International Journal of Applied Sciences, Research, and Innovation

OPEN ACCESS

FULL PAPER | IJASRI | OPEN

Robust Deep Architectures for Low-SNR Automatic Modulation Recognition in Nonstationary Wireless Channels Using Residual Channel-Attentive Feature Fusion

Ahmed Elgendy[1], **Mostafa Khalil**[2]

[1]Department of Electrical Engineering, Aswan University, Sahary Road, Aswan 81528, Egypt

[2]Department of Electrical Power and Machines Engineering, South Valley University, El-Gomhouria Street, Qena 83523, Egypt

Abstract

Wireless receivers increasingly operate in congested, fading, and interference-rich environments in which the received signal energy may approach or even fall below the effective noise floor. In such settings, automatic modulation recognition is no longer a simple pattern classification problem, because the observable waveform is jointly distorted by channel filtering, carrier frequency offset, phase uncertainty, timing mismatch, and nonstationary noise. Classical likelihood-based and handcrafted-feature pipelines remain informative, but their performance often degrades when the assumed signal model is violated or when discriminative structures are weakly expressed in short observation windows. This paper develops a technical framework for low-SNR automatic modulation recognition that combines signal-domain modeling with a residual channel-attentive deep architecture designed to preserve shallow waveform evidence while progressively refining class-discriminative latent features. The proposed formulation treats modulation recognition as inference over short in-phase and quadrature sequences and emphasizes the preservation of local envelope, phase-transition, and short-memory temporal signatures that are frequently lost in purely deep hierarchical compression. A grouped residual backbone is therefore augmented with cross-stage feature fusion and channel-reweighting operations that selectively amplify informative latent responses under adverse noise conditions. The paper further analyzes optimization under SNR uncertainty, calibration of posterior outputs, and the interaction between architectural depth, receptive field growth, and feature separability. The resulting perspective is not restricted to a single dataset or benchmark protocol; instead, it provides a general, mathematically grounded view of why residual aggregation, feature reuse, and adaptive channel emphasis are particularly relevant for low-SNR recognition, where class manifolds overlap and inter-class confusion is driven more by nuisance variability than by nominal constellation geometry.

1. Introduction

Automatic modulation recognition has long occupied an intermediate position between classical signal processing and modern machine learnings. It is a receiver-side inference task in which the goal is to determine the modulation format of a received waveform without relying on full demodulation, pilot knowledge, or protocol-level metadata. In noncooperative settings this capability supports spectrum awareness, adaptive interception, cognitive radio control, interference diagnosis, signal intelligence, and front-end configuration for downstream decoding or decision-making. The practical relevance of the problem has increased rather than diminished in recent years because wireless environments have become denser, waveform families more diverse, and hardware operating conditions more variable. Narrowband and wideband emitters coexist; analog and digital schemes can be encountered in the same band; and low-cost front ends introduce imperfections that interact with fading and non-Gaussian disturbances. Under these conditions the received complex baseband sequence often reflects not only the nominal modulation class but also multiple nuisance effects whose aggregate energy can obscure class structure.

The low-SNR regime is especially challenging because the standard intuition that modulation classes occupy distinct geometric regions in feature space becomes unreliable when the observable samples are dominated by stochastic perturbations. At moderate or high SNR, the constellation of a phase-shift keyed signal, the envelope variability of amplitude modulation, or the symbol transition statistics of continuous-phase frequency shift keying retain enough contrast for handcrafted descriptors or shallow classifiers to separate them. As the SNR decreases, however, those class-specific structures become progressively entangled with estimation error, channel coloration, synchronization mismatch, and finite-sample randomness. The resulting overlap is not merely a matter of additive dispersion around clean class templates. It also reflects the fact that a short observation window may emphasize different latent mechanisms across examples from the same class. One realization may preserve short-term phase continuity, another may exhibit envelope bursts after multipath superposition, and a third may be dominated by front-end offset. A robust recognizer in this regime must therefore avoid premature compression of the waveform and must reuse low-

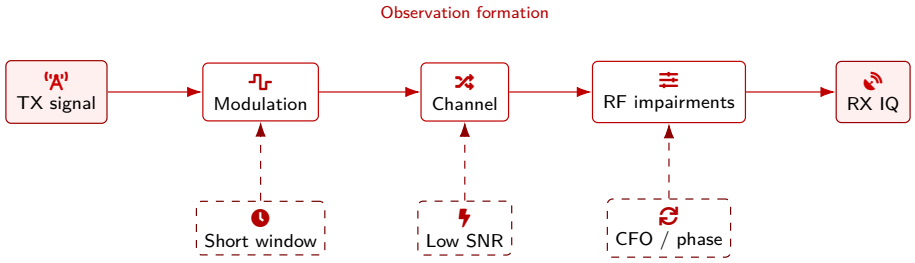


Figure 1. Signal formation chain for low-SNR modulation recognition. The received IQ window is shaped jointly by the nominal modulation, channel response, RF impairments, and weak observation conditions, motivating architectures that preserve fragile local evidence while remaining robust to nuisance distortion.

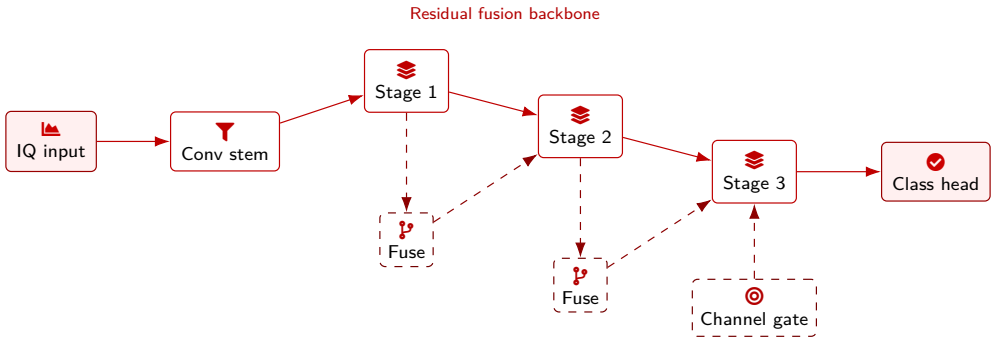


Figure 2. Overall residual channel-attentive AMR pipeline. A shallow stem lifts the IQ window into a latent space, successive grouped residual stages refine multiscale waveform features, cross-stage fusion restores useful early information, and channel gating emphasizes reliable latent responses before final modulation classification.

level evidence even after deeper abstractions are formed [6].

Historically, automatic modulation recognition was addressed through decision-theoretic and feature-engineered pipelines. Decision-theoretic methods start from probabilistic models of candidate modulations and derive test statistics based on likelihood functions, generalized likelihood ratios, or asymptotic approximations. These methods are analytically elegant and, under matched assumptions, can be close to optimal. Their practical limitations arise from the burden of nuisance parameter estimation, model mismatch, and computational scaling as the candidate set expands or the channel becomes more realistic. Feature-engineered approaches attempt to bypass exact likelihood evaluation by extracting descriptors such as higher-order cumulants, instantaneous amplitude and phase statistics, cyclostationary signatures, wavelet coefficients, or time-frequency textures, followed by a conventional classifier. Such methods reduce some modeling burden but shift the problem to handcrafted representation design. In high-noise or short-window settings the extracted features may become unstable, mutually redundant, or insufficiently expressive for classes with subtle differences.

Deep learning became attractive in this domain because it can learn directly from raw or minimally processed in-phase and quadrature sequences, thereby reducing the dependence on rigid front-end assumptions. Convolutional architectures, recurrent models, and hybrid networks demonstrated that latent features can be learned from waveform observations without explicit derivation of expert descriptors. Yet the low-SNR regime exposes weaknesses in naive applications of deep learning as well. If the network is made very deep, early details may be discarded before they can be reused by later layers. If the model is made very wide, parameter growth can exceed the information content of the training data and lead to overfitting, calibration degradation, or sensitivity to SNR imbalance [16]. If pooled representations dominate the architecture, the network may learn to emphasize coarse structures while suppressing weak but decisive local signatures such as phase increment consistency, envelope intermittency, or transient spectral asymmetry. These issues suggest that low-SNR modulation recognition is not simply a matter of increasing capacity; it is a matter of organizing capacity so that the model remains sensitive to weak, distributed, and scale-dependent evidence.

