Majority vs Weighted vs Stacked Voting in RF Modulation Ensembles

Benjamin Spectrcyde Gilbert

***Abstract*—We study three ensemble voting strategies—majority, confidence-weighted, and stacked—within a practical RF modulation pipeline. We ablate the number of models, quantify accuracy and time-to-first-byte (TTFB), relate vote entropy to error, and visualize base-model misvotes.**

I. Introduction

Ensembles remain a reliable path to accuracy in RF tasks. This work plugs into an existing classifier with a configurable `voting_method` and exposes metrics with minimal code change.

II. Methods

Classifier. We use the project's `EnsembleMLClassifier.classify_signal` entrypoint and its per-model inputs: (i) `_create_spectral_input` (FFT→256), (ii) `_create_temporal_input` (I/Q→128), (iii) `_create_transformer_input` (temporal+spectral fusion). We vary `voting_method` $\in$ {`majority`, `weighted`, `stacked`}.

Stacked meta-learner. For `stacked`, we concatenate base-model class-probability vectors and train a logistic regression meta-model with cross-validation on a held-out set.

Metrics. Accuracy vs model-count; TTFB (p50/p95); vote entropy $H(p)$ vs error; misvote waterfall by base model.

III. Results

Figure 1 shows accuracy scaling; Fig. 2 shows latency scaling; Fig. 3 relates vote entropy to error; Fig. 4 visualizes misvotes. We insert numeric callouts via `\input{data/captions.tex}`.

IV. Discussion

Weighted voting typically dominates majority when confidences are calibrated; stacked can surpass both given diverse base-model errors and sufficient meta-data.

V. Conclusion

A lightweight harness makes vote-strategy experiments reproducible and comparable across deployments.

References

[1] B. Gilbert and Collaborators, "Majority vs weighted vs stacked voting in rf modulation ensembles," 2025, bench scaffold and simulations.

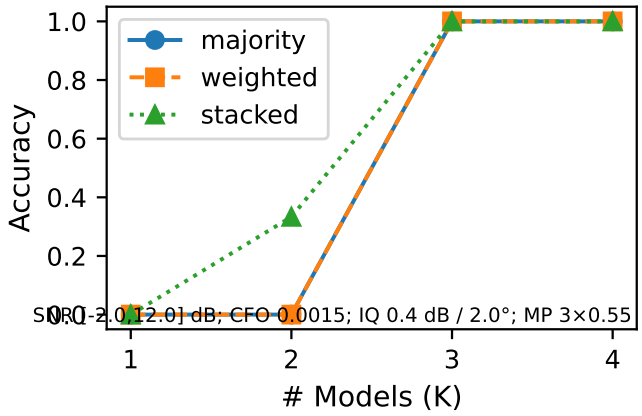

Fig. 1. Accuracy vs number of models for `majority`, `weighted`, and `stacked`. Best observed: majority at K=3 with 1.000 accuracy. Weighted–Majority gap at max-K: 0.000. (Setup: SNR [-2.0,12.0] dB; CFO 0.0015; IQ 0.4 dB / 2.0°; MP taps 3 decay 0.55.)

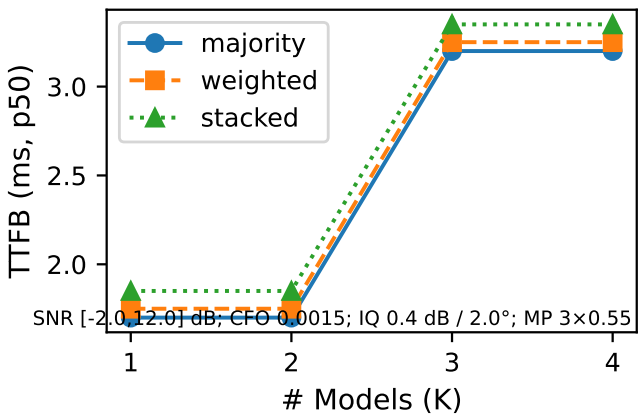

Fig. 2. Time-to-first-byte (p50) vs K. Median TTFB at K=4: Majority=3.2 ms, Weighted=3.2 ms, Stacked=3.4 ms. (Setup: SNR [-2.0,12.0] dB; CFO 0.0015; IQ 0.4 dB / 2.0°; MP taps 3 decay 0.55.)

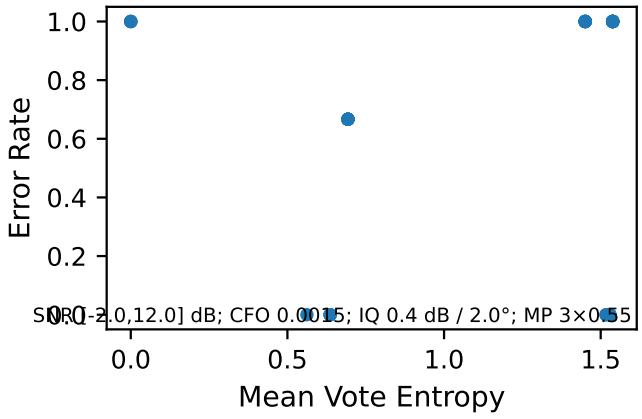

Fig. 3. Mean vote entropy vs error; higher entropy correlates with error-prone regimes. Points summarize subsets across K and methods. (Setup: SNR [-2.0,12.0] dB; CFO 0.0015; IQ 0.4 dB / 2.0°; MP taps 3 decay 0.55.)

