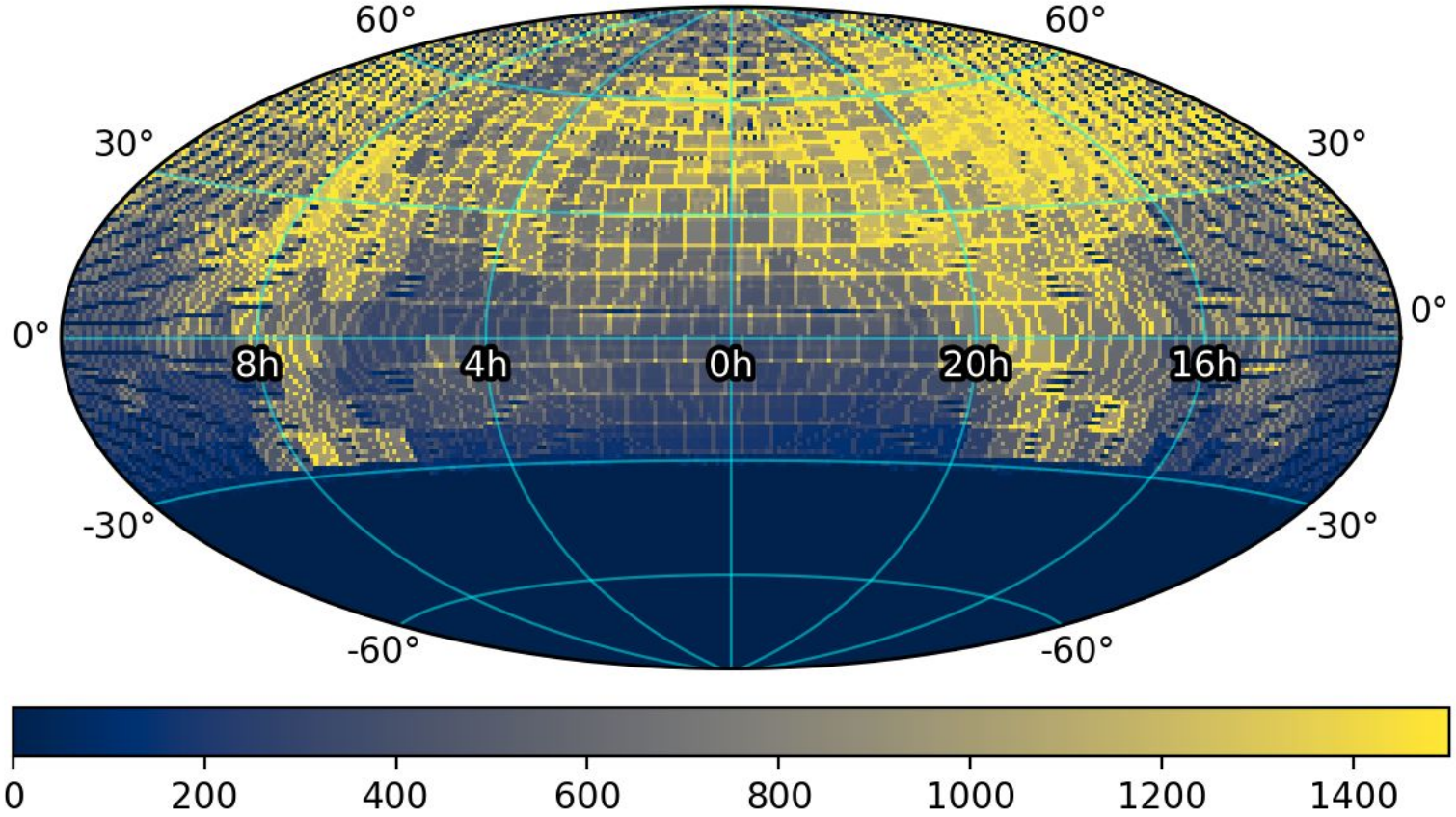
Deep Learning Applications in Astronomy



AY 119 · 2026

Ashish Mahabal — Caltech

Surveys, pipelines, decisions

From theory to practice

19 papers, multiple domains, mostly Caltech-connected.

Recurring theme: representation tricks matter more than fancy architectures.

The data deluge

ZTF: ~1M alerts per night → LSST/Roman coming

TESS: 200,000+ stars at 2-min cadence.

LIGO / Virgo / KAGRA: continuous strain, glitch-dominated.

CHIME / DSA-110: thousands of FRB candidates per day.

There is no human-only path forward. DL is the production layer.

Roadmap

1. Images — real/bogus, asteroids, subtraction-free
2. Light curves — DMDT, SCOPE, TESS
3. Spectra — SNIascore, CCSNscore
4. Gravitational waves — DMDT in LIGO + the GWSkyNet family
5. Radio transients — Frabjous (FRBs)
6. Language models — AstroAlertBench

1. Images

ZTF and the alert stream

Difference imaging: subtract a reference image, look for what's new.

Most candidates are artifacts: CR hits, optical ghosts, bad subtractions.

Allows a fast, high-recall gate before science pipelines spend time.

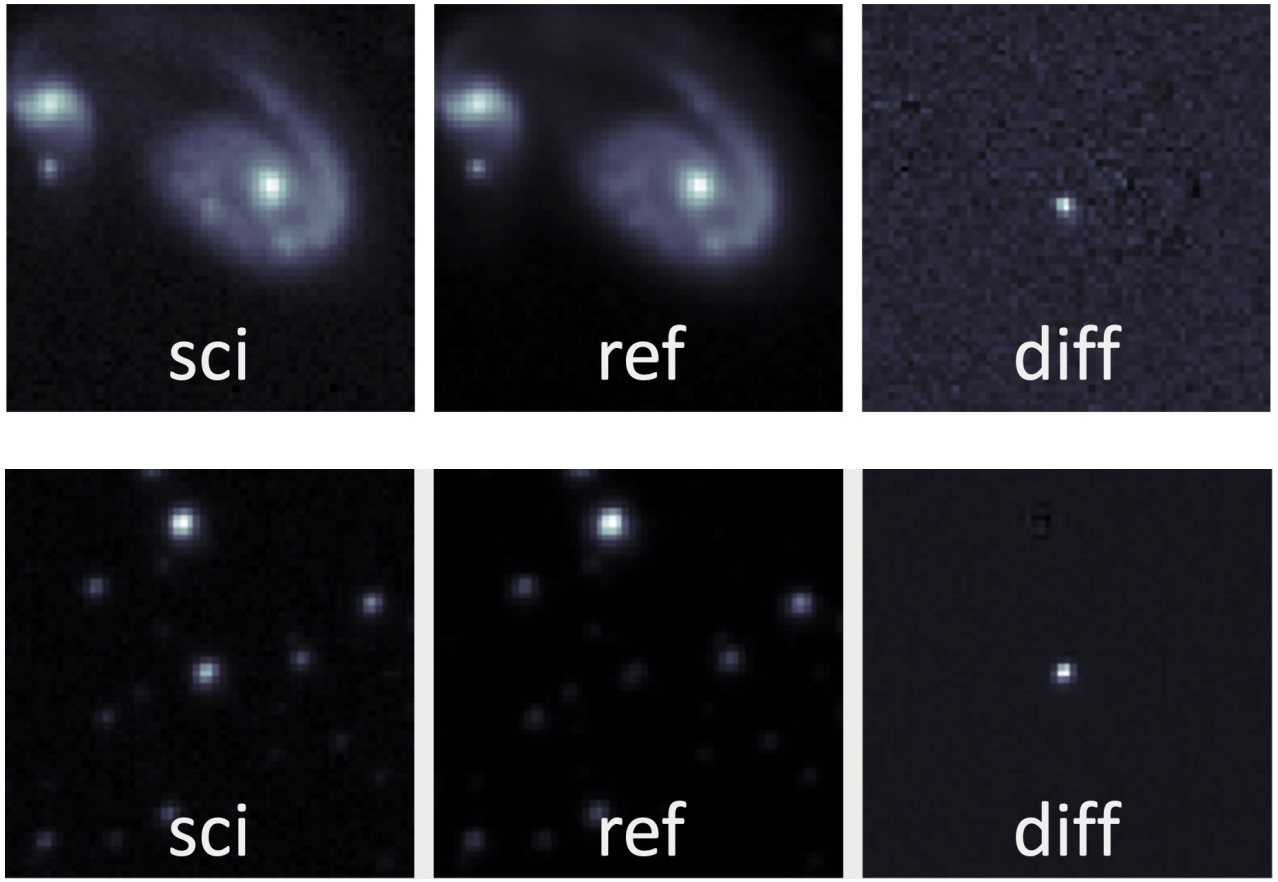
BRAAI — real or bogus (Duev+ 2019)

VGG6 CNN on (science, reference, difference) image triplets.