A useful design principle is that the recognizer should retain multiple levels of abstraction simultaneously. Shallow layers are often better at preserving waveform primitives, including short-run correlations, local oscillation rates, and abrupt energy changes. Intermediate layers can integrate those primitives over larger receptive fields and encode transition statistics, local stationarity, and cross-channel interactions between real and imaginary components. Deep layers are better suited to aggregating class-level evidence and suppressing sample-specific noise realizations. However, if the network only passes information forward in a strictly sequential manner, then one level of abstraction overwrites another. Residual connections partially address this by preserving identity information and stabilizing optimization, but ordinary residual designs do not automatically guarantee that the most informative shallow features remain explicitly available to later stages under severe noise. Cross-stage fusion becomes relevant because it permits the classifier to examine both local

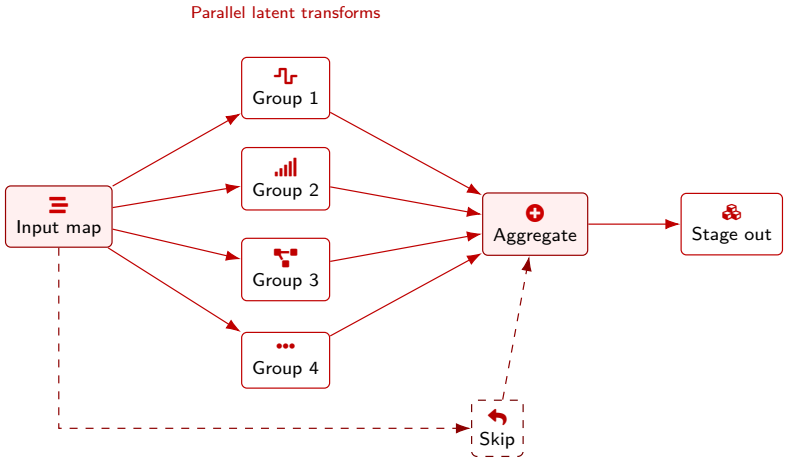


Figure 3. Grouped residual block used inside each stage. The same incoming feature tensor is processed by several structured branches that capture different short-memory waveform phenomena, after which the branches are aggregated and combined with a residual skip. This preserves optimization stability while allowing heterogeneous latent cues to coexist under noisy observation conditions.

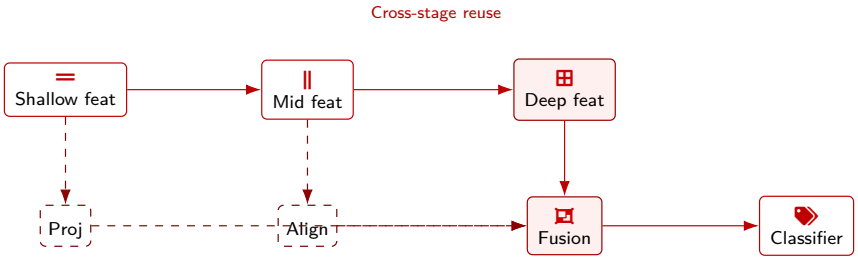


Figure 4. Cross-stage feature fusion mechanism. Earlier representations retain high-resolution local evidence, whereas deeper features encode broader class structure; projecting and aligning them into a common fusion space lets the classifier use both simultaneously. This is particularly helpful when decisive low-SNR cues survive only in shallow layers and would otherwise be discarded by sequential compression.

evidence and abstract summaries at once.

Another principle is selective emphasis. Not every latent channel learned by a deep architecture contributes equally to the discrimination of modulation classes in a given noise condition. Some channels may encode robust, class-aligned responses, whereas others may amplify nuisance structures, front-end artifacts, or dataset-specific biases [7]. A channel-attentive mechanism that estimates the relative utility of latent responses can be interpreted as a soft reliability allocator. Instead of forcing every feature map to contribute equally, the network learns to amplify channels whose activations remain informative across noisy observations while suppressing channels that respond indiscriminately. This form of adaptive reweighting is especially attractive in low-SNR recognition because the useful evidence is often sparse in the latent domain: only a subset of responses may preserve class information after corruption.

The technical objective of this paper is to formulate a research-oriented treatment of low-SNR automatic modulation recognition that integrates signal modeling, architectural design, and optimization analysis into a unified narrative. The paper does not treat neural architecture as a black box. Rather, it asks which structures are appropriate when the observation window is short, the SNR is poor, and the discriminative content of the signal is distributed across multiple scales. The answer developed here is a residual channel-attentive architecture with grouped transformations and explicit feature fusion across stages, together with an SNR-aware training objective and a careful interpretation of inter-class confusion. The guiding premise is that low-SNR recognition requires both preservation and suppression: preservation of weak but class-relevant waveform evidence, and suppression of latent responses that are dominated by nuisance variability.

The remainder of the paper elaborates this premise in a mathematically grounded manner. First, the signal and observation model is formalized so that the low-SNR recognition task is posed as inference under nuisance uncertainty. Next, the geometry of low-SNR feature learning is analyzed to explain why class overlap and representation collapse occur. A residual channel-attentive architecture is then developed to preserve shallow and deep information jointly [17]. The discussion proceeds to optimization and generalization under noise uncertainty, followed by an evaluation protocol that clarifies how such models should be judged beyond a single top-line accuracy number. The final sections interpret representative error patterns and consider computational deployment issues. Throughout, the tone is deliberately technical and neutral,

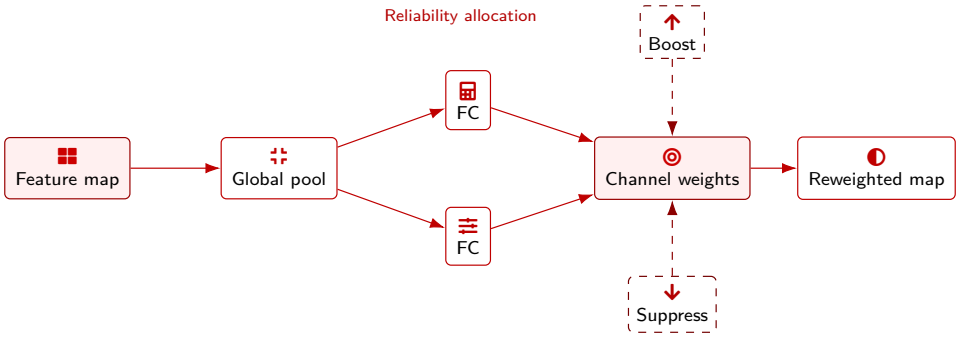


Figure 5. Channel-attentive reweighting block. A pooled descriptor summarizes the latent tensor, two compact fully connected transforms estimate channel importance, and the resulting gate amplifies robust discriminative responses while reducing nuisance-dominated channels. The mechanism acts as a soft reliability selector when only a subset of latent responses remains informative at low SNR.

Table 1. Signal Model Components and Effects

Component	Symbol	Description	Impact
Carrier Offset	Δf	Frequency mismatch	Phase drift
Phase Offset	ϕ_0	Initial phase	Rotation ambiguity
Channel Taps	h_ℓ	Multipath gains	Signal distortion
Noise	$\nu[n]$	Additive disturbance	SNR degradation

emphasizing mechanisms, limitations, and design trade-offs rather than broad claims.

2. Signal and Observation Model

Let the candidate modulation set be denoted by $\mathcal{M} = \{m_1, \ldots, m_K\}$, where each class corresponds to a distinct modulation family or format. A transmitted continuous-time complex envelope $s_m(t)$ associated with class $m \in \mathcal{M}$ can be written as a pulse-shaped symbol sequence, a continuous-phase waveform, or an analog amplitude or frequency modulation process, depending on the class. For a broad and practically relevant model, it is convenient to treat the received baseband signal as a nuisance-corrupted transformation of the nominal transmitted waveform. After carrier downconversion and sampling, the received sequence can be represented as

$$r[n] = e^{j(2\pi \Delta f n T + \phi_0)} \sum_{\ell=0}^{L-1} h_\ell s[n - d_\ell] + \nu[n] \tag{1}$$

where T is the sampling interval, Δf is residual carrier frequency offset, ϕ_0 is an unknown phase, $\{h_\ell\}$ and $\{d_\ell\}$ describe a finite impulse response channel with delay spread, and $\nu[n]$ denotes additive disturbance. The disturbance term is often approximated as circular complex Gaussian noise in benchmark analysis, but in practical receivers it may include colored interference, quantization residue, front-end imbalance, or bursty contamination. The classifier observes a finite window $x = [r[0], \ldots, r[N-1]]$ or, equivalently, a real-valued tensor formed by its in-phase and quadrature components.

