IOCU
2026
Conference

The 3rd International Online Conference on Universe

4–6 March 2026 | Online

MDPI

A deterministic Mini-Run Benchmark for Robust and Trustworthy Classification of Gravitational-Wave Glitches

Rudhresh Manoharan[1], *Brian Phillips[1]*, Tanish Chettiar[2], Gerald Cleaver[1]
Baylor University[1] Yale University[2]

INTRODUCTION & AIM

Transient noise artifacts in gravitational-wave detectors, commonly referred to as glitches, pose a significant challenge for reliable signal detection. Glitches can obscure or mimic true astrophysical signals and may bias inferred source parameters. Although traditional computational methods exist to mitigate glitches, these approaches are often computationally expensive and difficult to scale.

Machine learning offers a potentially more efficient alternative for automated glitch classification. Building on prior work, this project compares the performance of convolutional neural networks (CNNs) and vision transformers (ViTs) in classifying glitch types from spectrogram data.

However, achieving high accuracy alone is insufficient for deployment in gravitational-wave pipelines. This project therefore aims to develop a robust and trustworthy machine learning pipeline for glitch classification. Robustness demands stable performance under distribution shifts, while trustworthiness requires well-calibrated confidence estimates. These properties will become even more crucial as next-generation gravitational-wave detectors come online.

This project constructs a deterministic mini-benchmark using 10 balanced glitch classes, split into 1400/460/950 training, validation, and test samples. Stress test sets apply controlled signal or noise distortions at severity levels {0, 2, 4}, where 0 corresponds to no distortion, 2 to moderate distortion, and 4 to severe distortion, imitating real world glitch types the pipeline may be exposed to.

METHOD

Two baseline classifiers are evaluated. One is a convolutional neural network (CNN) and the other a Tiny Vision Transformer (ViT). This project looks to compare the results of these two models. All trials use the same fixed train/validation/test split with a deterministic seed to ensure reproducibility.

To simulate realistic detector variability, stress test sets are generated by applying signal strength modifications, noise perturbations, and distortion-like transformations. Each stress type is applied at severity levels {0, 2, 4}, where 0 corresponds to no stress, 2 to moderate stress, and 4 to severe stress. This will quantify how sensitive the models are to new glitch types, a situation that is likely to occur with new gravitational-wave detectors becoming operational.

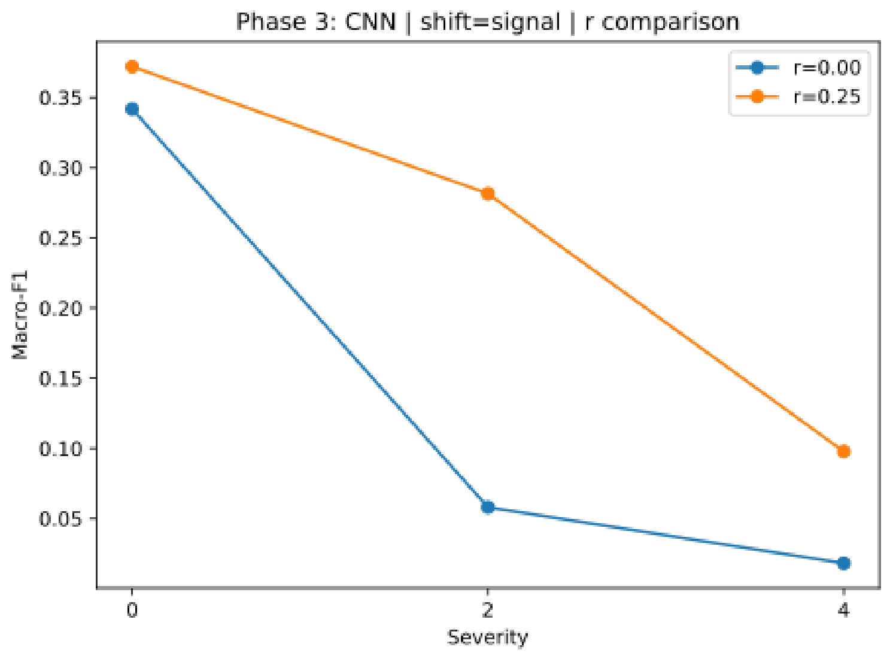
Although gravitational-wave glitches are available in the Gravity Spy dataset, the number of labeled samples per class remains limited. The lack of samples would make training on the previously mentioned models difficult. To account for this, synthetic training examples generated using generative adversarial networks (GANs) are introduced to expand and balance the training set. This also allows controlled evaluation of how additional data influences robustness under distribution shift. GANs is trained of the gravity spy dataset.

