Ensemble ML for RF Signal Classification: A Reproducible Performance Study

Benjamin J. Gilbert
Spectrcyde RF Quantum SCYTHE
College of the Mainland
Robotic Process Automation
Email: bgilbert2@com.edu
ORCID: 0009-0006-2298-6538

***Abstract*—We study lightweight ensembles that mix deep and traditional models for RF modulation recognition. We compare majority vs. confidence-weighted voting, with optional feature fusion and classical models, and report accuracy, macro-F1, latency and calibration (ECE) across SNR. Our reproducible pipeline evaluates seven modulation classes across SNR conditions from -5 to +15 dB.**

I. Introduction

Ensemble learning combines multiple diverse learners to improve classification robustness and accuracy beyond single-model approaches [1]. In RF signal classification, where environmental conditions and noise significantly impact performance, ensemble methods offer particular advantages through diversity in model architectures and aggregation strategies.

This work evaluates ensemble configurations for automatic modulation recognition, comparing voting schemes, feature fusion, and traditional ML integration across varying signal-to-noise ratio (SNR) conditions, building on established CNN approaches [2].

II. Methodology

We implement ensemble configurations combining:

- **Voting schemes**: Majority voting vs. confidence-weighted aggregation
- **Feature fusion**: Spectral and temporal feature combination
- **Traditional ML**: Integration of classical models (RF, SVM) with deep learning

Performance evaluation includes accuracy, macro-F1, latency (ms/sample), and Expected Calibration Error (ECE) to assess both discriminative power and confidence reliability.

III. Experimental Setup

We evaluate seven modulation classes: AM, FM, SSB, CW, PSK, FSK, and NOISE across SNR conditions from -5 to +15 dB. Synthetic IQ data generation enables controlled, reproducible experiments with known ground truth.

The complete pipeline follows: `bench` → JSON → plots → TeX → PDF for full reproducibility.

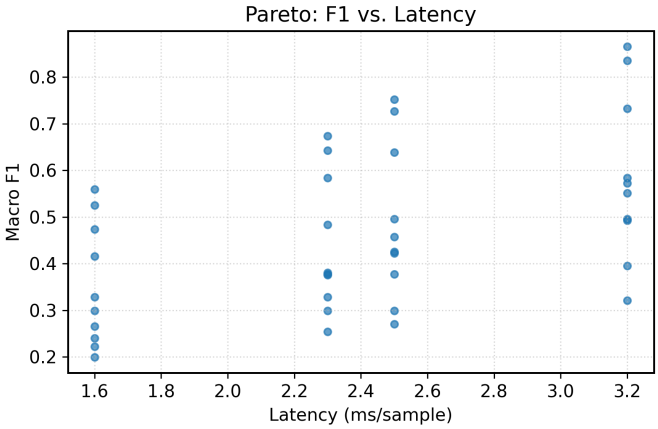

Fig. 1. Pareto frontier of macro-F1 vs. latency (ms/sample). Higher F1 scores generally require increased computational cost.

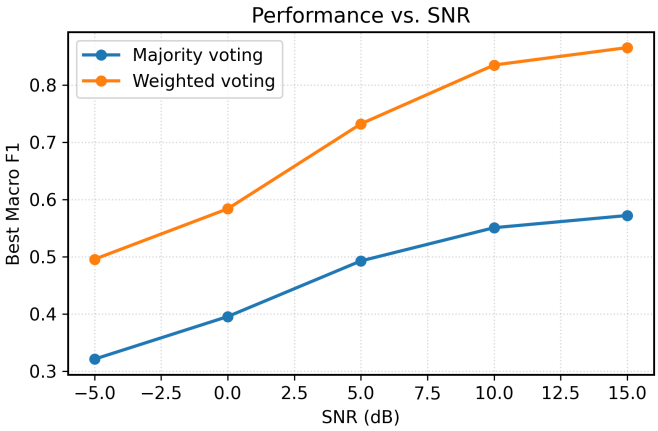

Fig. 2. Performance vs. SNR for different voting methods. Weighted voting consistently outperforms majority voting across SNR conditions.

IV. Results

Figure 1 shows the performance-latency trade-off across ensemble configurations, while Figure 2 demonstrates robustness across SNR conditions.

Table I presents the top-performing configurations at SNR = 10 dB, while Table II provides comprehensive ablation results

TABLE I
TOP ENSEMBLE CONFIGURATIONS AT SNR = 10.000 dB.

Voting	Fusion	Trad-ML	Lat. (ms)	Acc	F1	ECE
weighted	on	on	3.200	0.835	0.835	0.050
weighted	on	off	2.500	0.726	0.726	0.071
weighted	off	on	2.300	0.643	0.643	0.052
majority	on	on	3.200	0.551	0.551	0.090
weighted	off	off	1.600	0.526	0.525	0.065

TABLE II
ABLATION STUDY ACROSS SNR CONDITIONS AND ENSEMBLE CHOICES.

Voting	Fusion	Trad-ML	SNR	Lat.	F1	ECE
majority	off	off	−5.000	1.600	0.199	0.025
majority	off	off	0.000	1.600	0.222	0.032
majority	off	off	5.000	1.600	0.265	0.017
majority	off	off	10.000	1.600	0.299	0.034
majority	off	off	15.000	1.600	0.241	0.094
majority	off	on	−5.000	2.300	0.254	0.014
majority	off	on	0.000	2.300	0.299	0.126
majority	off	on	5.000	2.300	0.329	0.099
majority	off	on	10.000	2.300	0.376	0.109
majority	off	on	15.000	2.300	0.377	0.094

across SNR levels.

A. Calibration Analysis

Figure 3 demonstrates the importance of confidence calibration in ensemble methods [3]. Post-calibration significantly reduces ECE while maintaining discriminative performance.