The central difficulty is that the classification variable m is not observed directly, and the nuisance parameters are rarely known. A Bayes-optimal recognizer would compare posterior probabilities

$$p(m \mid x) \propto p(x \mid m) p(m)$$
$$p(x \mid m) = \int p(x \mid m, \eta) p(\eta \mid m)\, d\eta \tag{2}$$

where η collects frequency offset, phase, channel coefficients, timing uncertainty, and other latent variables. The integral is typically intractable or highly sensitive to modeling assumptions. In low-SNR settings this sensitivity becomes more severe because the likelihood surface can be relatively flat with respect to the class index when nuisance uncertainty dominates the signal energy. An effective recognizer must therefore approximate a discriminant that is robust to transformations induced by η while remaining sensitive to the limited residual structure that still encodes the modulation class [24].

A useful quantity is the window-level SNR, denoted by γ, which can be expressed in a simplified model as the ratio between expected signal energy and disturbance energy over the observation interval. If $x = s + \nu$ under an additive approximation, then one may write

$$\gamma = \frac{\mathbb{E}\|s\|_2^2}{\mathbb{E}\|\nu\|_2^2} \tag{3}$$

with the understanding that channel scaling and pulse shaping alter the effective signal term. At very small γ, the posterior entropy $H(M \mid X)$ rises because the observed sequence contains only weak class evidence. However, the degradation is not uniform across classes. Some modulations are inherently more robust to short-window classification because they produce distinctive local invariants even when their global structure is obscured. Continuous-phase schemes may preserve short-term phase continuity, amplitude-modulated ana-

Table 2. Challenges in Low-SNR Modulation Recognition

Challenge	Cause	Effect	Difficulty Level
Noise Dominance	Low γ	Feature masking	High
Channel Variability	Multipath	Pattern distortion	High
Short Windows	Limited samples	Weak statistics	Medium
Model Mismatch	Assumptions fail	Poor generalization	High

Table 3. Feature Types Across Network Depth

Layer Depth	Feature Type	Scale	Role
Shallow	Waveform primitives	Local	Preserve detail
Intermediate	Transition patterns		Context aggregation
Deep	Class abstractions	Global	Decision making
Fusion	Mixed features	Multi-scale	Robust inference

log waveforms may preserve envelope asymmetry, and higher-order quadrature amplitude modulations may collapse rapidly into diffuse clouds. Consequently, low-SNR recognition is not governed solely by average noise power but also by how noise interacts with class-specific invariants.

The in-phase and quadrature decomposition offers a compact representation of the observation:

$$\begin{aligned} x_{\mathrm{I}}[n] &= \Re\{r[n]\} \\ x_{\mathrm{Q}}[n] &= \Im\{r[n]\} \end{aligned} \tag{4}$$

and the two-channel tensor $X \in \mathbb{R}^{2\times N}$ is often used as network input. This representation is expressive because it retains amplitude, phase, and correlation information implicitly. Yet the discriminative cues are not uniformly distributed across time or channels. For instance, frequency offsets induce long-range phase drift, multipath can create local envelope distortions, and narrow observation windows may capture only a few symbol transitions [26]. Therefore the classifier should not assume that a single temporal scale or a single set of fixed summary statistics is sufficient.

Beyond additive noise, the channel itself can obscure class structure by mixing local patterns. Consider the local covariance

$$R_x[\tau] = \mathbb{E}\big[r[n]r^*[n-\tau]\big] \tag{5}$$

which in principle contains information about spectral occupancy, pulse shaping, and cyclostationary behavior. Under low SNR, finite-sample estimates of $R_x[\tau]$ are noisy, and channel filtering can reshape the correlation profile in a class-dependent but inconsistent manner. Similarly, higher-order cumulants that are often useful in classical recognition become statistically unstable for short windows and low energy. These observations motivate a learned representation that can discover which local and intermediate-order structures remain stable enough to support classification.

One may formalize the recognition problem as learning a parametric function $f_\theta : \mathbb{R}^{2\times N} \to \Delta^{K-1}$, where Δ^{K-1} denotes the probability simplex over K classes. The predicted label is

$$\hat{m}(x) = \arg\max_{m\in\mathcal{M}} f_\theta^{(m)}(x) \tag{6}$$

and the training objective seeks a parameter vector θ minimizing expected risk under a data-generating distribution that mixes over classes, channels, and SNR conditions. The complexity of low-SNR recognition arises because the support of this distribution is highly heterogeneous. Two samples from the same nominal class can differ more strongly in local appearance than two samples from different classes at a higher SNR. This heterogeneity encourages architectures that can preserve alternative explanatory cues in parallel rather than commit too early to a single abstraction path [29].

From an information-theoretic viewpoint, the latent representation should compress nuisance variability while preserving mutual information with the class label. Let $Z = \phi_\theta(X)$ be an intermediate representation. The ideal condition is that $I(Z;M)$ remains high even when $I(X;M)$ is low due to noise and uncertainty. In practice, however, deep compression can reduce both nuisance information and task-relevant information simultaneously. This is the central tension in low-SNR recognition: aggressive compression suppresses noise but risks erasing precisely those weak local signatures that distinguish one modulation from another. A carefully designed architecture must therefore use depth, residual reuse, and selective emphasis to navigate this tension.

3. Low-SNR Feature Geometry

The geometry of modulation features under adverse SNR is substantially different from the geometry that emerges in clean constellation diagrams or idealized textbook examples. In high-SNR settings, different modulation classes often form visually separable structures when represented in amplitude-phase, constellation, or spectral domains. Under low SNR, however, these structures contract toward a small number of noisy archetypes. Higher-order quadrature constellations become diffuse and approximately isotropic, narrow distinctions between phase-shift classes are blurred by phase noise and finite-window randomness, and analog waveforms may resemble one another because only coarse energy fluctuations remain observable. The relevant notion of class separability is therefore not clean-cluster distance but the persistence of local invariants under distortion.

A useful way to think about the low-SNR problem is to view each class as a manifold generated by nuisance transformations acting on an underlying waveform family [20]. Noise

Table 4. Architecture Components and Functions

Module	Operation	Purpose	Benefit
Stem	Convolution	Initial feature lift	Local extraction
Residual Block	Skip + transform	Stable learning	Gradient flow
Fusion Path	Cross-stage links	Feature reuse	Detail preservation
Attention	Channel weighting	Feature selection	Noise suppression

Table 5. Grouped Convolution Characteristics

Property	Dense Conv	Grouped Conv	Effect
Connectivity	Full		Modular learning
Parameters	High	Reduced	Efficiency
Diversity	Moderate	High	Multi-pattern capture
Coupling	Strong	Weak	Flexibility

thickens each manifold, and short-window sampling makes the observable patch of the manifold irregular. When SNR falls, the thickened manifolds for several classes intersect strongly in the ambient sample space. A network that attempts to separate the classes using only globally pooled deep features is likely to struggle because the pooled representation averages away sparse but discriminative events. Conversely, a network that uses only local features may remain too sensitive to incidental noise realizations. The most effective strategy is to maintain a hierarchy in which shallow responses preserve local motifs, while deeper responses aggregate those motifs into class-level hypotheses, and fusion mechanisms keep both available to the classifier.

The local motifs relevant for modulation recognition are diverse. Short-term phase increments may distinguish phase-continuous from phase-discontinuous families. Local envelope stationarity can help separate constant-envelope digital modulations from analog amplitude-based waveforms. Temporal smoothness and burst structure can provide indirect evidence about pulse shaping and symbol transitions. Cross-channel interactions between in-phase and quadrature streams contain information about instantaneous frequency and rotational behavior that is obscured if the channels are processed independently. These cues are weak at low SNR, but they are not uniformly absent. Instead, they appear intermittently, which argues for channel-attentive latent processing rather than uniform averaging.

If the network is represented as a sequence of nonlinear maps $z_\ell = \psi_\ell(z_{\ell-1})$, the separability of classes at layer ℓ depends on both the receptive field and the stability of the transformation to noise. A very shallow map may preserve input geometry too closely, leaving classes overlapped [9]. A very deep map may induce representation collapse if its effective invariances are stronger than the distinctions among modulation classes. The relevant architectural question is therefore not whether deeper is better, but at which layers class-specific structure becomes visible and whether those layers can still access the waveform evidence present near the input. Feature fusion across stages can be interpreted as preserving a family of projections onto different tangent approximations of the class manifolds.

