp. 345–345, 11 2022.

[25] I. A. Gepko, "Orthogonal ccsk based on complementary code sequences," Radioelectronics and Communications Systems, vol. 65, pp. 61–71, 9 2022.

[26] G. Castellanos, M. Deruyck, L. Martens, and W. Joseph, "Performance evaluation of direct-link backhaul for uav-aided emergency networks," Sensors (Basel, Switzerland), vol. 19, pp. 3342–, 7 2019.

[27] J. Nagai, K. Ishibashi, and Y. Yamao, "Signal quality improvement in downlink power domain noma with blind nonlinear compensator and frequency domain equalizer," IEICE Transactions on Communications, vol. E105.B, pp. 648–656, 5 2022.

[28] D. Pineda and C. Hernández, "Cognitive radio for tvws usage," TELKOMNIKA (Telecommunication Computing Electronics and Control), vol. 17, pp. 2735–2746, 12 2019.

[29] A. M. Evans, T. A.-M. Quazi, and H. Xu, "Ber performance of an mrt-hqam system over rayleigh fading channels with channel estimation errors and feedback delay," IET Communications, vol. 12, pp. 2150–2157, 9 2018.

[30] F. Firdaus, N. A. Ahmad, and S. Sahibuddin, "Fingerprint indoor positioning based on user orientations and minimum computation time," TELKOMNIKA (Telecommunication Computing Electronics and Control), vol. 17, pp. 1740–1749, 8 2019.

[31] C. Feng and T. Schneider, "Benefits of spectral property engineering in distributed brillouin fiber sensing," Sensors (Basel, Switzerland), vol. 21, pp. 1881–, 3 2021.