

Interference-Robust Automatic Modulation Recognition under Skewed Deployment Priors

Ruizhi Li^{1,a,*}, Ruisen Luo^{1,b}

¹*School of Computer Science and Artificial Intelligence, Southwest Minzu University, Chengdu, China*

^a*610626596@qq.com*

^{*}*Corresponding author*

Keywords: Automatic Modulation Recognition; Interference Robustness; Class Imbalance; Macro-Averaged Recall; Deployment Prior

Abstract: Deep learning-based automatic modulation recognition (AMR) models are typically evaluated on idealized synthetic benchmarks that assume interference-free conditions and balanced class distributions. However, real-world non-cooperative receivers inevitably suffer from severe structural interference (e.g., localized bursts or full-band jamming) and heavily skewed deployment priors. To investigate these practical mismatches, this paper systematically evaluates a compact convolutional-recurrent backbone on RadioML 2016.10a under both interference and class-imbalance scenarios. We introduce a lightweight sequence gate that adaptively recalibrates temporal feature channels with only 850 additional parameters, combined with IQ-plane augmentation and post-hoc logit calibration. Extensive experiments show that burst-interference robustness is largely a property of this convolutional-recurrent family and that the proposed sequence gate adds up to 4.1 points on top of the plain backbone, whereas domain-specific augmentation dominates resilience against full-band jamming. Furthermore, when trained under a long-tailed prior (imbalance ratio of 50), the standard model exhibits an alarming gap of 22.3% between deployment-weighted accuracy and macro-averaged recall, revealing severe performance degradation on minority classes that conventional metrics conceal. By incorporating the gate, augmentation, and a calibrated decision rule, the macro recall is substantially restored by up to 13.5% without sacrificing overall accuracy. This work provides empirical insights and an efficient baseline for deploying robust AMR models in non-ideal electromagnetic environments.

1. Introduction

Automatic modulation recognition (AMR) infers the modulation format of an intercepted waveform and underpins spectrum monitoring and electronic reconnaissance. Deep networks now dominate the field, and models trained on synthetic corpora such as the RadioML benchmark have become the standard reference[1], [2]. Almost all of these results rest on two assumptions that do not describe a receiver: that the waveform is corrupted only by additive Gaussian noise of known signal-to-noise ratio (SNR), and that the classes are equally frequent in training and in the test set

over which accuracy is averaged.

Structural interference breaks the first assumption. A burst from a nearby emitter or a deliberate jammer occupies a contiguous part of the observation window and is not Gaussian. Its waveform family is open: pulse width, duty cycle, bandwidth and centre frequency vary, and the affected window is unknown. Interference is therefore not a class that training can enumerate, so the method studied here does not model it and only makes the encoder less sensitive to it.

Class priors break the second assumption. Public corpora sample the modulation types in equal numbers and evaluation follows the same convention, so reported accuracy is implicitly an average under a uniform prior. At a receiver the prior is skewed and unknown, so a model that is strong on frequent classes can be reported as accurate while failing on the classes that matter. We therefore report macro-averaged recall, which is independent of the prior, next to the deployment-weighted accuracy, and measure the gap between them.

This paper makes three contributions. First, we evaluate a lightweight gate on the recurrent features of a gated convolutional backbone; it adds 850 parameters, needs no interference labels, and improves burst-interference accuracy by up to 4.1 points on the plain backbone and by up to 21.7 points on the augmented one. Second, we define a controlled protocol that injects burst interference and full-band jamming into the raw test snapshots at several amplitudes, so that robustness is measured rather than assumed. Third, we retrain the same backbones on a long-tailed set and show that the deployment-weighted accuracy and macro recall separate by up to 22.3 points, that augmentation with the gate limits the collapse of the tail classes, and that a calibrated decision rule recovers part of the lost macro recall without losing weighted accuracy.

The scope is narrow: all experiments use one synthetic benchmark, the interference and class priors are simulated in a controlled way, and the corrected decision rule assumes the deployment prior is known or estimable. We do not claim to solve class imbalance, only to measure it and reduce its effect at decision time.

2. Related Work

2.1. Backbones for modulation recognition

Convolutional models established the benchmark[1], and temporal convolution was later combined with recurrent layers to exploit the structure of the IQ window[2]. Attention-based variants replaced recurrence with Transformer blocks[3], and scale-aware modulation transformers were introduced to reduce the cost of that design[4]. These models are trained and compared on the same balanced, noise-only protocol, so the two mismatches studied here are outside their evaluation envelope.