Trained on ~32k vetted ZTF examples; <20 min on one GPU.

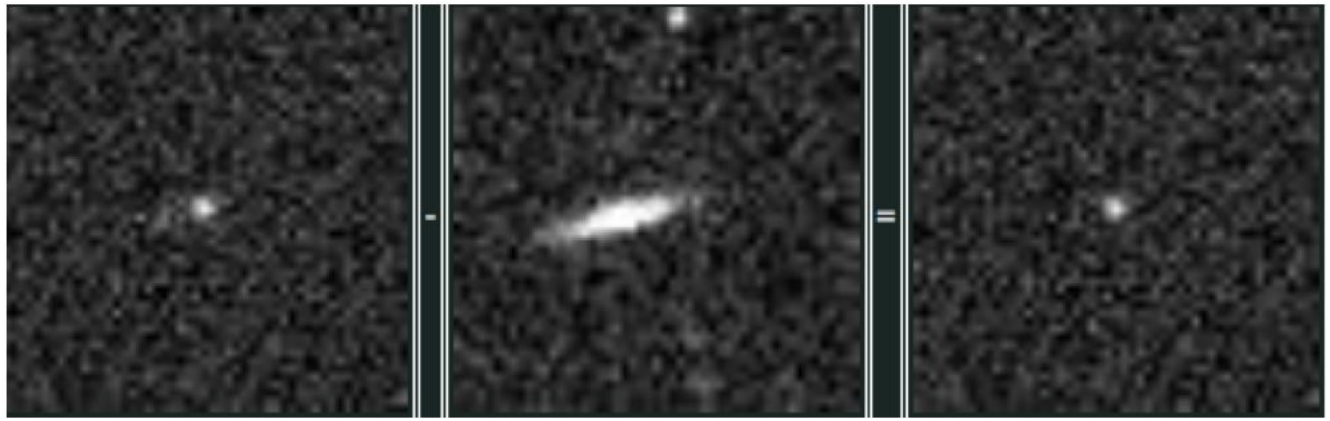
At threshold 0.5: 1.1% false-negative + 2.9% false-positive on test.

Every ZTF science pipeline sits behind this gate.

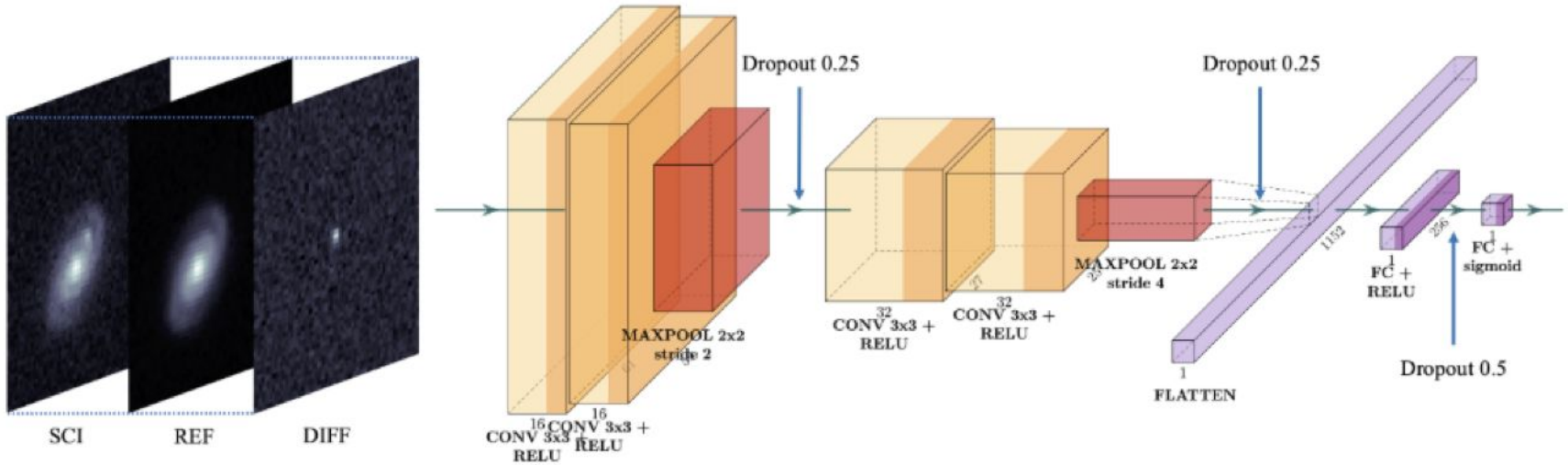


Real / bogus triplets — arXiv:1907.11259

ZTF Real-Bogus separation ('braai')