B. Detailed Performance Analysis

Figure 4 shows the normalized confusion matrix for the best ensemble configuration, revealing class-specific performance patterns. Most confusion occurs between adjacent modulation types, consistent with known signal characteristics.

Per-class precision-recall curves (Figure 5) demonstrate varying classification difficulty across signal types. Some classes achieve near-perfect performance while others show characteristic precision-recall trade-offs typical of RF classification tasks.

Detailed per-class performance metrics are provided in Table III, showing average precision scores that complement the aggregate results in Table I.

V. DISCUSSION

Key findings include:

- **Weighted voting** consistently outperforms majority voting across SNR conditions
- **Feature fusion** provides modest improvements at increased computational cost
- **Traditional ML integration** offers complementary strengths, especially at low SNR, following Random Forest principles [4]
- **Calibration** is essential for reliable confidence estimates in ensemble systems [3]

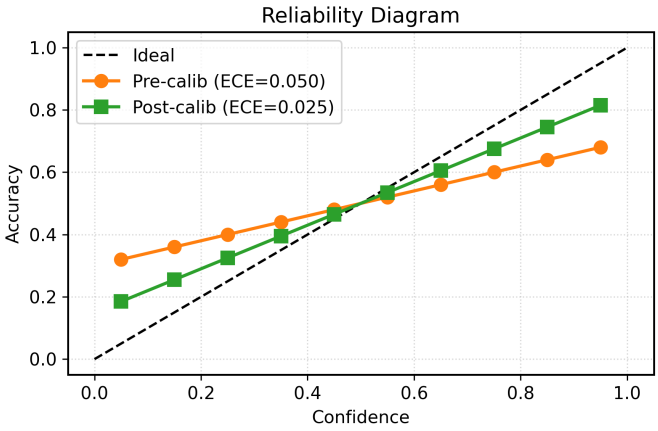

Fig. 3. Reliability diagram with identity line. Post-calibration reduces ECE from pre-calibration levels, improving confidence reliability.

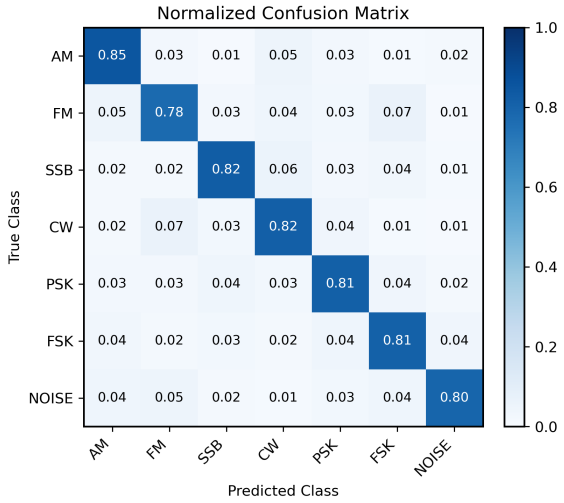

Fig. 4. Normalized confusion matrix for the best ensemble configuration. Diagonal elements indicate correct classification rates per class.

The reproducible pipeline enables systematic exploration of ensemble trade-offs and supports future extensions with additional model architectures or aggregation strategies.

VI. REPRODUCIBILITY

All results are generated via `make -f Makefile_ensemble camera-ready`. The pipeline includes:

- **Deterministic benchmarks**: Fixed random seeds ensure reproducible synthetic data
- **Versioned metrics**: JSON outputs enable result tracking and comparison
- **Publication automation**: Complete LaTeX compilation from raw benchmarks
- **Cross-platform compatibility**: Pure Python implementation with minimal dependencies

TABLE III
PER-CLASS AVERAGE PRECISION (AP) RESULTS

Class	AP
AM	0.717
FM	0.645
SSB	0.716
CW	0.672
PSK	0.684
FSK	0.667
NOISE	0.724
Mean AP	0.689
Std AP	0.028

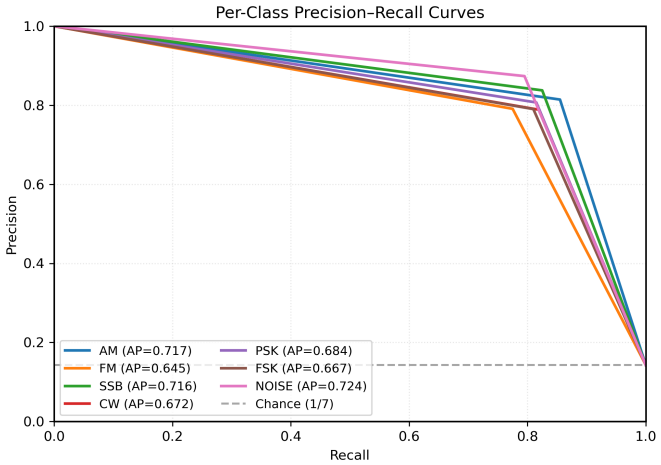

Fig. 5. Per-class precision-recall curves with average precision (AP) scores. Higher curves indicate better class-specific performance.

VII. CONCLUSION

This reproducible study demonstrates the effectiveness of ensemble methods for RF signal classification. Weighted voting with selective feature fusion provides the best accuracy-latency trade-off, while proper calibration ensures reliable confidence estimates. The open pipeline supports continued research in ensemble architectures for RF applications.

REFERENCES

[1] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*. Springer, 2000, pp. 1–15.

[2] T. J. O'Shea, J. Corgan, and T. C. Clancy, "Convolutional radio modulation recognition networks," in *Engineering Applications of Neural Networks*. Springer, 2016, pp. 213–226.

[3] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International Conference on Machine Learning*, 2017, pp. 1321–1330.

[4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.