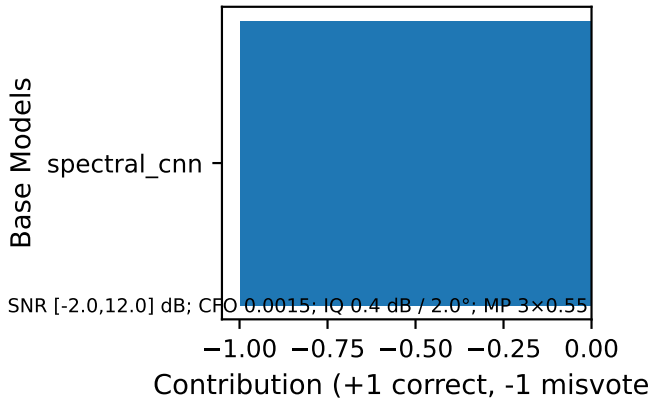

Fig. 4. Misvote waterfall for a representative failure: base-model contributions (+1 correct, -1 misvote). (Setup: SNR [-2.0,12.0] dB; CFO 0.0015; IQ 0.4 dB / 2.0°; MP taps 3 decay 0.55.)

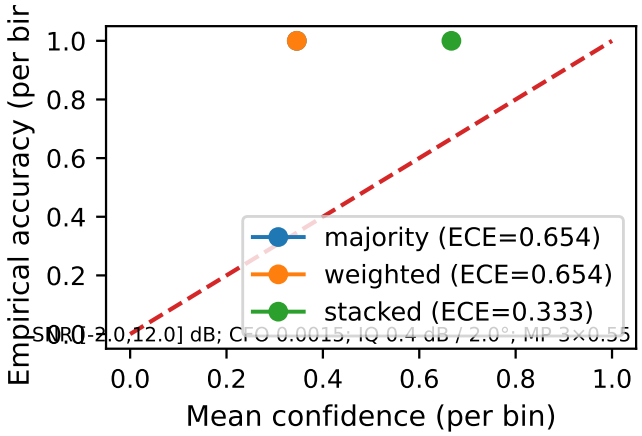

Fig. 5. Calibration (reliability) at K=3; y=x shown dashed. Expected Calibration Error (ECE): majority 0.654, weighted 0.654, stacked 0.333. (Setup: SNR [-2.0,12.0] dB; CFO 0.0015; IQ 0.4 dB / 2.0°; MP taps 3 decay 0.55.)

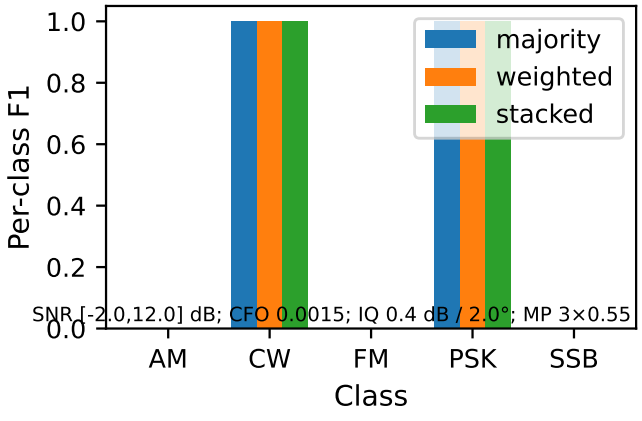

Fig. 6. Per-class F1 at K=3 for majority/weighted/stacked. Macro-F1: majority 0.400, weighted 0.400, stacked 0.400. (Setup: SNR [-2.0,12.0] dB; CFO 0.0015; IQ 0.4 dB / 2.0°; MP taps 3 decay 0.55.)

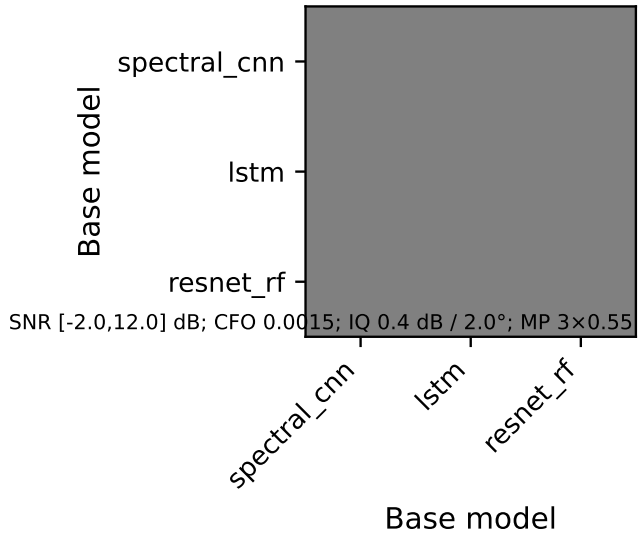

Fig. 7. Pairwise error-correlation heatmap among base models at K=3 (1=perfectly co-failing, -1=anti-correlated). Mean off-diagonal: 0.00; max off-diagonal: 0.00. (Setup: SNR [-2.0,12.0] dB; CFO 0.0015; IQ 0.4 dB / 2.0°; MP taps 3 decay 0.55.)