Sci **Ref** **Diff**



https://github.com/dmitryduev/braai

Duev, Mahabal, … arXiv 1907.11259

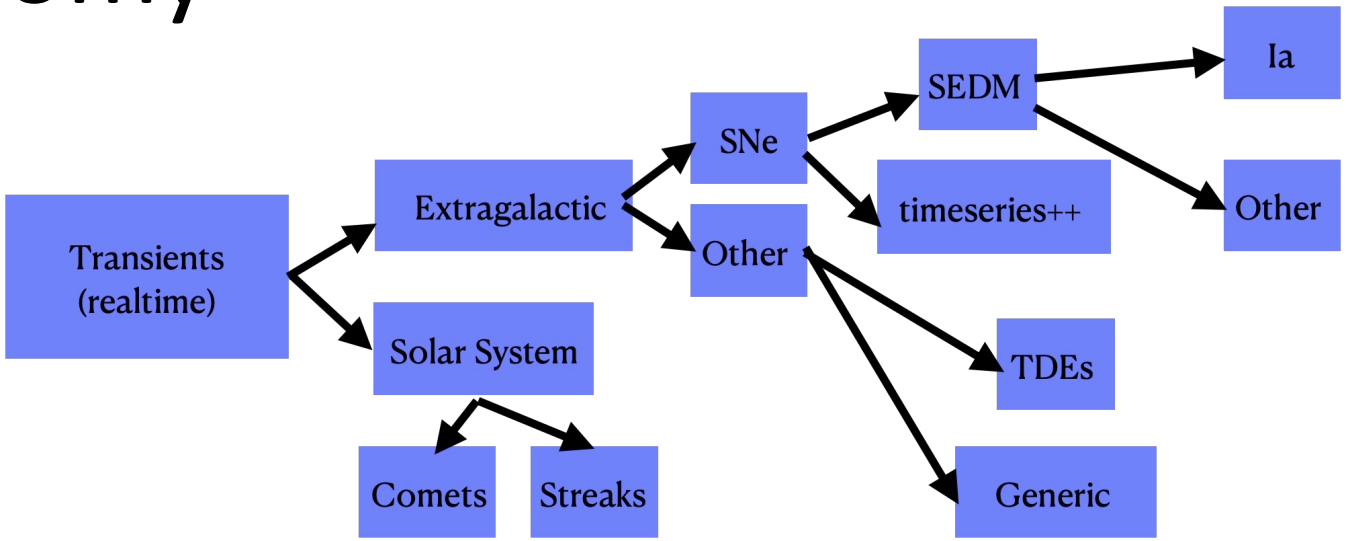
By area of astronomy

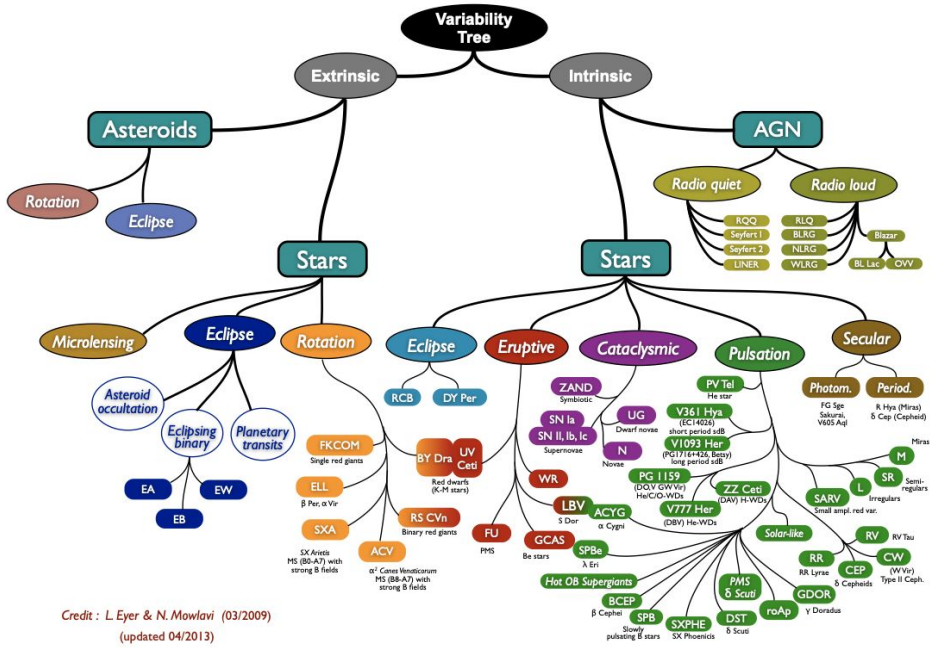
Transients

- Detection
- Classification

Variable stars

Solar System





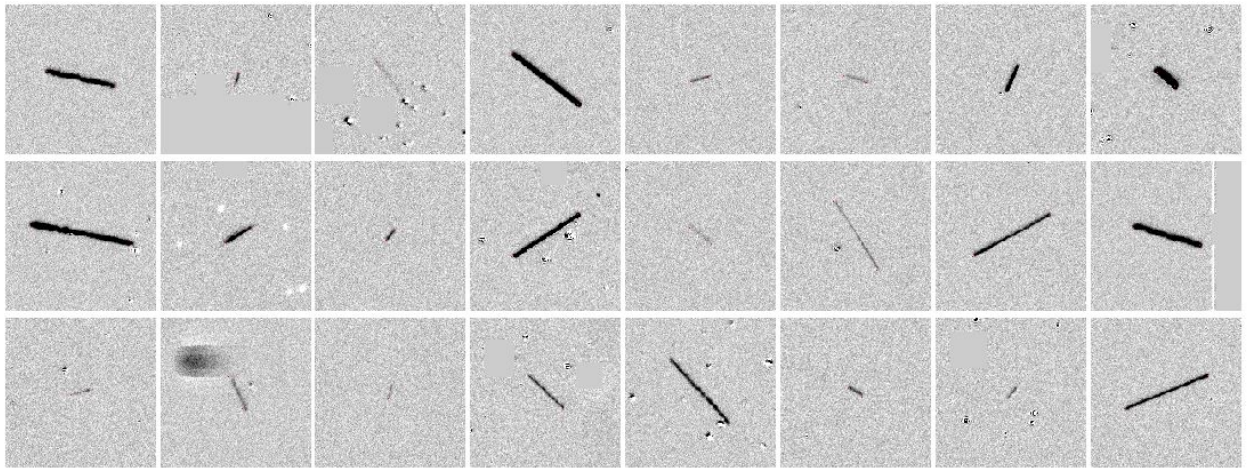
DeepStreaks — fast-moving asteroids (Duev+ 2019)

30-second exposures leave streaks for fast NEOs.

Trained on 1k real + 8k synthetic streaks + 6k bogus images.

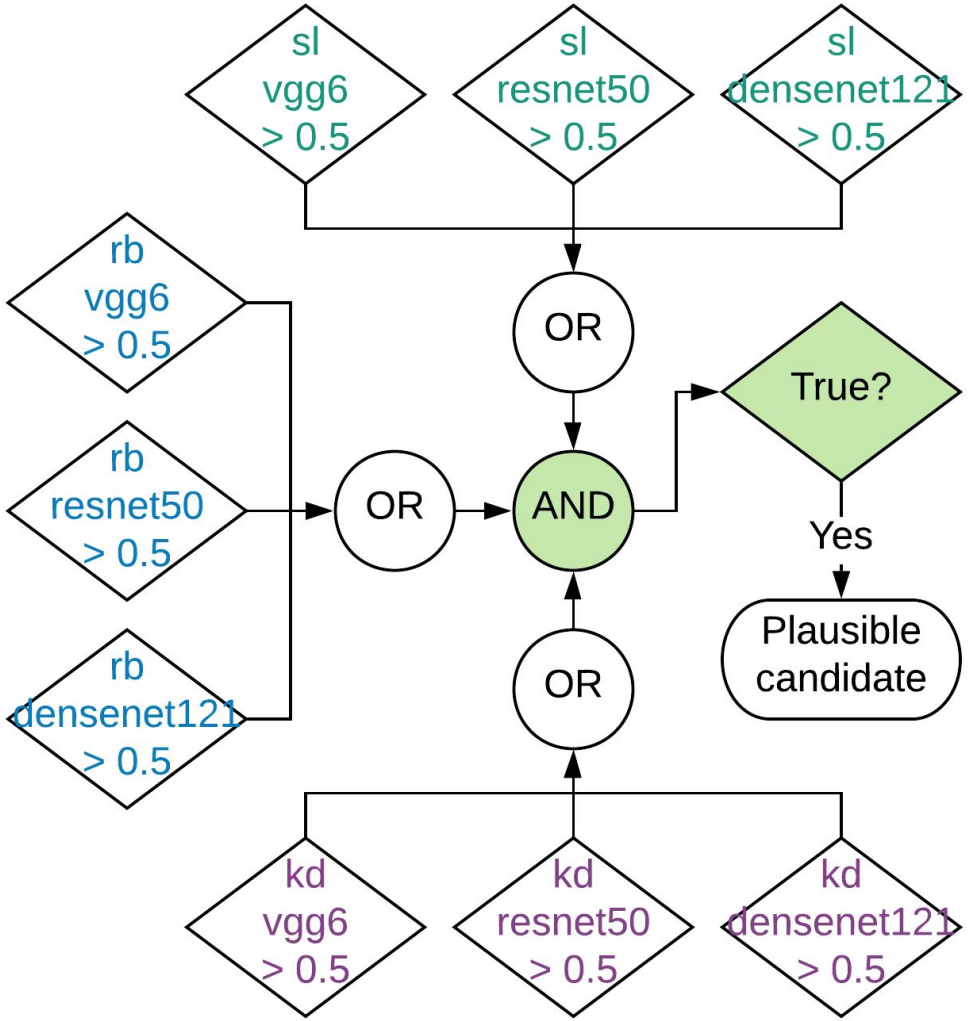
Ensemble: VGG6 + ResNet50 + DenseNet121, per task.

96–98% TPR; 50× fewer false positives than the RF baseline.



Real ZTF streaks — arXiv:1904.05920

DeepStreaks — pipeline architecture



Three-classifier ensemble decision flow — arXiv:1904.05920

TransiNet — subtraction-free detection (Sedaghat+ 2017)

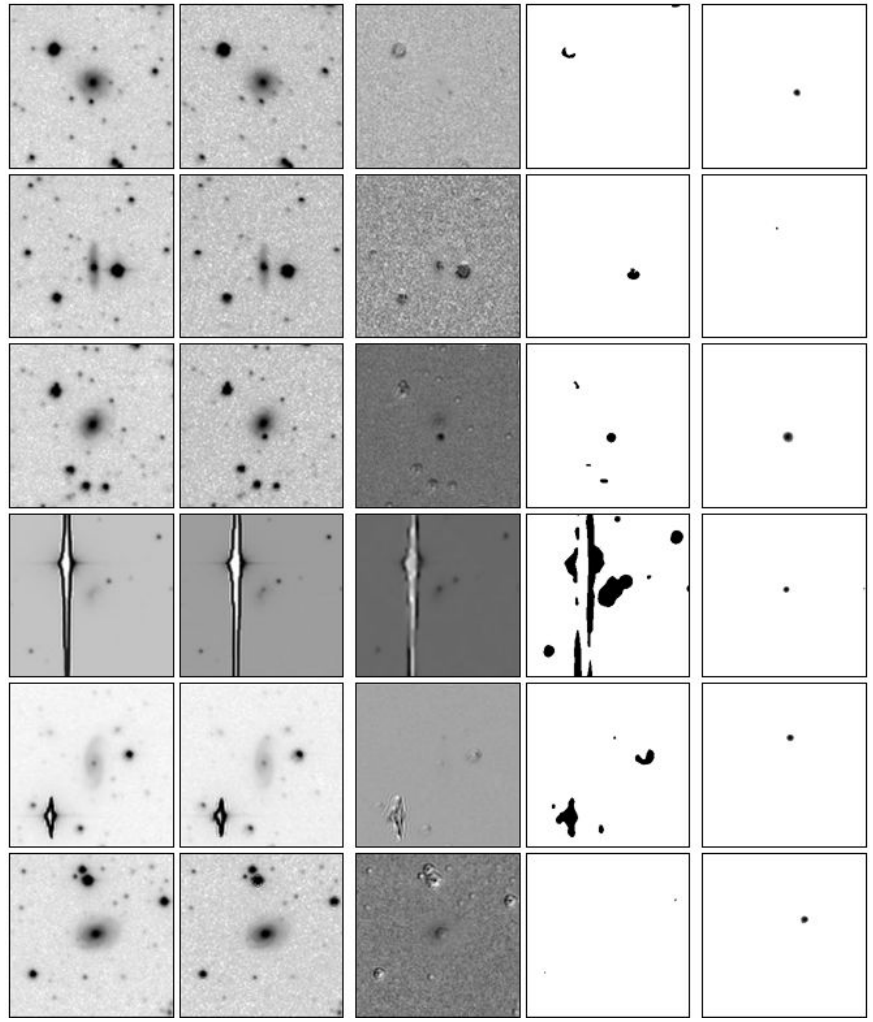
Skip the classical reference-subtraction pipeline entirely.