Normalization also plays an important role in feature geometry. Sample-wise energy normalization can reduce amplitude variability and improve optimization, yet excessive normalization may suppress envelope information that is itself class-discriminative for certain analog or amplitude-varying digital modulations. Similarly, phase derotation or frequency offset compensation can improve invariance but may impose assumptions that do not transfer across datasets or receiver configurations. A robust deep recognizer should learn partial invariances rather than rely entirely on rigid preprocessing. Grouped convolutions and channel-attentive gating are useful in this respect because they allow multiple latent transformations to coexist, some emphasizing local oscillatory behavior, others emphasizing envelope trends or cross-channel interactions.

The distinction between shallow and deep features can be formalized by examining feature covariance across layers. Let $z_\ell(x) \in \mathbb{R}^{C_\ell \times T_\ell}$ denote a latent tensor. Early layers tend to preserve high temporal resolution and localized activations, while later layers trade temporal detail for semantic abstraction. The covariance of latent channels,

$$\Sigma_\ell = \mathbb{E}\Big[\big(\bar{z}_\ell - \mu_\ell\big)\big(\bar{z}_\ell - \mu_\ell\big)^\top\Big] \tag{7}$$

where $\bar{z}_\ell$ is a pooled channel vector, often reveals that a few directions dominate as depth increases. If those dominant directions correspond to robust class evidence, abstraction is beneficial [1]. If they instead correspond to nuisance-aligned responses that are easy to learn from the training distribution, then the network becomes brittle. Channel attention is an attempt to reshape this covariance adaptively, boosting latent directions that remain predictive and diminishing those that only reflect incidental activation patterns.

In representative low-SNR studies using residual fusion and channel-aware designs on standard IQ benchmarks, the strongest improvements are often observed not at very high SNR but in the intermediate-to-poor regime where the baseline model is uncertain yet not completely information-starved. This pattern is consistent with the geometric interpretation just outlined. When the sample is nearly clean, most reasonable architectures can separate the classes, and extra fusion yields diminishing returns. When the sample is extremely corrupted, no architectural refinement can fully recover absent information. The largest gains appear where weak local evidence still exists and the model's task is to prevent it from being discarded. In that sense, low-SNR feature geometry favors architectures that postpone irreversible

Table 6. Optimization Strategies Under Noise

Method	Mechanism		Outcome
Cross-Entropy	Standard loss	Accuracy	Baseline
SNR Weighting	$\alpha(\gamma)$	Robustness	Low-SNR gain
Label Smoothing	Soft targets	Calibration	Reduced overconfidence
Augmentation	Perturbations	Invariance	Generalization

Table 7. Evaluation Metrics and Insights

Metric	Measures	Strength	Limitation
Accuracy	Correct predictions	Simple	Ignores imbalance
Recall	Class sensitivity	Fairness	Partial view
ECE	Calibration	Reliability	Needs bins
Confusion Matrix	Error patterns	Insightful	Complex

compression and permit selective reuse of early information.

4. Residual Channel-Attentive Architecture

The architecture considered here is designed around three interacting principles: aggregated transformation, cross-stage feature reuse, and channel-selective emphasis. The backbone begins with a shallow convolutional stem that receives the two-channel IQ tensor and lifts it into a moderate-dimensional latent space. The stem should be wide enough to capture multiple local waveform patterns but not so aggressive that it destroys temporal alignment. A small kernel along the time dimension can detect elementary oscillation, envelope, and transition signatures [30]. The subsequent feature extractor is organized into residual stages whose internal transformations are grouped. Grouped transformation is beneficial because it increases the diversity of learned filters without requiring a fully dense parameter expansion. Different groups can specialize in related but distinct latent phenomena, such as local phase evolution, short-memory amplitude patterns, or coupled I/Q oscillations.

Consider an input tensor $z_{\ell-1}$ to stage ℓ. A grouped residual transformation can be expressed abstractly as

$$u_\ell = \sum_{g=1}^{G} \mathcal{T}_{\ell,g}(z_{\ell-1})$$
$$z_\ell = \sigma\big(\mathcal{P}_\ell(z_{\ell-1}) + u_\ell\big) \tag{8}$$

where $\mathcal{T}_{\ell,g}$ is the gth grouped branch, $\mathcal{P}_\ell$ is the residual projection or identity map, and σ denotes the stage nonlinearity together with normalization. The grouped branches share a common input but learn different local transformations. In ordinary residual networks this construction already improves optimization and permits deeper models. For low-SNR modulation recognition, however, the ordinary stage output may still underutilize shallow evidence because only the transformed sum is propagated forward. The architecture is therefore extended with cross-stage fusion that directly reintroduces selected earlier features into later stages.

Let $s_1, s_2, \ldots, s_L$ denote the outputs of successive stages, with decreasing temporal resolution and increasing channel depth. A fusion path maps one or more shallow features into the dimensionality of a deeper stage and combines them through element-wise addition or concatenation followed by projection. If $\mathcal{F}_{a\to b}$ denotes the mapping from stage a to stage b, then a fused deep representation may be written as

$$\tilde{s}_b = s_b + \mathcal{F}_{a\to b}(s_a) \tag{9}$$

for selected pairs (a, b) with $a < b$. In practice, $\mathcal{F}_{a\to b}$ can include 1×1 convolutions, normalization, and temporal alignment operations that adjust channel dimension and sampling resolution. The critical point is not the exact implementation but the semantics of the path: a late stage is allowed to inspect a transformed image of earlier information rather than rely solely on what survived intermediate processing [25]. This is particularly useful when shallow cues such as local envelope fluctuations or short phase-consistency patterns are easily attenuated by repeated nonlinear transformations.

Feature fusion also changes the optimization landscape. Residual paths already facilitate gradient flow, but cross-stage fusion introduces additional short routes through which the loss can influence early filters. In low-SNR tasks this can bias early layers toward preserving information that remains useful after deeper integration. The architecture effectively says that the network will not decide in advance whether a waveform primitive matters; instead, it will keep a route available so that deeper layers can choose to reuse it. This is a more cautious strategy than simple stacking, and such caution is appropriate when the class evidence is weak.

Channel-selective emphasis is then applied to the fused representation. Let v_ℓ denote the latent tensor produced by a grouped residual or fused block. A channel-attentive gating mechanism first summarizes each channel spatially or temporally, then estimates a set of channel weights, and finally rescales the original channels. A compact formulation is

$$q_\ell = \mathcal{G}(v_\ell)$$
$$a_\ell = \sigma_{\text{sig}}\big(W_{\ell,2}\,\rho(W_{\ell,1}q_\ell)\big)$$
$$y_\ell = a_\ell \odot v_\ell \tag{10}$$

where $\mathcal{G}$ is a global pooling operator, ρ is a pointwise activation such as a rectifier, σ_{sig} is a sigmoid gate, and $\odot$ denotes channel-wise multiplication. The learned vector a_ℓ does not create new features; it reweights existing latent responses according to their estimated relevance. In the low-SNR regime

Table 8. Common Confusion Patterns at Low SNR

True Class	Confused With	Reason	Mitigation
AM-DSB	CPFSK	Smooth envelope	Fusion
QAM	PSK	Constellation blur	Attention
Analog FM	AM	Energy similarity	Multi-scale
FSK	Noise-like	Weak transitions	Residual reuse

Table 9. Complexity and Deployment Factors

Factor	Concern	Impact	Solution
Parameters	Memory	Hardware limits	Grouping
Latency	Real-time need	Delay	Short windows
Fusion Cost	Extra memory	Overhead	Sparse links
Attention		Moderate cost	Lightweight gates

such reweighting is useful because informative channels may be weak in magnitude but consistent in their task contribution, whereas nuisance-dominated channels may be strong yet nondiscriminative.

A representative benchmark study using 11 modulation classes, 128-sample IQ segments, and SNR values from -20 dB to 18 dB reported that combining residual fusion with channel reweighting improved mean recognition by about 4% at -10 dB and reduced a dominant AM-DSB-to-CPFSK confusion pattern, as described by Zou et al. (2022) [34]. That observation is technically important because it suggests that the benefit of the design does not arise primarily from parameter growth, but from the way feature reuse and channel emphasis reshape the decision process when the evidence is weak and class manifolds overlap [33].

The architectural role of grouped convolution deserves further analysis. In a dense convolution, every output channel mixes all input channels, which can produce expressive features but also increases parameter coupling. Grouped convolution partitions the channel space and learns several structured transformations in parallel. This induces a form of modularity that is useful when the relevant latent factors are heterogeneous. One group may respond to short-term rotational motion in the IQ plane, another to envelope bursts, another to symbol-boundary transients, and another to smoother analog variations. Because the groups are aggregated only after separate transformation, the model can represent several explanatory hypotheses for the same noisy input without collapsing them immediately. Such diversity is valuable when the classifier faces uncertainty about which cues are reliable in a given sample.

