Astrophysical or Terrestrial: Machine learning classification of gravitational-wave

Avyukt Raghuvanshi, Michael Coughlin, Andrew Toivonen
University of Minnesota | School of Physics and Astronomy

Abstract

Rapid follow-up of gravitational-wave candidates is essential for multi-messenger astronomy, but terrestrial “glitches” generate a handful of false alarms each observing run. Current significance metrics (sky localization, FAR, p_{astro}) come from individual pipelines and can disagree. We introduce a multi-pipeline machine-learning classifier that ingests signal-vs-noise and coherence Bayes factors (LOGBSN, LOGBCI) from BAYESTAR[1] plus preferred SNR and FAR from four low-latency searches. Trained on the 50 000-injection Mock Data Challenge[2] and tested on real O3 events, it significantly reduces false positives relative to p_{astro} alone. Our score can guide more reliable electromagnetic and neutrino follow-up.

Data Sets

- Training (MDC[2]): 40 days of O3 strain with $5{\times}10^4$ simulated BNS/NSBH/BBH injections yields ~5 600 true positives and ~1 800 noise triggers.
- Testing (O3 real-events): Confirmed “astrophysical” candidates from the GWTC catalog versus formally retracted alerts labeled “terrestrial.”

This split ensures robust evaluation on both simulated and genuine alert streams.

Key Features

From each low-latency search (GstLAL[3], MBTA[4], PyCBC[5], SPIIR[6]) we extract:
- Preferred SNR and log FAR of the trigger passing the FAR threshold.
- LOGBSN (signal vs. noise Bayes factor) and LOGBCI (coherent vs. incoherent Bayes factor) from the BAYESTAR[1] sky-map FITS.

If a pipeline does not recover an event, its four features are flagged “–1,” encoding non-detection and naturally down-weighting that pipeline’s contribution.

Feature Importance

Permutation tests on the RF model rank features :

1. LOGBSN: $\log_{10}$[P(data | signal) ÷ P(data | noise)] -Bayes factor comparing signal vs. noise hypotheses
2. LOGBCI: $\log_{10}$[coherent-signal BSN ÷ incoherent-signal BSN] - Coherence test for multi-detector consistency
3. Preferred SNR: Peak signal-to-noise ratio from the top pipeline - measure of signal strength above detector noise
4. No. of Pipelines: Number of independent searches that recovered the event - cross-pipeline validation
5. FAR: (false-alarm rate) of the preferred trigger - expected rate of noise events with similar properties

This ordering highlights the key role of coherent Bayes odds and cross-pipeline agreement in distinguishing real gravitational wave signals.

Model Training and Architecture

We use a two-stage ensemble:
1. Per-pipeline sub-models (KNN, RF with 300 trees at max depth 5, and a 7-layer MLP with ReLU activations) each output a preliminary “astro” score or –1.
2. A superevent aggregator ingests the four pipeline scores to produce the final probability.

Hyperparameters were tuned via 10-fold cross-validation on the MDC[2]. Class imbalance is addressed by KDE-based resampling of rare terrestrial cases.