2.2. Interference and label-noise robustness

Recognition under strong interference has been addressed with lightweight Transformer networks[5], and at the physical layer deep learning has been applied to anti-jamming beamforming [6]. These methods either add interference-specific machinery or require a training set that already contains the corruption. The gate used here carries no interference labels and is trained only with the classification loss.

2.3. Imbalanced recognition

Class imbalance is a mature topic: re-weighting by the effective number of samples[7], loss shaping[8] and systematic studies of how imbalance affects convolutional training[9] are standard

tools, and comprehensive surveys of long-tailed learning are available[10]. Prior-aware decision rules are the natural remedy at inference time[11]. What is less common is to report how far the weighted accuracy and the macro-averaged recall of the same model separate when the deployment prior drifts, which is the measurement we add here.

3. Method

3.1. Backbone

The backbone follows the compact convolutional-recurrent design used for this benchmark[2]: a one-dimensional convolution with 80 filters and kernel 12, max pooling, two gated recurrent units with 40 units each, Gaussian dropout, a flatten layer and a dense classifier, 52,011 trainable parameters in total. The recurrent stage produces a sequence of $T = 58$ vectors of dimension $C = 40$, so the flattened representation has 2,320 dimensions. All models in this paper share this backbone and differ only in the two components described next.

3.2. Sequence gate

A gate is inserted after the second recurrent unit. It follows the channel recalibration principle of squeeze-and-excitation networks[12], except that the statistics are pooled over the time axis of the recurrent output instead of over spatial locations: it derives one weight per feature channel through a bottleneck of size 10 and applies that weight to every time step.

$$w = \sigma(W_2 \text{ReLU}(W_1 \bar{h})) \quad (1)$$

where $\bar{h}$ is the average over time of the recurrent output and the resulting weight vector is applied to every time step. The two matrices have sizes 40 by 10 and 10 by 40, so the gate adds 850 parameters, or 1.6% of the backbone. Because the weight vector is shared across time steps, the gate does not localise interference in time; it recalibrates the channel response of the recurrent representation according to the statistics of the current input, which is what a short perturbation of the waveform changes first. The gate carries no interference-specific parameter and is trained end to end with the classification loss.

3.3. Training recipe

Every batch is augmented with a random rotation of the IQ plane and a multiplicative gain drawn from a normal distribution with mean 1 and standard deviation 0.1. The rotation matches the carrier phase uncertainty of a non-cooperative receiver, and the gain jitter prevents the encoder from relying on absolute amplitude.

3.4. Deployment-prior calibration

A classifier trained on a balanced set estimates the posterior of the balanced problem. If the deployment prior differs, the decisions that maximise the deployment-weighted accuracy are obtained by adding the log prior to the logits,

$$y^* = \arg \max_y [z_y - \lambda \ln \pi_y] \quad (2)$$

where z are the logits, π is the prior of the deployment population and λ is a scalar. Setting $\lambda = 1$ applies the Bayes rule for that prior and maximises weighted accuracy; positive or negative deviations trade weighted accuracy against macro recall. We select λ on a held-out long-tailed split,

never on the test set.

4. Experimental Results

4.1. Setup

RadioML 2016.10a contains eleven modulation classes with 1,000 snapshots per class, and SNR from -20 dB to 18 dB in steps of 2 dB. For every class and SNR block the first 80% of the snapshots are used for training and the last 20% for testing, giving 44,000 test snapshots; the normalisation statistics come from the training split only. Four backbones are trained with one recipe: 60 epochs, Adam with a learning rate of 2×10^{-3} , batch size 128 and a plateau schedule, and they differ only in the presence of the gate and of the augmentation. All three comparison models follow the same evaluation protocol, the two learnt baselines being retrained with this recipe.

Interference is injected into the raw test snapshots and the features are extracted again with the same frozen encoder. Burst interference adds a pulse of width between 4 and 16 samples to a random sub-window, with amplitude equal to k times the standard deviation of that snapshot, for k in $\{5, 10, 20\}$. Full-band jamming adds noise of the same relative amplitude over the whole window, for k in $\{0.25, 0.5, 1, 2\}$. Three metrics are reported: accuracy for $\text{SNR} \geq 0$ dB, which is a prior-weighted average over the balanced test set; macro recall, the unweighted mean of the per-class recalls; and the deployment-weighted accuracy, which uses the deployment prior instead of the uniform one.

For the prior study the same four backbones are retrained on a long-tailed training set. Class c of the eleven classes receives a fraction $\rho^{-(c/10)}$ of its samples, with $\rho = 50$ and a floor of 300 samples per class in each SNR block; the validation and test sets are left balanced, so that any change in macro recall is caused by the training distribution rather than by the test set.