Encoder–decoder CNN: raw image pair → clean detection map.

Trained on ~100 CRTS SNe + Galaxy Zoo synthetics; L1 loss at all scales.

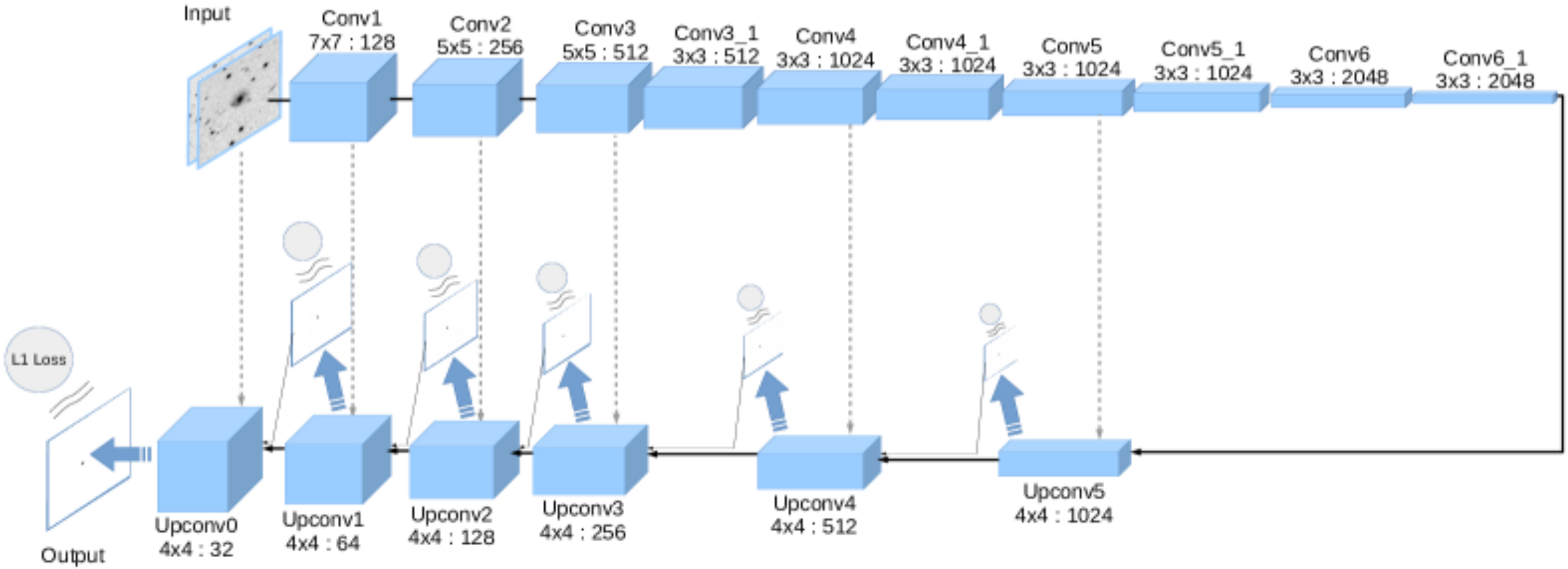
75.5% recall at 98.4% precision on real SNe.

Hint of what foundation models could do for whole pipelines.



ZOGY vs TransiNet on a real SN — arXiv:1710.01422

TransiNet — encoder-decoder architecture



Conv encoder / up-conv decoder with skip connections — arXiv:1710.01422

2. Light curves

The light curve representation problem

Irregular sampling, gaps, heteroscedastic errors.

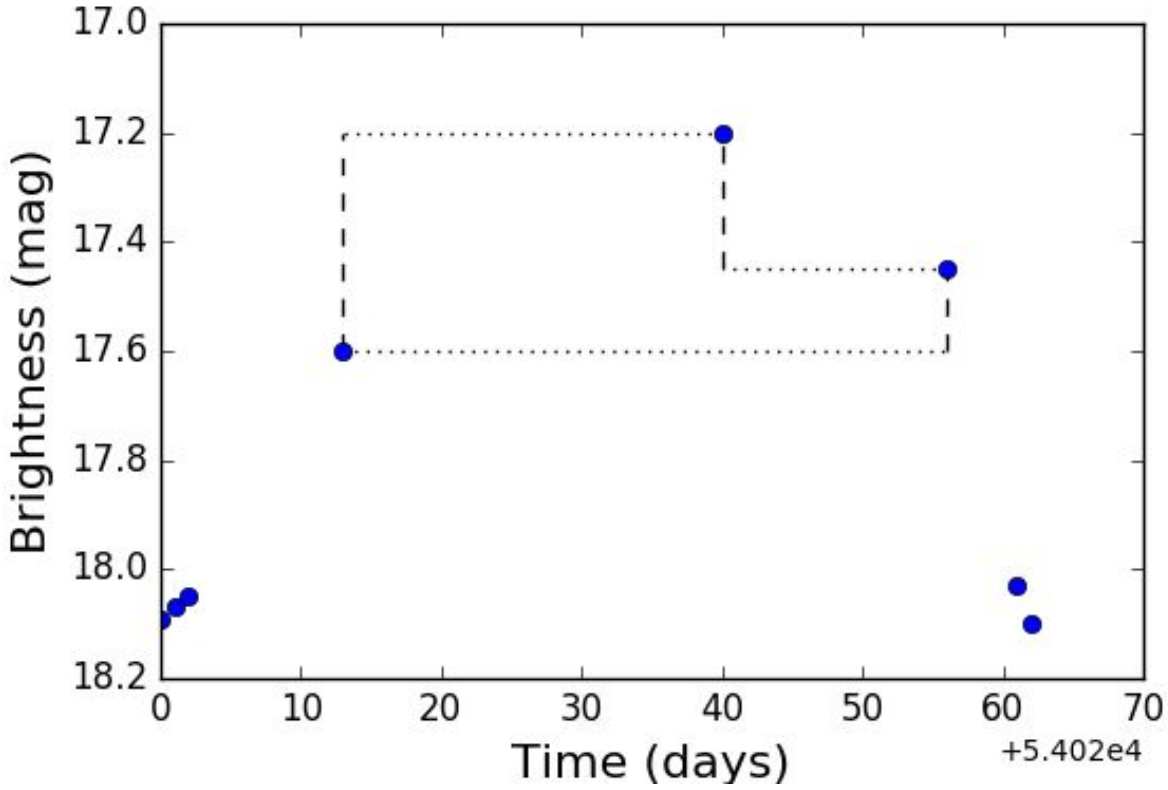
Can't feed naïvely into a CNN.

Three options: hand-crafted features, RNNs / Transformers, or re-package as images.

Our path: re-package as images — and then a CNN.

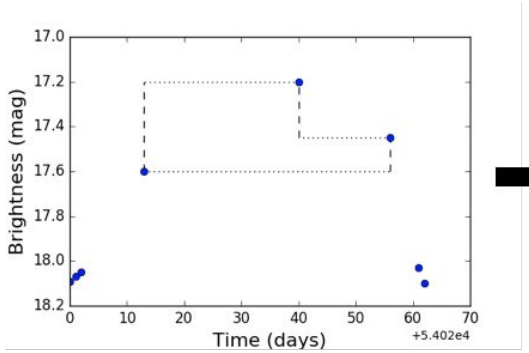
DMDT images (Mahabal+ 2017)

- For every pair of points compute (Δmag, Δt).
- Bin into a 23×24 grid ⟶ a grayscale image.
- Now a standard CNN classifies variable star types.
- 83% accuracy on 7-class CRTS; shallow CNN ≈ deep CNN.
- Representation > architecture.