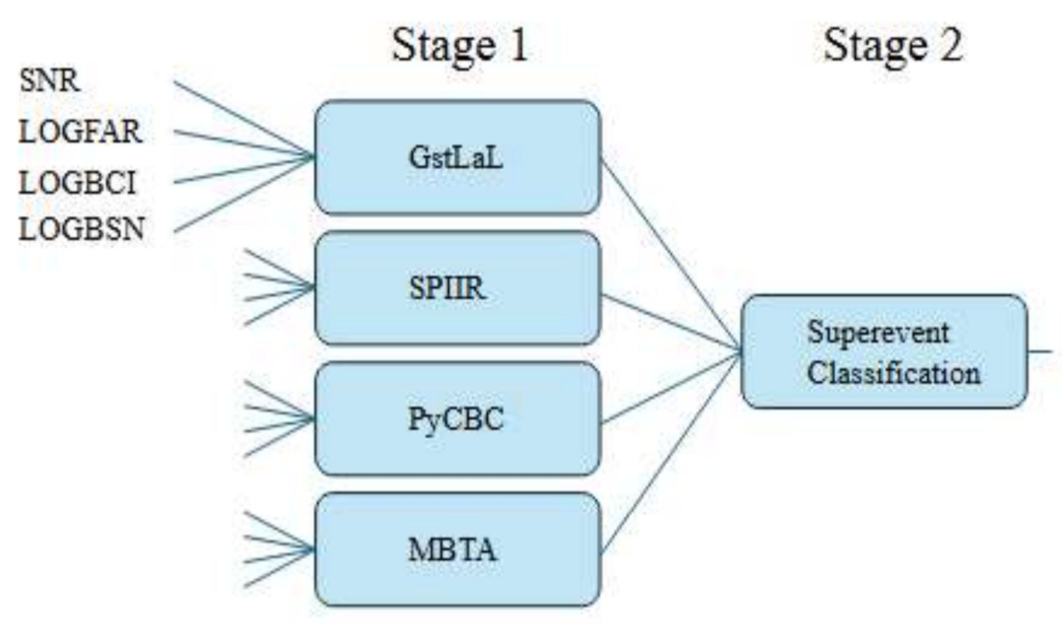

Fig.1 Diagram of model architecture. If a pipeline is not used in a superevent, a value of -1 is sent to super-event classification.

Discussion and Future Work

Limitations:

Without direct strain-data quality inputs, loud detector glitches can occasionally mimic real signals. MDC[2] FAR values are biased by the high injection density, reducing FAR’s discriminative weight.

Next steps:

- Run a pre-O5 MDC[2] using pure O4 noise (no injections) to correct FAR calibration.
- Incorporate spectrogram-based or conditional glitch-flag metrics into feature set.
- Deploy this ML classifier alongside p_{astro} and GWSkyNet in real-time O5 and beyond alert streams.