4.2. Interference robustness

Table 1 reports accuracy for $\text{SNR} \geq 0$ dB and Figure 1 shows the corresponding curves on the clean test set. On clean inputs the three comparison models reach 79.78%, 83.65% and 88.99%, and the largest of them, AMCNet with 466.1k parameters, is within 0.2 points of the 52.0k-parameter baseline and within 1.7 points of the proposed model. The augmentation is the component that lifts the clean accuracy of that baseline to 91.24%, while the proposed combination reaches 90.69%.

Under burst interference the comparison models degrade. At $k = 20$ they retain 19.46%, 43.83% and 62.04%, and the proposed model retains 78.61%, which is 16.57 points above the best of them. The contribution of the gate grows with the strength of the burst on both the plain and the augmented backbone. The augmented backbone alone degrades fastest, from 91.24% to 56.92%, and the gate is what restores it, as shown in Table 2 and Figure 2(a).

Under full-band jamming AMCNet stays the strongest comparison model, reaching 70.72% at $k = 1$ and 47.61% at $k = 2$, and the proposed model follows it closely with 70.27% and 40.65%. The augmentation is the component responsible for this behaviour, because the phase and gain jitter covers the perturbation that the jammer introduces, and the proposed combination degrades more slowly than the plain backbone at every jamming amplitude, as shown in Figure 2(b). The two components therefore cover different failure modes rather than duplicating each other.

Table 1. Accuracy for $\text{SNR} \geq 0$ dB (%) on the balanced test set. B and J denote the burst and jamming amplitude, and the lower jamming amplitudes are shown in Figure 2.

Method	Params	Clean	B10	B20	J1	J2
CNN	266.0k	79.78	31.40	19.46	40.65	23.89
ResNet	106.1k	83.65	60.69	43.83	64.33	37.60
AMCNet	466.1k	88.99	75.41	62.04	70.72	47.61
Proposed	52.9k	90.69	80.51	78.61	70.27	40.65

Table 2. Component ablation of the proposed baseline, accuracy for $\text{SNR} \geq 0$ dB (%). Baseline is the 52k-parameter backbone, +TAG adds the gate and +Aug adds the augmentation.

Method	Params	Clean	B20	J2
Baseline	52.0k	88.82	76.20	25.12
+TAG	52.9k	89.56	80.25	35.74
+Aug	52.0k	91.24	56.92	39.12
Proposed	52.9k	90.69	78.61	40.65

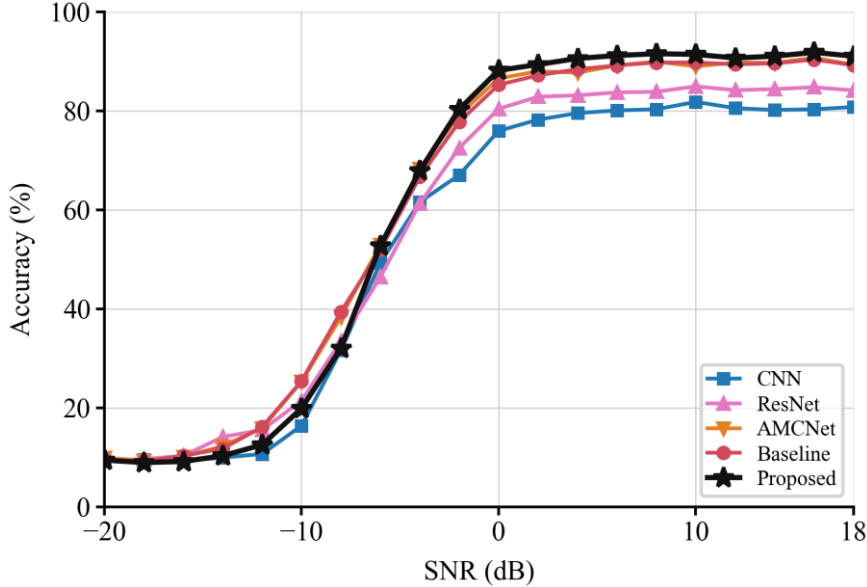


Figure 1. Accuracy on the clean test set as a function of SNR for the compared models.

4.3. Skewed deployment priors

Table 3 contrasts the plain long-tailed backbone with the model that combines the gate and the augmentation. It reports the worst per-class recall next to the macro and the deployment-weighted averages, so that the collapse of the tail and its effect on the two averages can be read together. Training on the skewed set concentrates the capacity on the frequent classes: the recall of QPSK drops from 97.95% with balanced training to 2.35% for the plain backbone, and WBFM, the hardest class, drops from 42.45% to 1.75%. Macro recall follows, from 88.82% to 68.11%. Augmentation and the gate recover most of this loss: the two components together raise QPSK to 92.20%, lift the worst class from 1.75% to 10.05%, and raise macro recall to 79.61%, 11.50 points above the plain backbone.