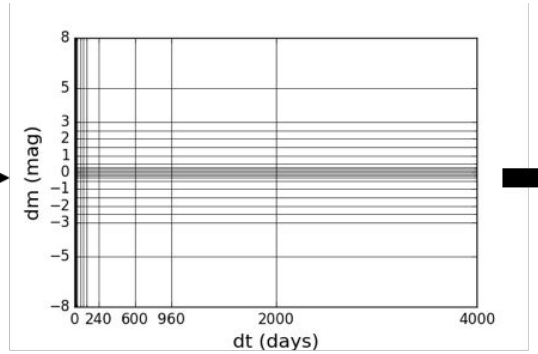


DMDT construction — arXiv:1709.06257

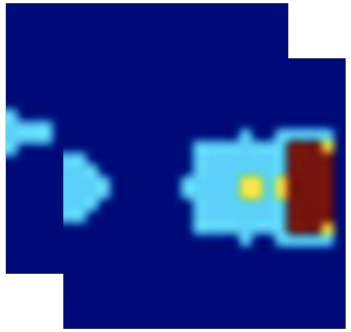
DMDT — stacked class composites



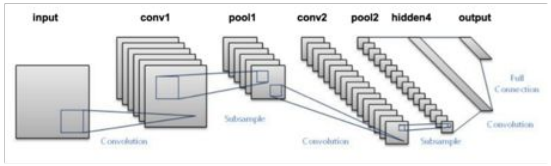
Light curves



Density representation



Equi-area images



Convolutional Neural Network

dmdt images

RR Lyrae

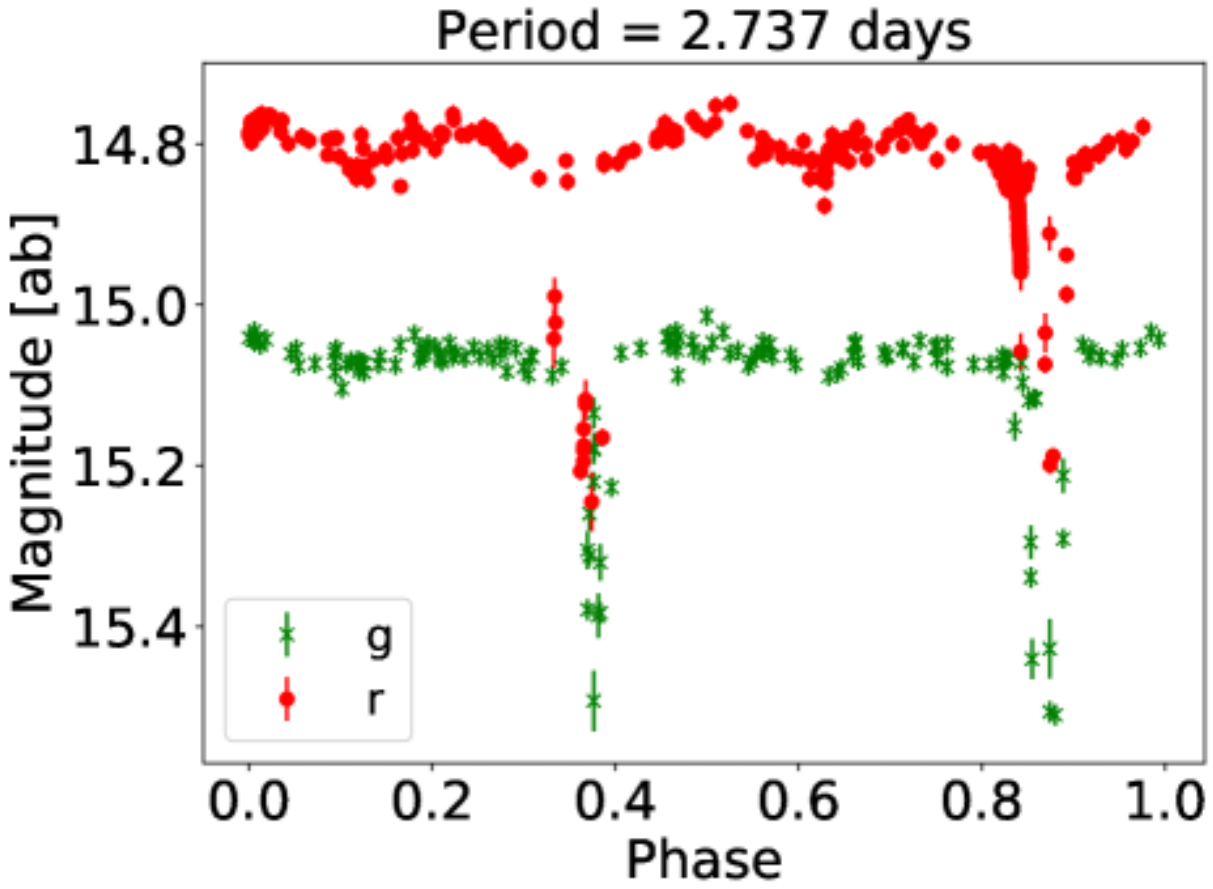
RS CVn

LPV

Per-class DMDT averages: each class has a distinct fingerprint — arXiv:1709.06257

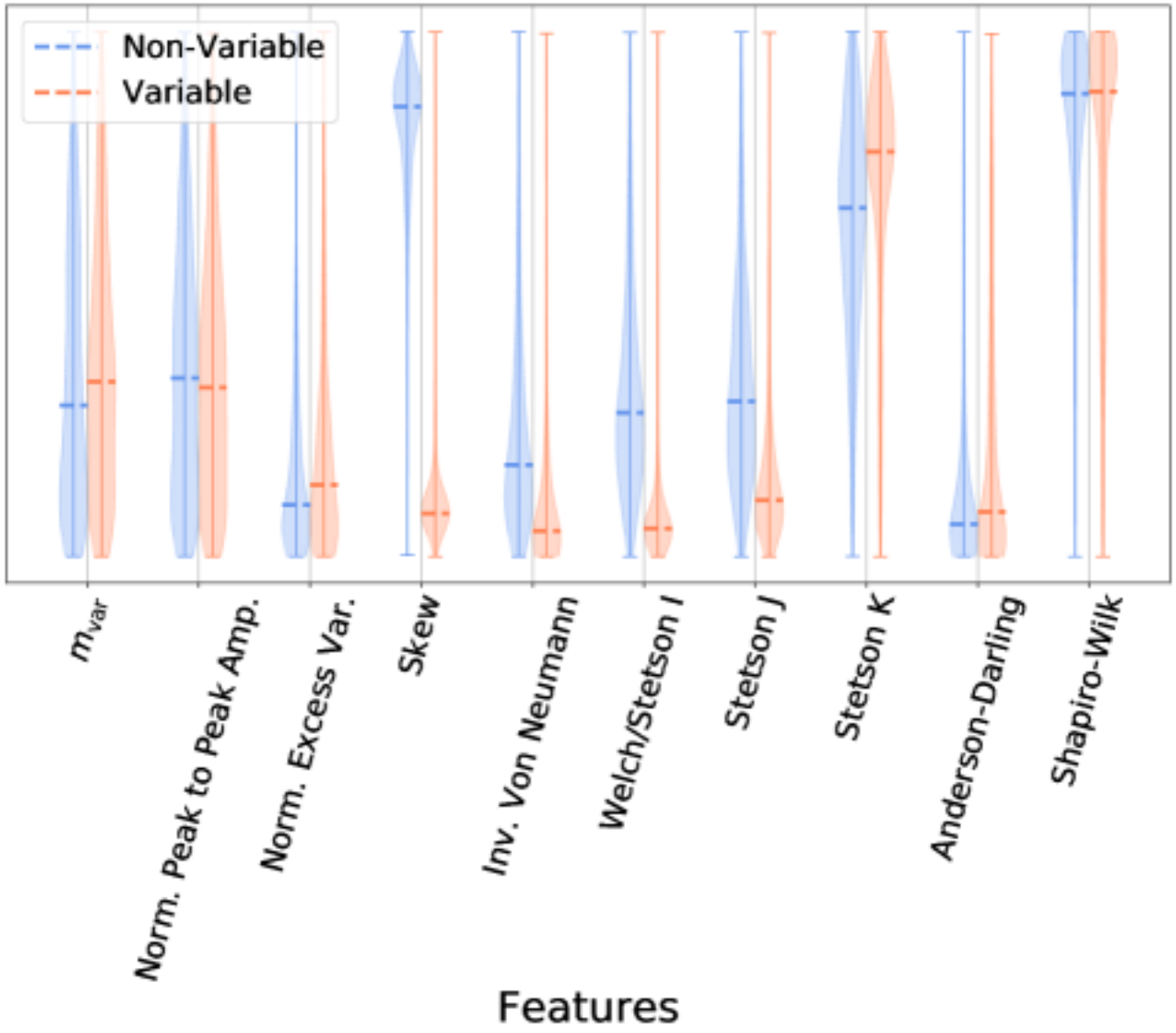
ztfperiodic — period-finding at scale (Coughlin+ 2020)

- Before classification: variability stats + a period.
- GPU Conditional Entropy + Lomb–Scargle, run in parallel.
- 37 statistical features per light curve, aliasing mitigation.
- ~1 second / light curve ⟶ all 1.3B ZTF sources reachable.
- The plumbing under SCOPE.