Cross-stage fusion and channel attention interact in a complementary way. Fusion restores access to early information, while attention decides how much of that restored information should matter in context. Without fusion, attention can only rescale whatever the deep stack has already preserved. Without attention, fusion may reintroduce useful details together with irrelevant ones, leaving the classifier to disentangle them downstream. The combined mechanism is therefore more than additive [21]. Fusion broadens the candidate evidence set, and attention performs reliability weighting over that set. This combination is particularly appropriate for short-window modulation recognition because local and global cues coexist, and the relative importance of each varies with SNR, channel condition, and modulation family.

The classifier head operates on the final latent tensor after one or more pooling operations and produces class logits. Global average pooling is often preferred over large fully connected layers because it reduces parameter count and encourages channel semantics, but pooling alone should not be allowed to erase too much temporal information before the final decision. A moderate design therefore uses pooling after several feature-extraction stages rather than immediately after the stem. Let $h(x)$ denote the final pooled feature vector and W_c the classifier weights. The logit for class m is

$$g_m(x) = W_c^{(m)} h(x) + b_m \tag{11}$$

and the softmax posterior follows in the usual way. The interpretive value of this formulation is that each class logit is a weighted sum of channels whose activations have already been filtered by grouped residual transformations, fusion, and channel attention. The hope is not that the network discovers a single canonical descriptor for each modulation, but that it learns a robust mixture of descriptors whose usefulness varies smoothly across noise conditions.

From a signal-processing perspective, the architecture can be seen as an adaptive multiscale estimator. Shallow filters behave somewhat like learned local detectors for oscillation, transition, or envelope patterns [31]. Middle layers aggregate these detections over larger temporal neighborhoods, effectively estimating short-memory dependencies and latent waveform regimes. Deep layers form class-aligned summaries, but they do so with continued access to early evidence and with channel-wise reliability control. This interpretation explains why such architectures are appropriate for low-SNR recognition. The signal energy available for classification is not concentrated in a single representation domain. It is spread across weak local features and broader contextual patterns, both of which must remain available until the final decision is formed.

5. Optimization and Generalization Under Noise Uncertainty

Architectural design alone does not determine low-SNR recognition performance. The training objective, sampling

Table 10. Robustness Enhancements and Future Directions

Technique	Idea	Benefit	Limitation
Self-Supervision	Pretraining	Better features	Data demand
Equivariance	Built-in invariance	Stability	Design complexity
Domain Augmentation	Realistic noise	Transferability	Tuning needed
Uncertainty Modeling	Confidence control	Safer decisions	Added complexity

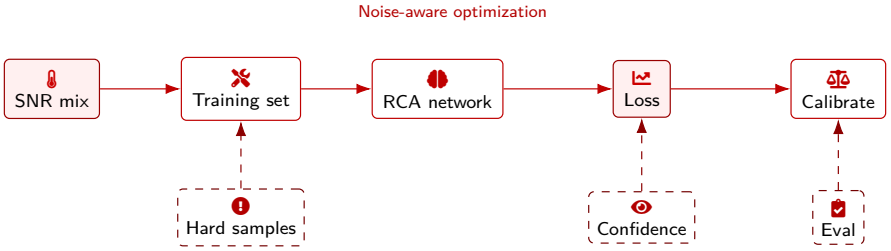


Figure 6. Training and evaluation workflow under SNR uncertainty. Mixed-SNR samples expose the model to diverse observation difficulty, challenging examples influence representation learning, posterior confidence is monitored during optimization, and calibration is applied so that output probabilities remain useful even when class evidence is weak or uneven across operating conditions.

strategy, and regularization regime shape which invariances the model will learn and how its confidence behaves across SNR. Let the training set be $\{(x_i, y_i, \gamma_i)\}_{i=1}^{N}$, where y_i is the modulation label and γ_i is the associated SNR or an SNR proxy. A natural baseline loss is cross-entropy,

$$\mathcal{L}_{\mathrm{CE}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log f_\theta^{(y_i)}(x_i) \tag{12}$$

but this objective treats all examples equally and does not account for the fact that extremely low-SNR samples are both more difficult and often more informative for robustness. If the data distribution is imbalanced across SNR or if the optimization is dominated by easy high-SNR samples, the learned classifier may achieve respectable average accuracy while remaining poorly calibrated exactly where robustness matters most.

One remedy is SNR-aware weighting. Define a nonnegative weighting function $\alpha(\gamma)$ that increases the relative importance of difficult SNR regions without overwhelming the optimization [18]. The weighted objective becomes

$$\mathcal{L}_{\mathrm{W}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \alpha(\gamma_i) \log f_\theta^{(y_i)}(x_i) \tag{13}$$

where $\alpha(\gamma)$ may be piecewise constant, smoothly decaying, or learned through validation. This objective does not create information that is absent from the samples, but it does bias representation learning toward features that remain predictive under poor conditions. In practice, excessive weighting of the lowest SNR bins can destabilize training because the labels there are intrinsically harder to infer from short windows. A balanced design therefore uses moderate weighting together with curriculum or mixed-SNR batching so that the model first learns stable coarse distinctions and then refines them under harder conditions.

Regularization is also central to low-SNR generalization. The usual ℓ_2 penalty on parameters controls overall scale, but it does not directly address sensitivity to noise perturbations. A complementary perspective examines the local Lipschitz behavior of the classifier. If a network with Jacobian $J_\theta(x)$ has large operator norm in typical regions of the input space, then small noise realizations may produce disproportionately large logit changes. The perturbation of the representation can be approximated by

$$\|\phi_\theta(x+\delta) - \phi_\theta(x)\|_2 \le \|J_\theta(x)\|_2 \, \|\delta\|_2 \tag{14}$$

for sufficiently small δ. Architectural features such as residual connections and normalized grouped transformations can improve optimization stability, but explicit regularization of sensitivity remains beneficial. Data augmentation with stochastic phase rotation, amplitude scaling, timing jitter, or mild frequency drift can be interpreted as encouraging local invariance along nuisance directions [12]. Importantly, the augmentation should be physically plausible. Arbitrary perturbations can misalign the training objective with the actual channel model and reduce transfer.

Label smoothing can improve calibration by preventing the classifier from becoming overconfident on noisy or ambiguous samples. If ε is a small smoothing factor and K is the number of classes, the target distribution becomes

$$\tilde{y}_{i,m} = (1-\varepsilon)\,\mathbf{1}[m = y_i] + \frac{\varepsilon}{K} \tag{15}$$

and the loss is evaluated against $\tilde{y}_i$ rather than a one-hot vector. In low-SNR recognition this has two effects. First, it reduces the tendency of the model to assign extreme posterior mass to one class when the evidence is weak. Second, it encourages the latent representation to preserve graded similarity structure among classes, which is useful for confusion analysis and downstream decision fusion. The disadvantage is that too much smoothing can blunt separability at higher SNR, so the hyperparameter should be chosen conservatively.

A more structured view of optimization decomposes risk by SNR:

$$\mathcal{R}(\theta) = \int \mathcal{R}(\theta \mid \gamma)\, p(\gamma)\, d\gamma$$
$$\mathcal{R}(\theta \mid \gamma) = \mathbb{E}\big[\ell(f_\theta(X), Y) \mid \Gamma = \gamma\big] \tag{16}$$

where ℓ is the per-sample loss. This decomposition clarifies that a single global accuracy number can obscure very different behaviors across SNR. Two classifiers may have similar average performance but opposite robustness characteristics, with one dominating at high SNR and another dominating in the adverse region [8]. For low-SNR research, it is therefore more meaningful to inspect the conditional risks, their slopes with respect to SNR, and the class-specific confusion patterns they induce.

The interplay between feature fusion and optimization deserves emphasis. Fusion paths increase the number of active information routes through the network, which can regularize training implicitly by preventing any single deep path from monopolizing the decision. Yet fusion can also introduce redundancy. If shallow features are fused indiscriminately at every stage, the classifier may overfit to local artifacts or develop shortcut dependencies on features that are stable in the training data but not semantically causal. The corresponding design principle is selective fusion rather than maximal fusion. Cross-stage connections should be used where the scales are complementary, not where they merely duplicate information. In low-SNR tasks, fusion from early or mid-level stages into deeper stages is often more useful than exhaustive stage-to-stage densification.