The same model looks very different under a deployment prior. With a head-heavy prior of ratio $\rho = 50$ the deployment-weighted accuracy of the plain long-tailed model is 90.37% while its macro recall is 68.11%, a gap of 22.26 points, and the gap narrows to 14.54 points for the model that

combines augmentation and the gate. Figure 2(c) traces both metrics as the prior becomes more skewed: the weighted accuracy rises by more than twenty points while macro recall stays flat, which is exactly the effect that a single accuracy number hides. The separation is not an artefact of the tail collapse alone: the per-class recalls of the plain long-tailed model still span 1.75% to 99.90%, and it is that spread which the uniform average conceals.

Table 3. Recall for $\text{SNR} \geq 0$ dB of the plain long-tailed backbone (Baseline) and of the model that combines the gate and the augmentation (Proposed), both retrained on the long-tailed set. All entries, and the per-class and ablation values quoted in the text, come from the same per-class evaluation: macro recall is the unweighted mean of the eleven per-class recalls, the weighted accuracy uses a head-heavy deployment prior of ratio 50, and the gap is the difference between the two. The last two rows apply the calibrated decision rule of (2).

Metric	Baseline	Proposed
Macro recall (%)	68.11	79.61
Worst-class recall (%)	1.75	10.05
Classes below 50% recall (of 11)	4	2
Weighted accuracy at $\rho = 50$ (%)	90.37	94.16
Gap, weighted – macro (points)	22.26	14.54
Weighted accuracy after calibration (%)	91.18	94.77
Macro recall after calibration (%)	71.80	81.64

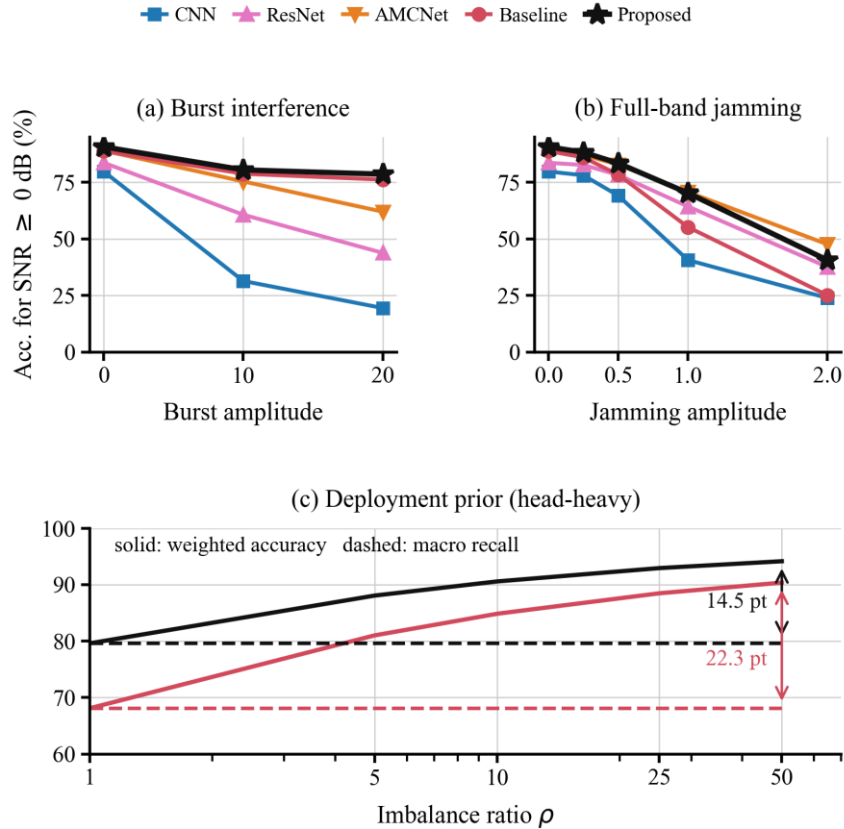


Figure 2. (a) Burst interference and (b) full-band jamming, accuracy for $\text{SNR} \geq 0$ dB against interference amplitude. (c) Under a head-heavy deployment prior the weighted accuracy rises with the imbalance ratio while macro recall stays flat; the arrow marks the 14.5 point gap of the model that combines augmentation and the gate.