Phase-folded periodic variables — arXiv:2009.14071

ztfperiodic — GPU scaling



Analysis throughput vs. GPU count, across facilities — arXiv:2009.14071

SCOPE — Source Classification Project (van Roestel+ 2021)

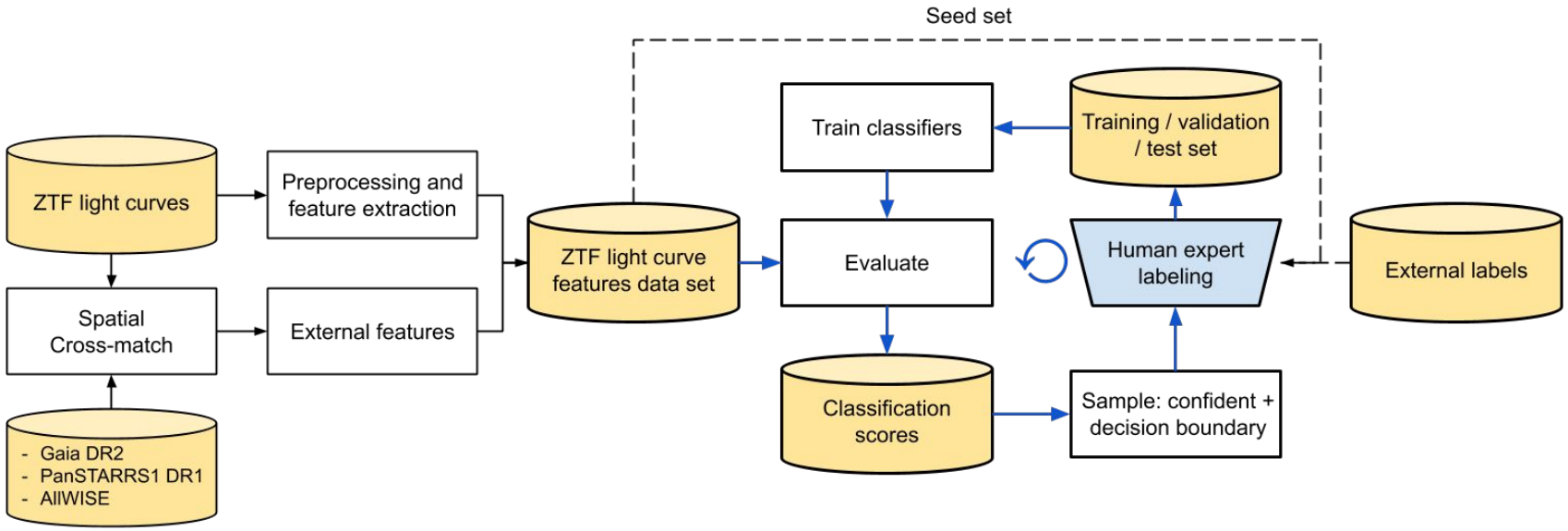
Many dichotomous binary classifiers, not one big multi-class head.

Two taxonomies in parallel: ontological + phenomenological.

40 ZTF features + DMDT + Gaia/Pan-STARRS/AllWISE crossmatches.

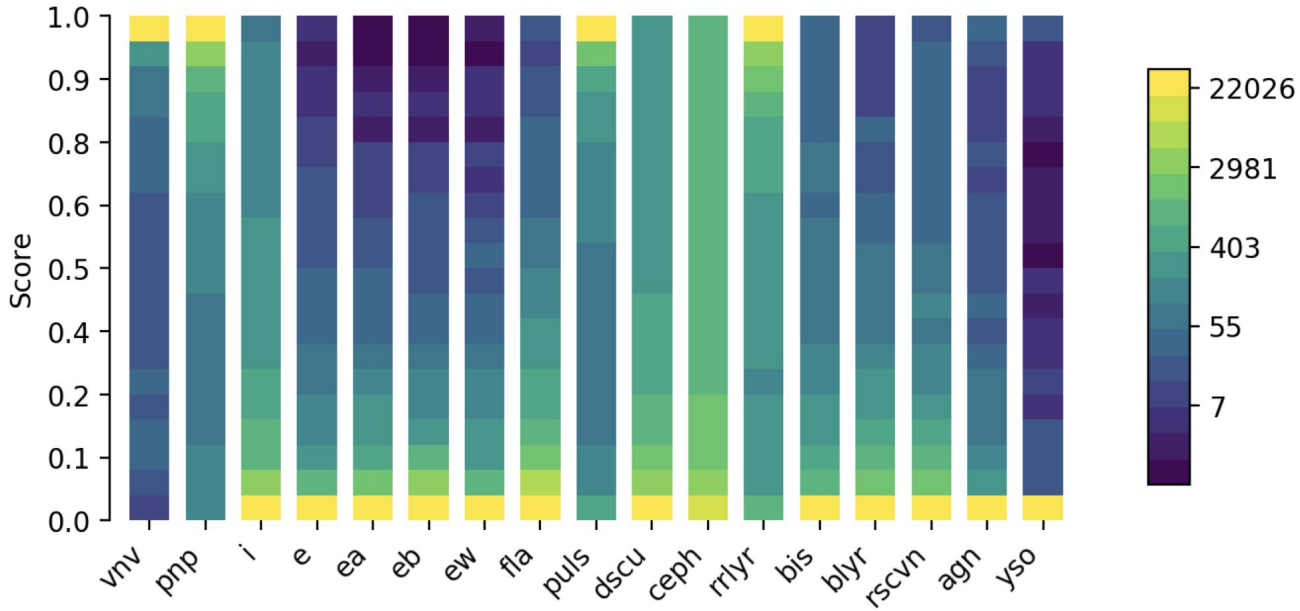
DNN and XGBoost both trained; active-learning loop via SkyPortal.

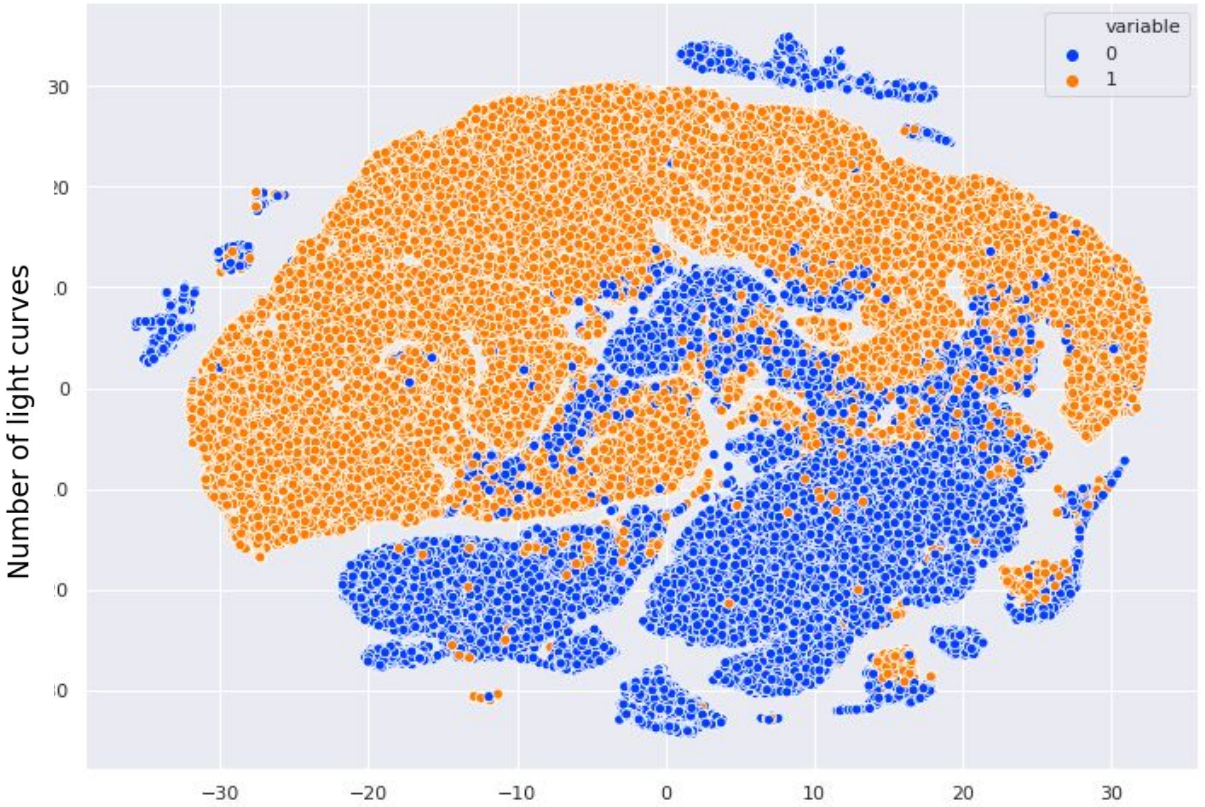
89% precision for RR Lyrae extraction in production.