Another important issue is posterior calibration. The softmax output is often interpreted as a probability, but neural networks can be miscalibrated, especially on noisy or out-of-distribution samples. Temperature scaling provides a simple post hoc correction:

$$p_T(m \mid x) = \frac{\exp(g_m(x)/T)}{\sum_{k=1}^{K} \exp(g_k(x)/T)} \tag{17}$$

with $T > 1$ typically softening overconfident predictions [19]. In low-SNR modulation recognition, calibration matters because the classifier may be used as an upstream component in a larger receiver chain. A poorly calibrated recognizer can mislead adaptive processing modules that rely on posterior uncertainty. The architecture described earlier tends to improve class separability, but calibration still requires explicit assessment because higher accuracy does not automatically imply better probabilistic reliability.

Domain shift remains a major challenge even when a model is well optimized on a benchmark. Real receivers differ in gain control, front-end filtering, oscillator stability, and interference patterns. A model trained on synthetic or semi-synthetic data can learn brittle assumptions about these factors. Techniques such as feature normalization, augmentation, and moderate channel attention can reduce sensitivity, but none fully solve the problem. The generalization target should therefore be framed carefully: the goal is not universal invariance, but improved robustness within a class of plausible receiver and channel conditions. This framing helps avoid overinterpreting gains measured on a single standardized corpus.

6. Experimental Protocol and Evaluation Criteria

A technically meaningful evaluation protocol for low-SNR automatic modulation recognition must do more than report overall accuracy. Because the core scientific question concerns robustness under adverse conditions, the protocol should resolve performance by SNR, by class, and by error type. The starting point is a dataset of short IQ sequences labeled by modulation class and, if available, tagged by nominal SNR [2]. Such datasets should include both analog and digital classes, because low-SNR confusion often occurs not only within a family but also across families when the observable waveform degenerates to a small set of weak temporal patterns. Observation windows should be fixed in length to avoid confounding architecture behavior with variable context size, and train-test partitioning should be stratified by class and SNR wherever possible.

The primary scalar metric is top-1 accuracy, but it should be reported as a function of SNR rather than in aggregate alone. The curve $\mathrm{Acc}(\gamma)$ reveals whether an architecture merely improves already easy cases or meaningfully shifts the decision boundary in the adverse region. Complementary metrics include macro-averaged recall, balanced accuracy, confusion entropy, expected calibration error, and class-conditional precision at selected SNR values. Macro-averaging is particularly important when some classes become nearly indistinguishable at low SNR. A model with high aggregate accuracy may still fail systematically on a subset of modulations whose decision regions are poorly represented in the latent space.

Confusion matrices provide more insight than accuracy curves because they expose which classes are being collapsed into which alternatives. In low-SNR modulation recognition, the dominant failure modes are rarely random. They reflect structural similarity under corruption. A continuous-phase signal may absorb predictions from a constant-envelope phase or frequency modulation family because noisy local dynamics appear smoother than they truly are. An analog amplitude-modulated waveform may be confused with a wideband frequency-modulated waveform if only coarse energy or spectral occupancy cues survive. A useful evaluation therefore inspects the evolution of the confusion matrix across several SNR levels, rather than only at the highest or lowest condition.

Ablation analysis is indispensable when the model includes grouped residual paths, cross-stage fusion, and channel attention [3]. The contribution of each mechanism should be isolated by comparing a baseline residual backbone against variants that add one mechanism at a time. The point of the ablation is not merely to show that every module helps, but to determine where it helps. Feature fusion may improve mostly in the moderately low-SNR regime by preserving weak local evidence. Channel attention may reduce specific confusion pairs by suppressing nuisance-aligned latent responses. Grouped transformations may increase representational diversity without large parameter growth. These effects should be examined both globally and conditionally on SNR.

Training and evaluation should also consider calibration and confidence structure. Reliability diagrams, negative log-likelihood, and temperature-scaled posteriors can reveal whether the model knows when it is uncertain. This is crucial for low-SNR operation because a practical receiver may

prefer abstention or deferred processing when class evidence is insufficient. A high-accuracy but overconfident model can be less useful than a slightly less accurate model whose uncertainty correlates better with correctness. Entropy of the posterior distribution is a natural diagnostic in this context. If the model remains sharply peaked even where classes are heavily overlapped, then the learned representation may still be relying on unstable shortcuts [27].

Robustness testing should extend beyond the nominal training distribution. Even when the benchmark is synthetic, additional perturbation sweeps can approximate realistic mismatch conditions. These include mild frequency-offset drift, amplitude scaling, impulse contamination, front-end clipping, sample-rate mismatch, and colored interference. The objective is not to build a universal stress test, but to examine whether the architecture's gains are tied narrowly to one corruption model or whether they persist under adjacent conditions. A model that improves only on nominal white-noise corruption but collapses under slight oscillator mismatch is of limited practical interest.

Complexity must also be measured. Parameter count, multiply-accumulate cost, peak activation memory, and latency per inference window should be reported alongside recognition metrics. Low-SNR models are often deployed in resource-constrained environments where the classification head sits near the front end of a real-time receiver. Grouped residual designs can be efficient relative to dense expansions, but cross-stage fusion and attention introduce overhead that should be quantified rather than assumed negligible. Complexity-aware evaluation is especially important when two models achieve similar recognition curves but differ substantially in latency or energy consumption.

Finally, evaluation should distinguish between within-distribution robustness and transfer robustness. Within-distribution robustness concerns the SNR and nuisance range present during training [4]. Transfer robustness concerns deployment conditions that shift these distributions moderately. The former can be measured with standard train-test splits, while the latter benefits from controlled perturbation studies or cross-receiver validation. A careful protocol reports both when possible, because an architecture that preserves shallow and deep information simultaneously may transfer better precisely because it relies less on one brittle feature scale.

7. Results Interpretation and Error Analysis

When a residual channel-attentive model is evaluated under the protocol described above, the most informative findings are usually qualitative patterns rather than isolated headline numbers. The first pattern concerns the shape of the accuracy-versus-SNR curve. Architectures that only add raw capacity tend to improve more visibly in the high-SNR regime, where the signal is already separable and additional parameters can refine boundaries. Architectures that preserve and reweight multiscale information tend to produce their most meaningful gains in the intermediate-to-low SNR region, where the signal contains recoverable but fragile evidence. This is exactly the region in which premature compression harms performance and where cross-stage feature reuse becomes valuable.

The second pattern concerns class-specific confusion. In low-SNR conditions, confusion is not evenly spread over all classes. Instead, prediction mass accumulates around a few attractor classes whose latent signatures are broad, smooth, or easy to activate under noise. Continuous-phase and constant-envelope families often play this role because the model can interpret noisy local phase continuity as weak evidence for such classes [32]. Analog amplitude-based waveforms can also become attractors when the network overweights coarse envelope energy. The role of channel attention is to reduce indiscriminate activation of these broad attractors by suppressing channels that respond strongly but nonspecifically. In effect, attention sharpens the latent evidence budget so that ambiguous samples are not all drawn toward the same few classes.

Feature fusion has a different but complementary effect. When a class is defined partly by subtle local dynamics that do not survive deep pooling, a fused model can recover those dynamics from earlier layers and use them to counter a misleading deep-level hypothesis. Consider a phase-modulated signal whose global appearance under low SNR resembles another phase family but whose short-term transition statistics remain slightly different. A pure deep stack may compress both classes into a similar latent region. A fused stack can preserve access to the earlier transition evidence and shift the final logit in the correct direction. This does not guarantee perfect recognition, but it reduces the probability that the classifier commits to a wrong class based solely on an over-smoothed deep representation.

Another interpretive lens is the posterior entropy. In a well-behaved low-SNR recognizer, entropy should rise as SNR decreases, but not uniformly across classes. Some classes remain relatively identifiable because their local or envelope features are robust. Others lose distinctiveness quickly [22]. If the model's entropy profile is flat across all classes or remains low even in difficult conditions, the representation may not reflect true uncertainty. Residual fusion and channel attention can improve this profile indirectly by making the evidence pathways more faithful. When the relevant features are preserved and nuisance responses are suppressed, the network becomes less likely to issue spuriously certain predictions. This is important not only for decision quality but also for downstream fusion with demodulators or tracking modules.

Error analysis should also consider temporal support. Short windows amplify the effect of random symbol alignment and partial observation. A classifier may fail not because the modulation is intrinsically indistinguishable, but because the selected window captures a nonrepresentative fragment. Cross-stage fusion mitigates this to some degree by allowing the model to combine very local patterns with broader context, yet the fundamental limit remains. This suggests that some errors attributed to architecture are actually errors of insufficient observation time. In practice, a robust recognizer should ideally support sequential integration across multiple windows, either explicitly through a temporal aggregator or implicitly through decision smoothing. The present architecture is compatible with such extensions because its output is a calibrated posterior over classes.