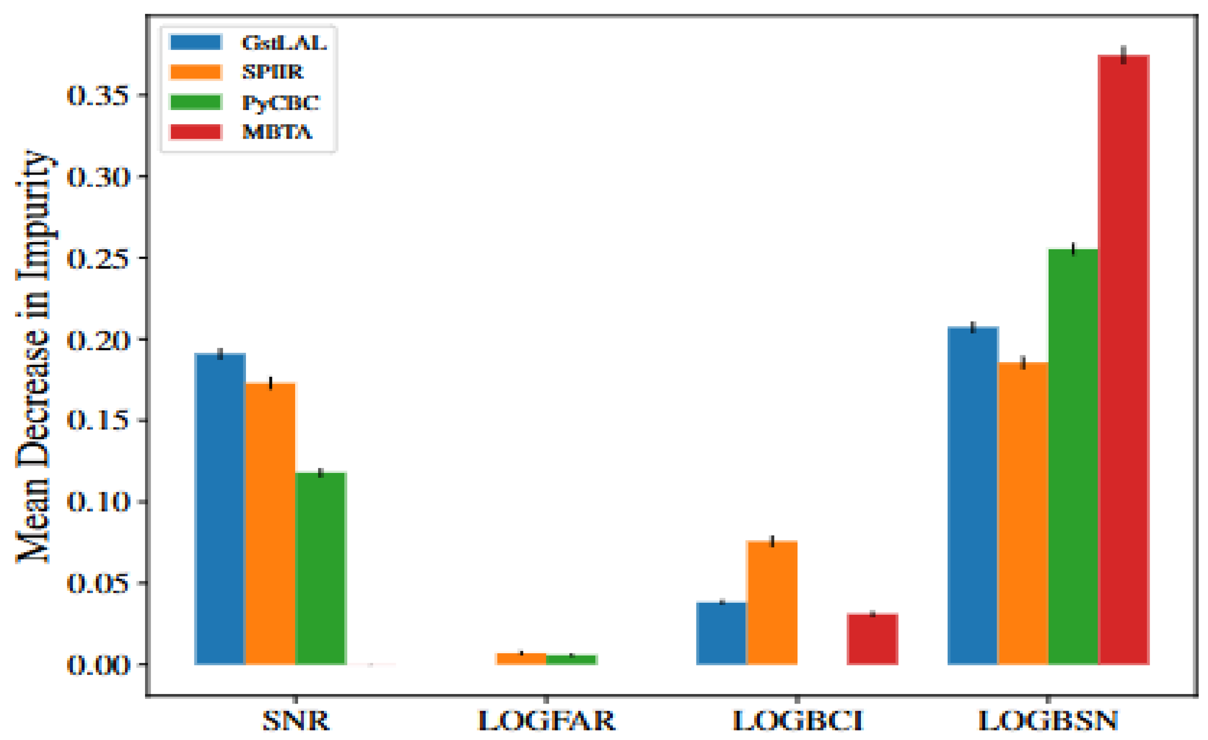

Fig.2 Permutation importance for the random forest classifier for each pipeline.

Results

- MDC[2] (train): RF/NN AUC = 0.96; KNN AUC = 0.93; accuracy ≃ 0.91.
- O3 (test): RF/NN AUC ≈ 0.94; KNN AUC ≈ 0.88; accuracy ≃ 0.88.

Compared to p_{astro} baselines, our RF/NN reduce false-positive glitch acceptance by ~15% while maintaining high detection efficiency for confirmed astrophysical events.