Active-learning pipeline — arXiv:2102.11304

SCOPE — feature-space structure





t-SNE: ontological classes cluster cleanly in feature space — arXiv:2102.11304

SCOPE at full ZTF scale (Healy+ 2023)

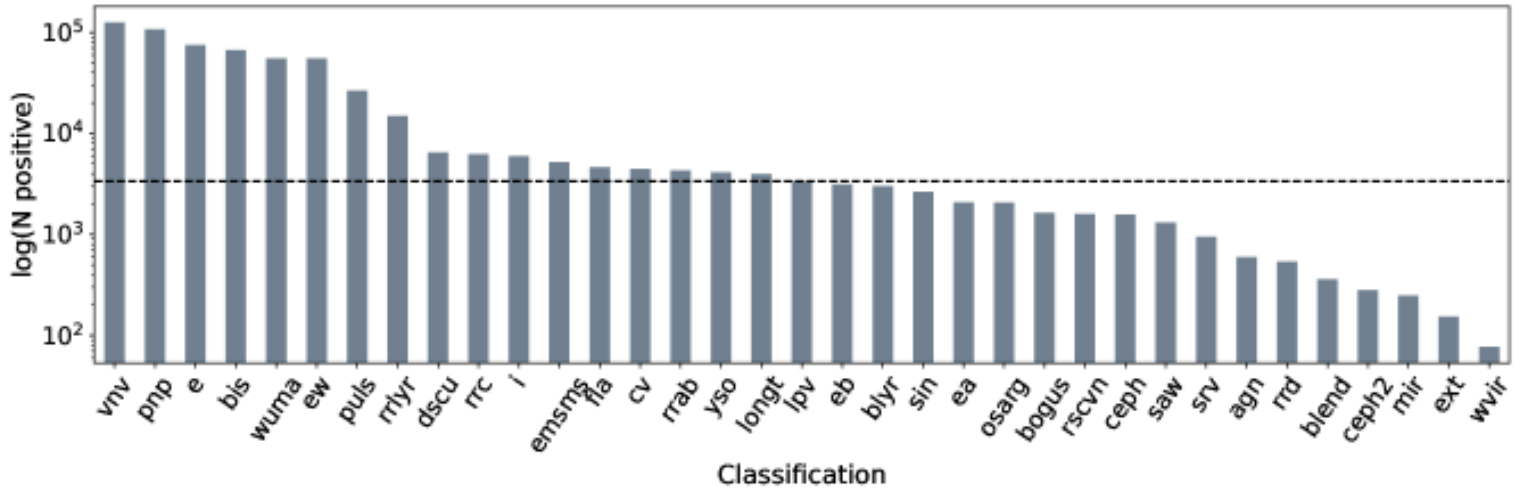
Same dichotomous framework, now applied to all of ZTF.

35+ classifiers; 170k hand-labelled light curves after iterative AL.

112,476,749 classifications produced (40 fields).

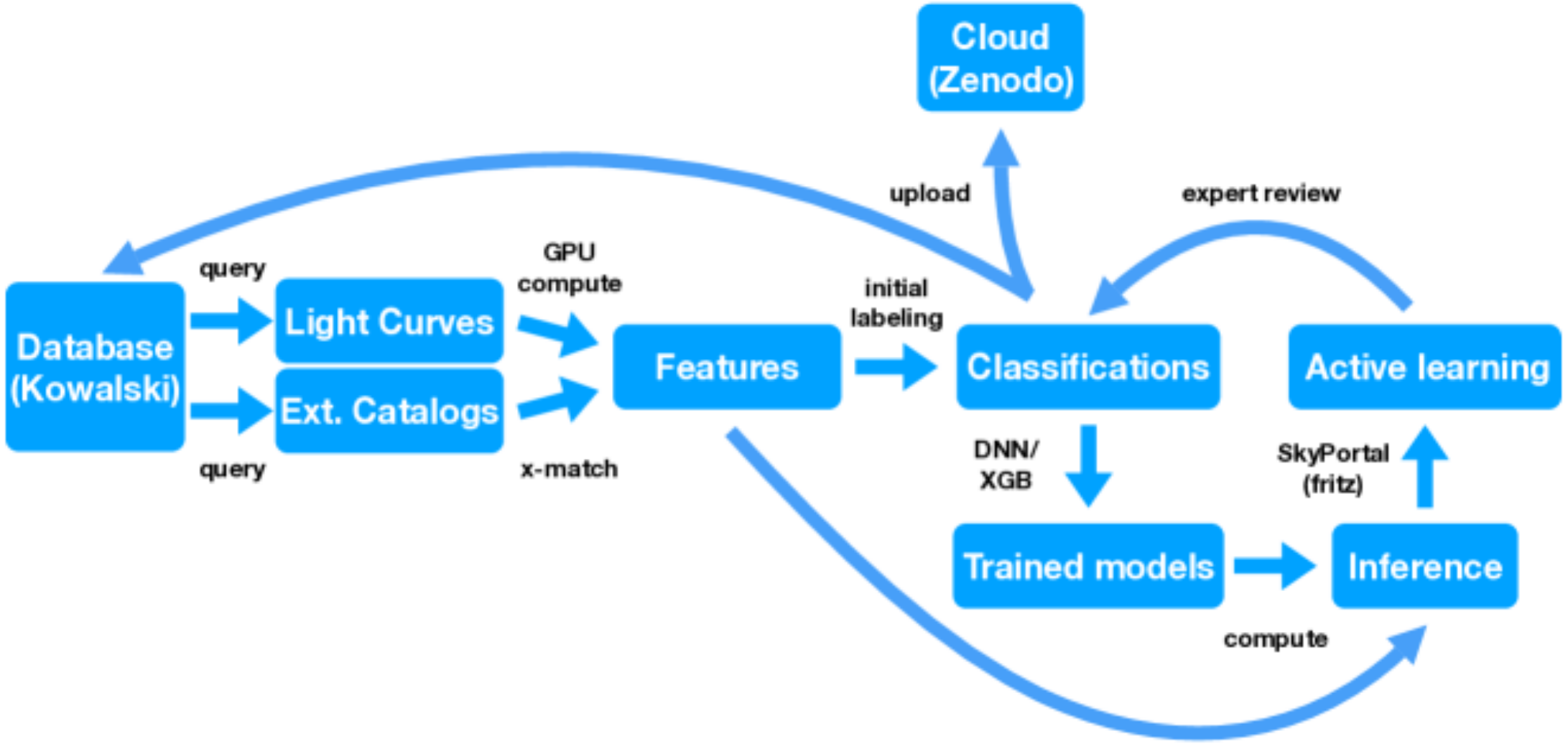
Period determination + Fourier power dominate feature importance.

Public catalog — the deliverable for the community.