The distribution of false positives is equally revealing [28]. Suppose a model frequently predicts one class across multiple unrelated source classes at very low SNR. This behavior indicates that the learned feature space contains a dominant basin that is easy to enter under uncertainty. Channel attention can reduce this effect if the basin is caused by overactive

nuisance channels. However, if the basin reflects a genuine lack of information in the samples, then no architectural modification can eliminate it completely. The practical response is not to insist on certainty but to incorporate abstention or top-k decision logic. Low-SNR automatic modulation recognition benefits from this realism: there are conditions under which the waveform does not support a reliable single-class decision from a short segment, and a good recognizer should expose rather than hide that fact.

A subtle but important aspect of error interpretation concerns analog versus digital classes. The distinction between these families is often treated as coarse and therefore easy, but low SNR complicates the picture. Analog waveforms can exhibit smoother trajectories that seem easier to recognize, yet front-end distortions and limited observation windows may erase the very characteristics that mark them as analog. Conversely, certain digital modulations can mimic analog smoothness when pulse shaping and noise interact. This is one reason why confusion analysis is more valuable than family-level accuracy summaries. The model should be studied at the specific class level, where one can observe whether improvements arise from genuine class discrimination or merely from better family separation [10].

The geometry of the final latent space can be explored qualitatively through dimensionality reduction or class-center analysis, though such visualizations should not be overinterpreted. In low dimensions, clusters may appear separable even when decision margins are poor in the full feature space, or they may appear tangled even though the classifier operates effectively on nonlinear boundaries. More informative are class-wise margin statistics and logit-gap distributions. If residual fusion increases the average correct-class margin specifically in the low-SNR regime, then its effect is not merely cosmetic. Similarly, if channel attention decreases the variance of nuisance-aligned logits across classes, then it is functioning as intended.

The limitations of the architecture should be stated clearly. First, the model still depends on the representativeness of the training distribution. If deployment conditions introduce interference types or front-end distortions not seen in training, the gains from fusion and attention may shrink. Second, attention can occasionally suppress weak but useful channels if the gating network is itself biased by the training set. Third, grouped convolutions reduce parameter coupling but do not solve all issues of feature redundancy. Finally, the approach remains fundamentally supervised. It assumes labeled waveform data across relevant SNR values, which can be expensive or unrealistic outside synthetic benchmarks [14]. These limitations do not negate the value of the architecture, but they define the context in which its behavior should be interpreted.

8. Complexity, Robustness, and Deployment Considerations

The practical attractiveness of a low-SNR modulation recognizer depends not only on its discrimination quality but also on its computational profile. A front-end recognition module may operate on streaming data, on embedded hardware, or under strict latency and power budgets. Dense deep networks can be problematic in such settings because large convolutional stacks and wide fully connected layers raise both inference cost and memory traffic. Grouped residual designs partly alleviate this by increasing representational diversity without fully dense channel mixing at every layer. This gives a favorable trade-off between expressivity and computational burden, particularly when the number of groups is chosen to match the channel dimensionality of each stage.

Cross-stage fusion adds overhead because earlier feature maps must be retained or recomputed for later use. The cost is not only arithmetic but also memory. In resource-constrained hardware, activation memory can be more limiting than parameter count. A practical implementation therefore benefits from sparse fusion, where only selected stages are bridged and only compact projected versions of shallow features are passed forward. This preserves the conceptual benefit of feature reuse while keeping the memory footprint manageable. Similarly, channel attention introduces a modest but nonzero cost through global pooling and small gating networks. The overhead is usually acceptable when compared with the benefit in difficult conditions, but it should still be profiled carefully [15].

Latency is another deployment concern. Automatic modulation recognition may be required before later receiver stages commit to demodulation or adaptive filtering. If the model consumes long windows or requires multiple forward passes for uncertainty estimation, the system-level response time may become unacceptable. The present architecture is compatible with short-window processing, which is advantageous. However, short windows increase epistemic uncertainty in some cases, and the system designer must choose between early decisions and delayed but more reliable integration across windows. A sensible deployment strategy is to use the network on short windows but allow posterior accumulation over time when the confidence is low. This preserves responsiveness for easy cases while providing a mechanism for cautious integration when the signal is ambiguous.

Quantization and pruning are relevant for embedded deployment. Because grouped residual and attentive models already organize features into modular channels, they can often tolerate structured pruning better than monolithic dense architectures. Channel-wise importance estimates derived from the learned gates may even inform pruning decisions, though such use should be validated empirically because a channel that appears weak on average may still carry essential information in specific SNR ranges. Mixed-precision inference is also attractive. Early convolutions that process raw IQ data may benefit from higher precision, whereas deeper layers and gating pathways can often be quantized more aggressively with limited degradation [23]. The main caution is that calibration may deteriorate under aggressive quantization even if top-1 accuracy remains similar.

Robustness under real receiver mismatch remains a primary challenge. Synthetic datasets usually model additive noise, frequency offset, and simple fading more cleanly than actual hardware produces them. Real front ends introduce image leakage, nonlinear distortion, automatic gain control transients, and local oscillator imperfections whose joint effect can alter the latent statistics learned by the network. The architecture developed here offers some protection because fusion preserves multiple evidence scales and attention can downweight some nuisance-dominated channels. Yet this should not be conflated with full domain robustness. Deployment in realistic systems benefits from continued adaptation,

domain augmentation during training, or unsupervised alignment strategies that recalibrate feature statistics without requiring exhaustive relabeling.

There is also a question of interpretability. Although deep modulation recognizers need not be fully transparent to be useful, some degree of diagnostic visibility is important in high-consequence applications. Channel attention provides a limited interpretive handle because the gating values indicate which latent channels were emphasized for a given input. Combined with saliency over time or feature ablation, this can reveal whether the model is relying on local phase transitions, envelope bursts, or broader temporal patterns. Such diagnostics are not proofs of causal reasoning, but they can help identify dataset artifacts, channel-specific brittleness, or undesirable shortcut behavior [11]. From an engineering standpoint, this diagnostic capacity is valuable because it guides retraining and architecture revision.

Future robustness improvements are likely to arise from hybrid designs rather than purely larger networks. One direction is self-supervised pretraining on unlabeled IQ streams, allowing the model to learn invariances and temporal structure before supervised fine-tuning for modulation classes. Another is explicit equivariance to carrier phase rotation or frequency drift, which could reduce the amount of data required to learn these transformations implicitly. A third is uncertainty-aware decision layers that combine class prediction with abstention or open-set detection, thereby handling conditions where the observed signal does not resemble any trained class. These directions do not replace residual fusion and channel attention; rather, they complement them by broadening the set of conditions under which the latent features remain useful.

Deployment decisions should reflect the specific operational objective. If the recognizer is used to narrow a candidate set for downstream demodulation, then calibrated top-k performance may matter more than top-1 accuracy. If it is used for spectrum monitoring, stable class posteriors over time may be more important than single-window decisiveness. If it is used on edge hardware, latency and energy dominate. The residual channel-attentive architecture fits well within this broader design space because it improves the quality of short-window features without demanding extreme model size. Its value lies less in universal optimality than in the balance it strikes among feature preservation, robustness, and efficiency [5].

9. Conclusion

Low-SNR automatic modulation recognition is fundamentally a problem of inference under weak evidence and strong nuisance variability. The main technical challenge is not simply to build a larger classifier, but to preserve the fragile waveform structures that remain informative after noise, channel distortion, and finite-window uncertainty have obscured the clean modulation geometry. This paper developed a research-oriented treatment of that challenge by combining a signal-level observation model with a residual channel-attentive deep architecture. The key design logic is that shallow waveform cues and deeper semantic abstractions should coexist rather than compete, and that latent channels should not contribute uniformly when only a subset remains reliable under adverse conditions.

The analysis showed why grouped residual transformations, cross-stage feature fusion, and channel-selective emphasis are well matched to low-SNR recognition. Grouped transformations diversify feature extraction without excessive parameter coupling. Fusion preserves access to local evidence that may otherwise be lost in a deep hierarchy. Channel attention acts as a reliability allocator that suppresses nuisance-dominated responses while amplifying channels that retain class information. These mechanisms do not eliminate the intrinsic ambiguity of very noisy short observations, but they provide a principled way to use the available evidence more effectively.

Equally important, the paper emphasized that evaluation should be resolved by SNR, class, confusion structure, calibration, and computational cost. A single accuracy number is insufficient for this problem. Robustness claims must be tied to the specific conditions under which they hold, and deployment decisions should account for latency, memory, and mismatch across receivers. Within that neutral framing, residual channel-attentive feature fusion emerges as a technically coherent direction for low-SNR modulation recognition, particularly when the goal is to improve discrimination without sacrificing the multiscale structure of the received waveform [13].