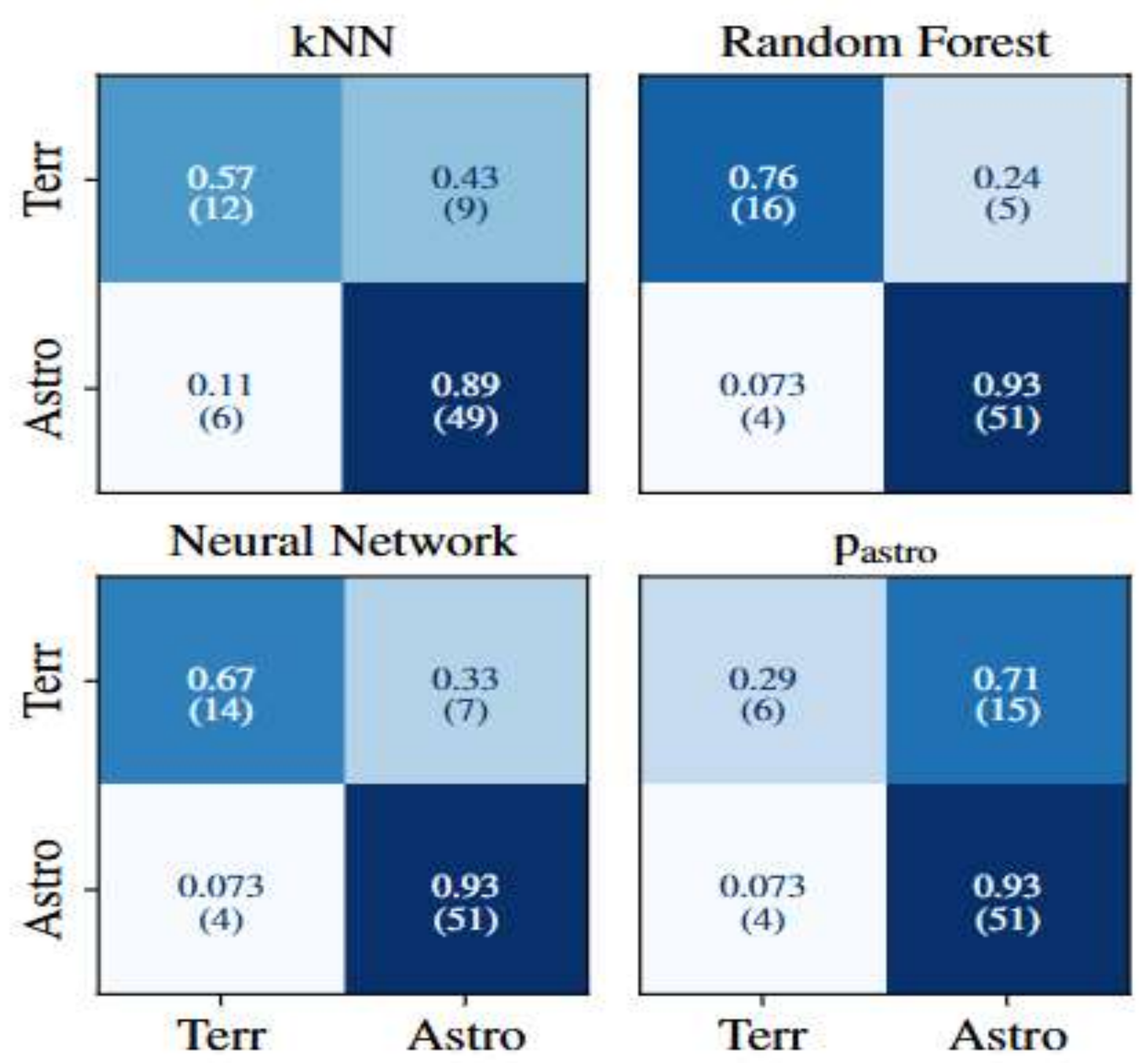

Fig. 3 Confusion matrix of model performance on real GW events from the O3 dataset.

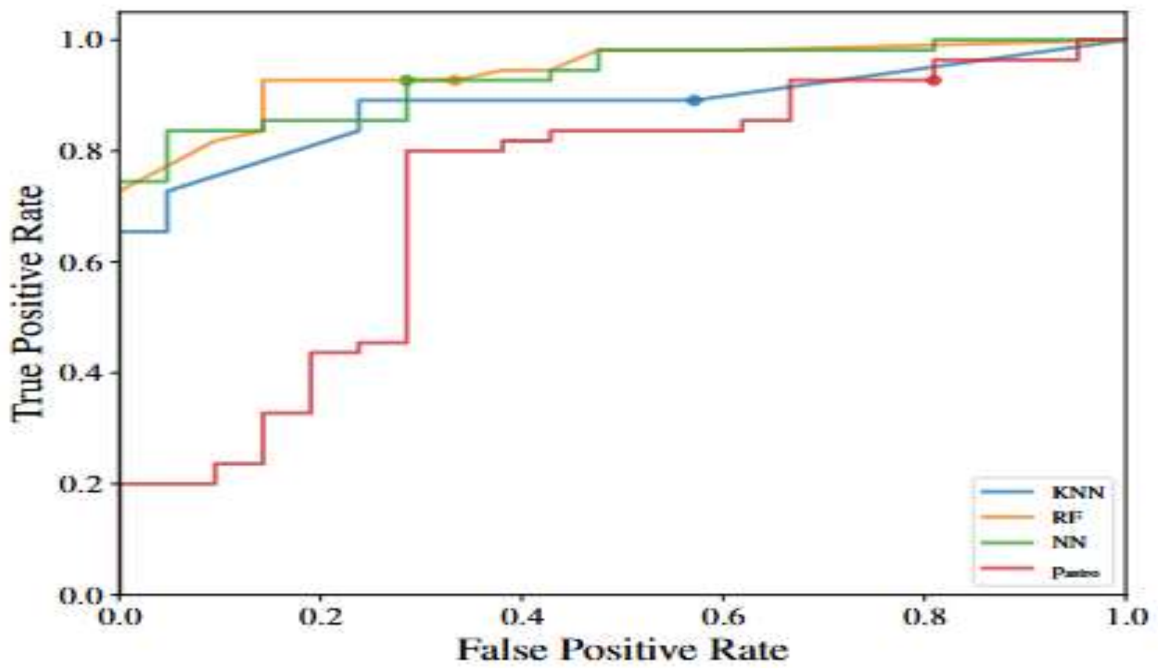

Fig.4 ROC curves obtained from the testing the on O3dataset.