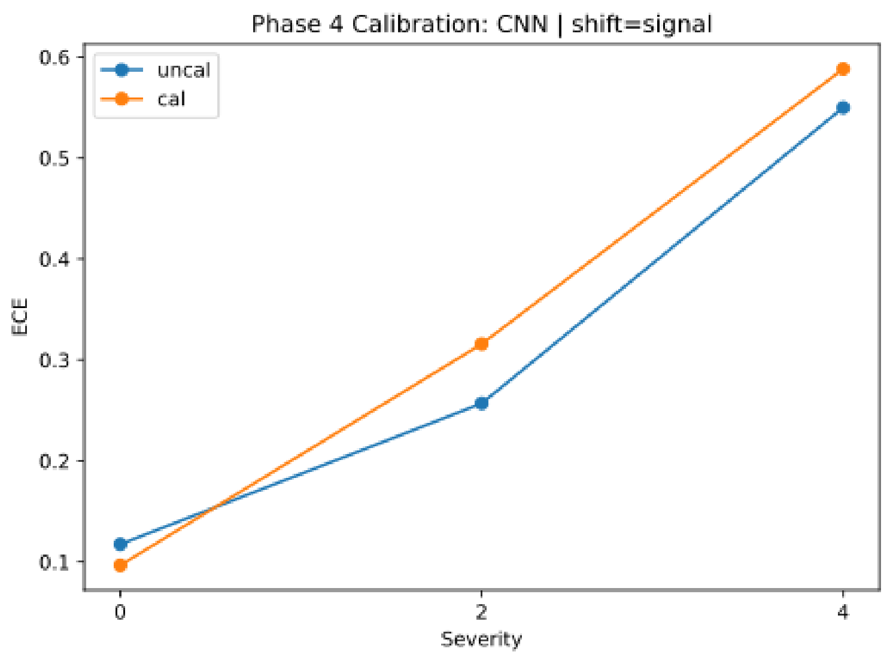
Because operational deployment requires reliable confidence reporting, model outputs are calibrated using temperature scaling. The temperature parameter is fit exclusively on validation data and evaluated using Expected Calibration Error (ECE) under both clean and shifted conditions.

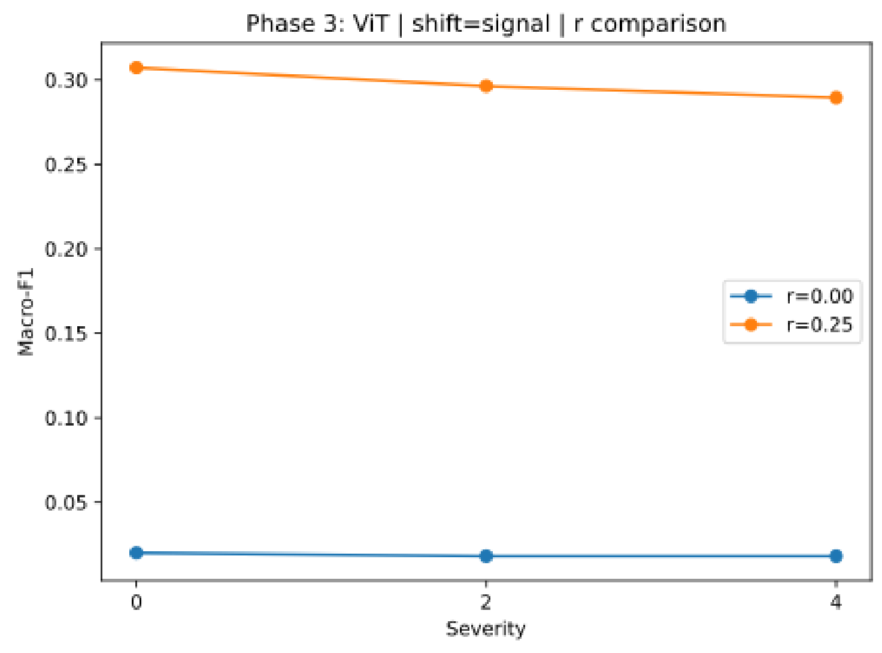
The pipeline consists of four structured phases:

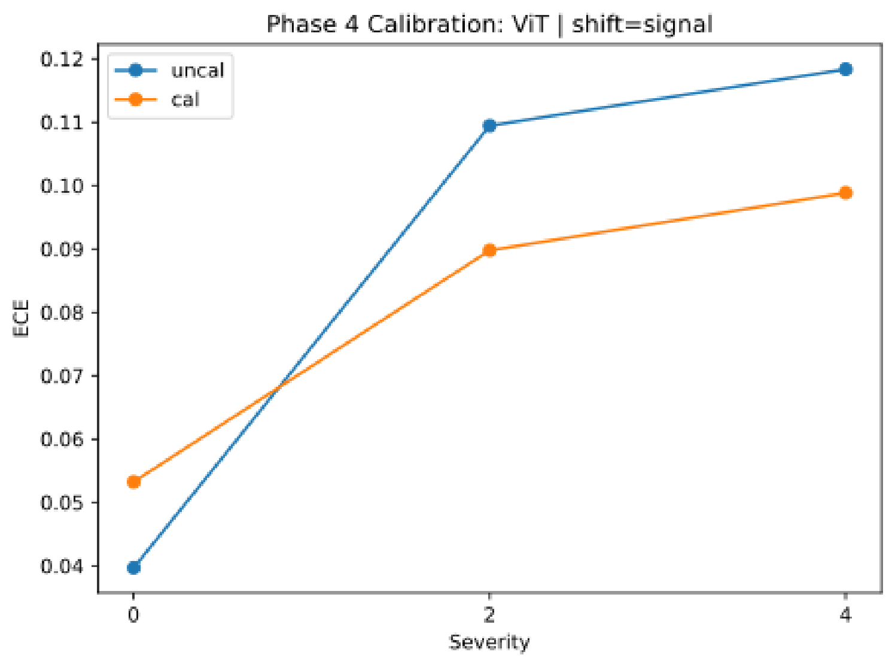
1. Baseline training and clean evaluation
2. Stress testing under controlled transformations
3. Synthetic augmentation analysis
4. Calibration under distribution shift is applied after training

RESULTS & DISCUSSION









The top figure shows Macro-F1 performance under signal shift for the CNN baseline. Performance degrades with increasing severity, confirming that signal visibility changes represent a dominant failure mode. Synthetic augmentation ($r = 0.25$) improves robustness across all severity levels.

The figure below that presents calibration behavior for the CNN under the same shift conditions. Expected Calibration Error (ECE) increases under distribution shift, indicating overconfident predictions as signal distortion grows. Temperature scaling reduces ECE on clean data and partially mitigates miscalibration under shift. This highlights the importance of a confidence report, especially as new gravitational-wave detectors come online, and new glitches appear in the dataset.

A similar robustness trend is observed for the Tiny ViT. Without augmentation, performance is severely degraded in the low-data regime. However, synthetic augmentation does improve Macro-F1 scores, suggesting that transformer-based models are more data-dependent in this setting.

Calibration trends for the ViT reveal that although temperature scaling reduces ECE, calibration remains sensitive to signal shift severity. Across both architectures, signal shift produces the most significant performance degradation, underscoring the importance of robustness evaluation prior to deployment in gravitational-wave pipelines.

CONCLUSION

In the limited-data regime, CNNs demonstrate stronger baseline performance. However, synthetic augmentation substantially improves transformer-based models, indicating that ViTs are more data-dependent but viable with sufficient training examples. The graphs indicate ViT is undertrained on this dataset.

Signal shift emerges as the dominant failure mode across architectures, highlighting the sensitivity of classifiers to variations in signal visibility. Calibration analysis further shows that while temperature scaling improves confidence estimates on clean data, reliability degrades under distribution shift.

Together, these results demonstrate that accuracy alone is insufficient for deployment. Robustness and calibrated confidence must be jointly evaluated before integrating machine learning classifiers into gravitational-wave monitoring pipelines.

This mini-benchmark establishes a controlled and reproducible foundation for scaling to larger datasets and more comprehensive robustness studies in preparation for next-generation gravitational-wave detectors.

FUTURE WORK / REFERENCES

Zevin et al., *Gravity Spy: Integrating Advanced LIGO Detector Characterization, Machine Learning, and Citizen Science*, CQG (2017).