References

[1] A novel recognition algorithm based on holder coefficient theory and interval gray relation classifier. *KSII Transactions on Internet and Information Systems*, 9(11), 11 2015.

[2] Soumia Sid Ahmed, Zoubeida Messali, Larbi Boubchir, Ahmed Bouridane, Sergio Marco, and Cédric Messaoudi. Iterative bayesian denoising based on variance stabilization using contourlet transform with sharp frequency localization: application to eftem images. *BMC biomedical engineering*, 1(1):1–12, 6 2019.

[3] Faouzi Bellili, Souheib Ben Amor, Sofiene Affes, and Ali Ghrayeb. Maximum likelihood joint angle and delay estimation from multipath and multicarrier transmissions with application to indoor localization over ieee 802.11ac radio. *IEEE Transactions on Mobile Computing*, 18(5):1116–1132, 5 2019.

[4] Liu Bin, Feng Youqian, Yin Zhonghai, and Fan Xiangyu. Radar signal emitter recognition based on combined ensemble empirical mode decomposition and the generalized s-transform. *Mathematical Problems in Engineering*, 2019(1):1–15, 4 2019.

[5] Ballard J. Blair and James C. Preisig. Comparison and analysis of equalization techniques for the time-varying underwater acoustic channel. *The Journal of the Acoustical Society of America*, 126(4):2249–2249, 10 2009.

[6] Andrea Bocchieri, Enzo Pascale, Lorenzo Mugnai, Quentin Changeat, and Giovanna Tinetti. Observability of exo-atmospheres in emission using ariel, 7 2021.

[7] Holly Dale, Christopher J. Baker, Michail Antoniou, Mohammed Jahangir, George Atkinson, and Stephen Harman. Snr-dependent drone classification using convolutional neural networks. *IET Radar, Sonar & Navigation*, 16(1):22–33, 7 2021.

[8] Junjie Ding, Jiao Zhang, Yiran Wei, Feng Zhao, Chaoyang Li, and Jianjun Yu. Comparison of geometrically shaped 32-qam and probabilistically shaped 32-qam in a bandwidth-limited im-dd system. *Journal of Lightwave Technology*, 38(16):4352–4358, 8 2020.

[9] Steven J. Gibbons, Johannes Schweitzer, Frode Ringdal, Tormod Kværna, Svein Mykkeltveit, and Berit Paulsen. Improvements to seismic monitoring of the european arctic using three-component array processing at spits. *Bulletin of the Seismological Society of America*, 101(6):2737–2754, 12 2011.

[10] Pansoo Kim, In-Ki Lee, Deock-Gil Oh, and Joon-Gyu Ryu. *Robust initial access technique of spread spectrum based on DVB-RCS2 standard for mobile application*, pages 217–221. Institution of Engineering and Technology, 11 2019.

[11] Smt. D.Lalitha Kumari* and Dr. M.N.Giri Prasad. Performance of mimo systems over weibull fading channel using maximal ratio combining. *International Journal of Innovative Technology and Exploring Engineering*, 9(2):4410–4415, 12 2019.

[12] Seunghee Lee, Peter C. Doerschuk, and John E. Johnson. Multiclass maximum-likelihood symmetry determination and motif reconstruction of 3-d helical objects from projection images for electron microscopy. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 20(7):1962–1976, 2 2011.

[13] Yunseong Lee, Chanhong Park, Taeyoung Kim, Yeongyoon Choi, Kiseon Kim, Dongho Kim, Myung-Sik Lee, and Dong-Keun Lee. Source enumeration approaches using eigenvalue gaps and machine learning based threshold for direction-of-arrival estimation. *Applied Sciences*, 11(4):1942–, 2 2021.

[14] Qiufeng Li, Gengsheng Luo, Guo Chen, and Pan Huang. Application of empirical mode decomposition for ultrasonic testing of coarse-grained materials. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, 11(9):5322–5328, 9 2013.

[15] Xuan LI and Xiaochuan Ma. Joint doppler shift and time delay estimation by decovonlution of generalized matched filter, 3 2021.

[16] Bingfan Liu, Baixiao Chen, and Minglei Yang. Parameter estimation and crb analysis of 1-bit colocated mimo radar. *IET Radar, Sonar & Navigation*, 15(6):592–604, 4 2021.

[17] Yingli Lu, Kim A. Connelly, Yuesong Yang, Subodh B. Joshi, Graham A. Wright, and Perry Radau. Semi-automated analysis of infarct heterogeneity on de-mri using graph cuts. *Journal of Cardiovascular Magnetic Resonance*, 14(1):1–3, 2 2012.

[18] Helge Lüders, Andreas Minwegen, Benedikt Eschbach, and Peter Vary. An improvement of umts lte 64qam performance. *European Transactions on Telecommunications*, 21(5):443–448, 7 2010.

[19] A. Malyali, M Rivi, Filipe B. Abdalla, and Jason D. McEwen. Radio galaxy detection in the visibility domain. *Monthly Notices of the Royal Astronomical Society*, 486(2):2695–2704, 4 2019.

[20] Ikuo Matsuo. Localization of the moving object by echolocation. *The Journal of the Acoustical Society of America*, 131(4):3423–3423, 4 2012.

[21] Sachin Shivaji Nair. Fast and non-invasive raman imaging of material distributions at interfaces, 11 2021.

[22] Ilya Pyatin, Juliy Boiko, and Oleksander Eromenko. Design and simulation of synchronization systems in telecommunications with qpsk. In *2020 IEEE 2nd International Conference on Advanced Trends in Information Theory (ATIT)*, pages 98–103. IEEE, 11 2020.

[23] Dayana. R and Kumar R. Performance analysis of universal filtered multi-carrier -isotropic orthogonal transform algorithm for next generation wireless networks. *International Journal of Engineering & Technology*, 7(4.19):109–113, 11 2018.

[24] Carl L. Rodriguez, Benjamin Farr, Will M. Farr, and Ilya Mandel. Inadequacies of the fisher information matrix in gravitational-wave parameter estimation. *Physical Review D*, 88(8):084013–, 10 2013.

[25] Anas M. Salhab and Salam A. Zummo. Performance analysis of switch-and-examine diversity systems with interference and rayleigh fading channels. *Wireless Communications and Mobile Computing*, 16(8):884–894, 1 2015.

[26] Lalitha Sankar, Gerhard Kramer, and Narayan B. Mandayam. Dedicated-relay vs. user cooperation in time-duplexed multiaccess networks. *Journal of Communications*, 6(4):330–339, 7 2011.

[27] Abrar Siddique, Tahesin Samira Delwar, and Jee-Youl Ryu. A novel optimized v-vlc receiver sensor design using ga in automotive applications. *Sensors (Basel, Switzerland)*, 21(23):7861–, 11 2021.

[28] Sofie Tapper, Anders Tisell, and Peter Lundberg. How does motion affect gaba-measurements? order statistic filtering compared to conventional analysis of mega-press mrs. *PloS one*, 12(5):e0177795–, 5 2017.

[29] Mingjian Tuo and Xingpeng Li. Long-term recurrent convolutional network-based inertia estimation using ambient measurements, 12 2021.

[30] Carlos Uribe, Edward C. Jones, Silvia D. Chang, S. Larry Goldenberg, Stefan A. Reinsberg, and Piotr Kozlowski. In vivo 3t and ex vivo 7t diffusion tensor imaging of prostate cancer: Correlation with histology. *Magnetic resonance imaging*, 33(5):577–583, 2 2015.

[31] Wen Xu and Arthur B. Baggeroer. Quantitative ambiguity analysis for matched-field parameter estimation. *The Journal of the Acoustical Society of America*, 110(5):2716–2716, 11 2001.

[32] Yankai Xu, Siyuan Cao, Xiao Pan, Wei Liu, and Hongling Chen. Random noise attenuation using a structure-oriented adaptive singular value decomposition. *Acta Geophysica*, 67(4):1091–1106, 5 2019.

[33] Huanhai Yang. Detection and tracking of infrared dim small image sequence moving target. *The Open Automation and Control Systems Journal*, 7(1):1698–1704, 10 2015.

[34] Binghang Zou, Hanzhi Yan, Faquan Wang, Yucheng Zhou, and Xiaodong Zeng. Research on signal modulation classification under low snr based on resnext network. *Electronics*, 11(17):2662, 2022.