4.4. Calibrated decisions

Finally we apply the decision rule of (2) with the threshold selected on a held-out long-tailed split built from the validation set with the same prior. The last two rows of Table 3 report the result. For both models the selected threshold improves the macro recall and the weighted accuracy at the same time, and for the model that combines augmentation and the gate macro recall rises from 79.61% to 81.64% while weighted accuracy rises from 94.16% to 94.77%. The rule is not universally favourable: on the gated backbone alone the weighted accuracy moves by 0.27 points in the other direction while macro recall gains 6.00 points, which is the expected trade-off when the threshold is pushed further towards the rare classes.

The recovery is partial, and the reason is visible in Table 3: the classes that collapse are those that are intrinsically confusable in this benchmark, so no decision rule can restore them. What the calibration does recover is the part of the loss that comes from the decision boundary rather than from the representation.

4.5. Discussion and limitations

Three observations follow. First, the two robustness axes are independent: the augmentation is what protects the model against broadband perturbation, the gate is what protects it against temporally localised perturbation, and neither component dominates on both. Second, a single accuracy number is not a sufficient summary once the prior can drift, because the same weights can be reported as 68.11% or 90.37% depending on which classes are frequent. Third, the cheap part of the remedy is at decision time: a threshold selected on a held-out split recovers 2.0 to 6.0 points of macro recall, whereas recovering the remaining loss requires changing the training distribution, which is what the augmentation and the gate do at a cost of 850 parameters.

The limitations are equally clear. The interference is simulated, and the burst amplitude is uniform across the window while a real jammer may sweep or be frequency selective. The calibration of (2) assumes that the deployment prior is known or can be estimated from a small unlabelled sample, which is realistic for a monitoring receiver but not guaranteed. Finally, the study uses one synthetic benchmark and one long-tailed ratio; both should be repeated on measured signals before the numbers are generalised.

5. Conclusion

This paper studied two ways in which a modulation recogniser meets a receiver that its benchmark did not describe. Under burst interference a lightweight gate on the recurrent features improves accuracy for $\text{SNR} \geq 0$ dB by up to 21.7 points depending on the backbone and the interference amplitude, while under full-band jamming the phase and gain augmentation is the dominant factor. Under a skewed deployment prior the weighted accuracy and the macro recall of the same model separate by up to 22.3 points, the combination of augmentation and gate limits the collapse of the tail classes, and a decision threshold selected on a held-out split recovers part of the remaining loss. Future work will repeat the protocol on measured signals and replace the global gate with a module that localises the interfered portion of the window.

References

- [1] T. J. O'Shea, T. Roy, and T. C. Clancy, "Over-the-air deep learning based radio signal classification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 1, pp. 168–179, 2018. DOI: 10.1109/JSTSP.2018.2797022.
- [2] G. Gao, X. Hu, B. Li, W. Wang, and F. M. Ghannouchi, "CrossTLNet: A multitask-learning-empowered neural network with temporal convolutional network-long short-term memory for automatic modulation classification,"

Electronics, vol. 12, no. 22, p. 4668, 2023. DOI: 10.3390/electronics12224668.

[3] W. Zhang, K. Xue, A. Yao, and Y. Sun, "CTRNet: An automatic modulation recognition based on Transformer-CNN neural network," *Electronics*, vol. 13, no. 17, p. 3408, 2024. DOI: 10.3390/electronics13173408.

[4] Y. Huo, C. Wang, J. Liang, K. Yue, and W. Li, "SMTrans: An efficient automatic modulation recognition network based on the scale-aware modulation transformer," *Physical Communication*, vol. 74, p. 102966, 2026. DOI: 10.1016/j.phycom.2025.102966.

[5] D. Liu, P. Liu, J. Chen, K. Xiong, B. Jing, and D. Liao, "Automatic modulation recognition based on a lightweight Transformer network under strong interference conditions," *Engineering Reports*, vol. 7, no. 9, 2025. DOI: 10.1002/eng2.70405.

[6] O. Kwon, H. Lee, M. Debbah, and I. Lee, "Deep learning-based anti-jamming beamforming designs against adversarial jamming attacks," *IEEE Transactions on Wireless Communications*, vol. 25, pp. 20900–20912, 2026. DOI: 10.1109/TWC.2026.3713089.

[7] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9260–9269. DOI: 10.1109/CVPR.2019.00949.

[8] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007. DOI: 10.1109/ICCV.2017.324.

[9] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018. DOI: 10.1016/j.neunet.2018.07.011.

[10] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep long-tailed learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10795–10816, 2023. DOI: 10.1109/TPAMI.2023.3268118.

[11] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, "Long-tail learning via logit adjustment," in *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2007.07314.

[12] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141. DOI: 10.1109/CVPR.2018.00745.