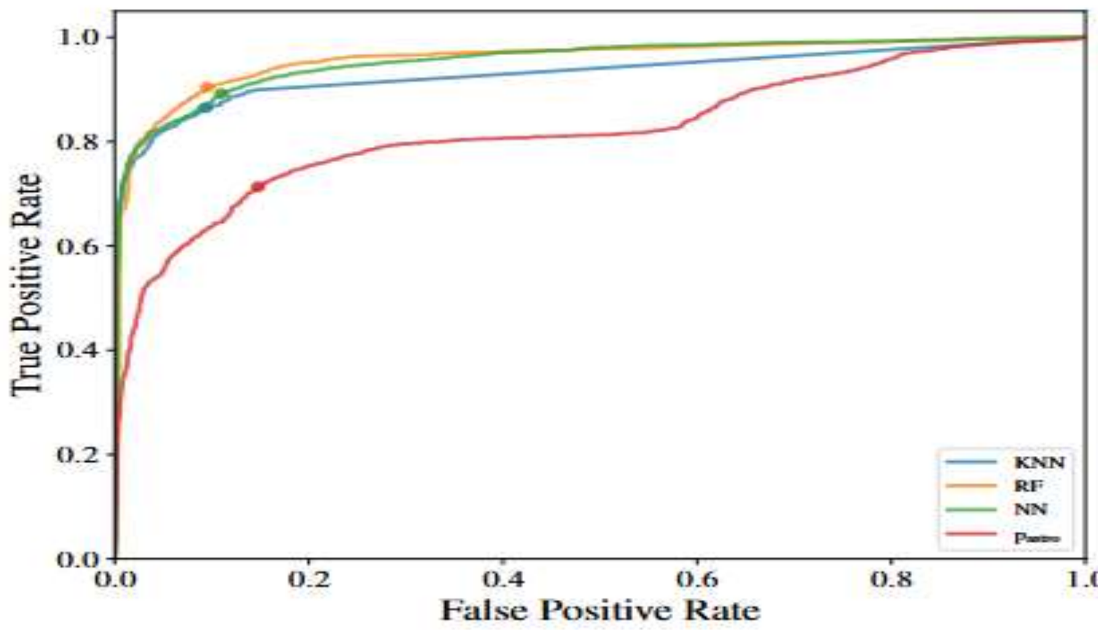

Fig.5 ROC curves obtained from the MDC[2].

Acknowledgements

We thank our mentors Michael Coughlin and Andrew Toivonen for guidance, and S. Sharma-Chaudhary for internal review. This work builds on the LVK Mock Data Challenge [2] and the GstLAL [3], PyCBC [4], MBTA [5], SPIIR [6], and BAYESTAR [1] pipelines.
Support was provided by NSF grants PHY-2117997, PHY-2308862, PHY-2409481, and the LIGO Laboratory.

References:
[1] S. Sharma-Chaudhary et al., “Mock Data Challenge design,” 2023.
[2] J. Messick et al., “The GstLAL Search Analysis Methods,” Phys. Rev. D 93, 042001 (2016).
[3] T. Dal Canton et al., “PyCBC: Open-source gravitational-wave analysis,” Phys. Rev. D 102, 063029 (2020).
[4] C. Aubin et al., “MBTA: Multi-Band Template Analysis,” Class. Quantum Grav. 37, 215009 (2020).
[5] M. Chu et al., “SPIIR: Low-latency gravitational-wave search,” Phys. Rev. D 102, 024076 (2020).
[6] L. Singer et al., “BAYESTAR: Rapid sky localization,” Astrophys. J. Lett. 829, L15 (2016).

Contact & code: raghu119@umn.edu