Precision/recall across classifiers — arXiv:2312.00143

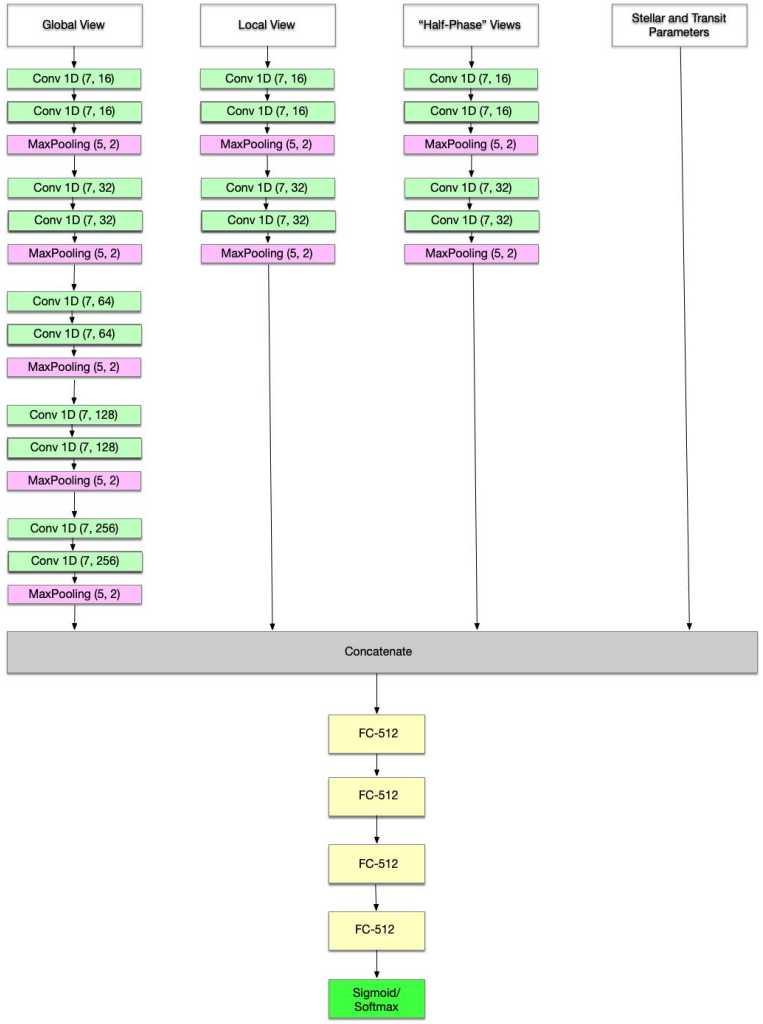
SCOPE — end-to-end workflow



From raw light curves through active learning to public catalog — arXiv:2312.00143

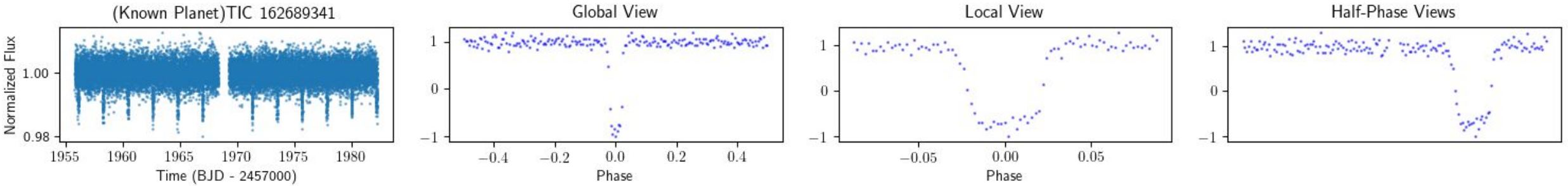
TESS exoplanet transits (Rao+ 2021)

- 2-min cadence light curves; Transit-Least-Squares pre-screen.
- Three input views into one CNN: global (201 bins) + local (81) + half-phase.
- Half-phase view kills eclipsing-binary contaminants.
- AUC 0.908; recovers 91.9% of known planets at score > 0.7.
- 38 new vetted candidates across 7 sectors.



Three-view CNN architecture — arXiv:2101.09227

TESS — phase-folded example



A known planet's phase-folded light curve — arXiv:2101.09227

3. Spectra

SNIascore — real-time Type Ia typing (Fremling+ 2021)

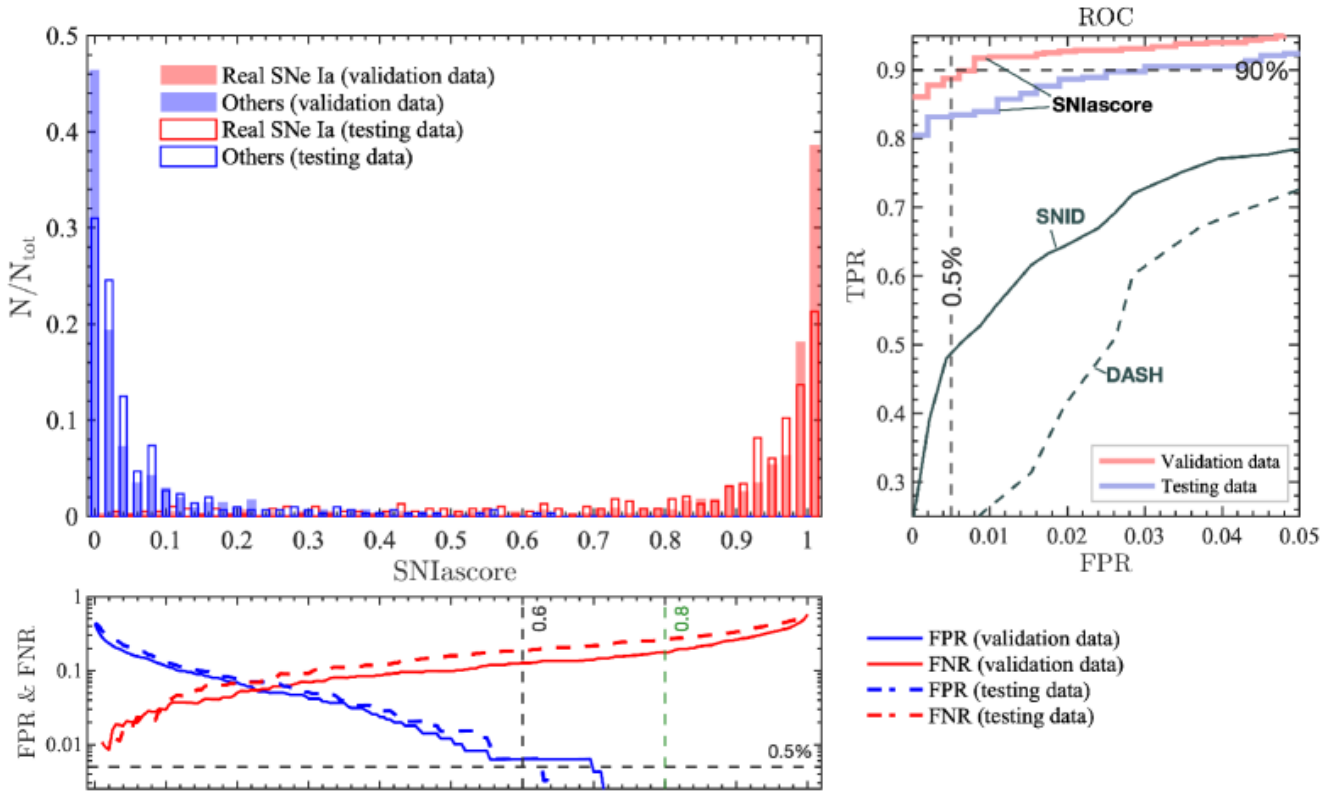
Low-resolution (R ~ 100) SEDM spectra ⟶ SN Ia / not-Ia.

BiLSTM + GRU + BiLSTM with 40–45% dropout.

Joint redshift regression head (custom MBE loss).

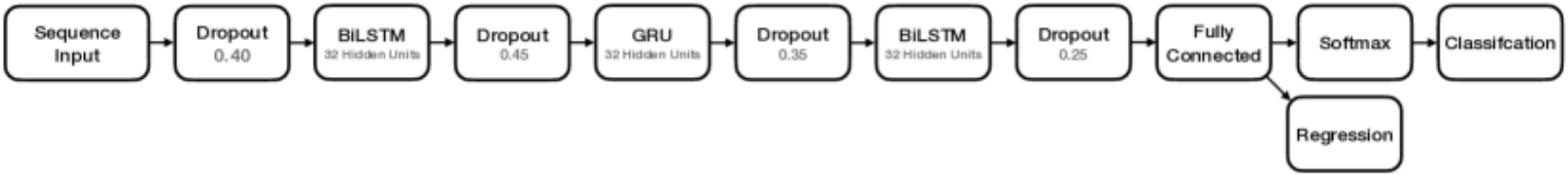
90% TPR at 0.6% FPR on validation; runs inside the live BTS workflow.

Frees humans for the rare / weird objects.



Score distributions + ROC vs. SNID / DASH — arXiv:2104.12980

SNIascore — RNN architecture



BiLSTM → GRU → BiLSTM stack with heavy dropout — arXiv:2104.12980

CCSNscore — core-collapse SNe (Sharma+ 2024)

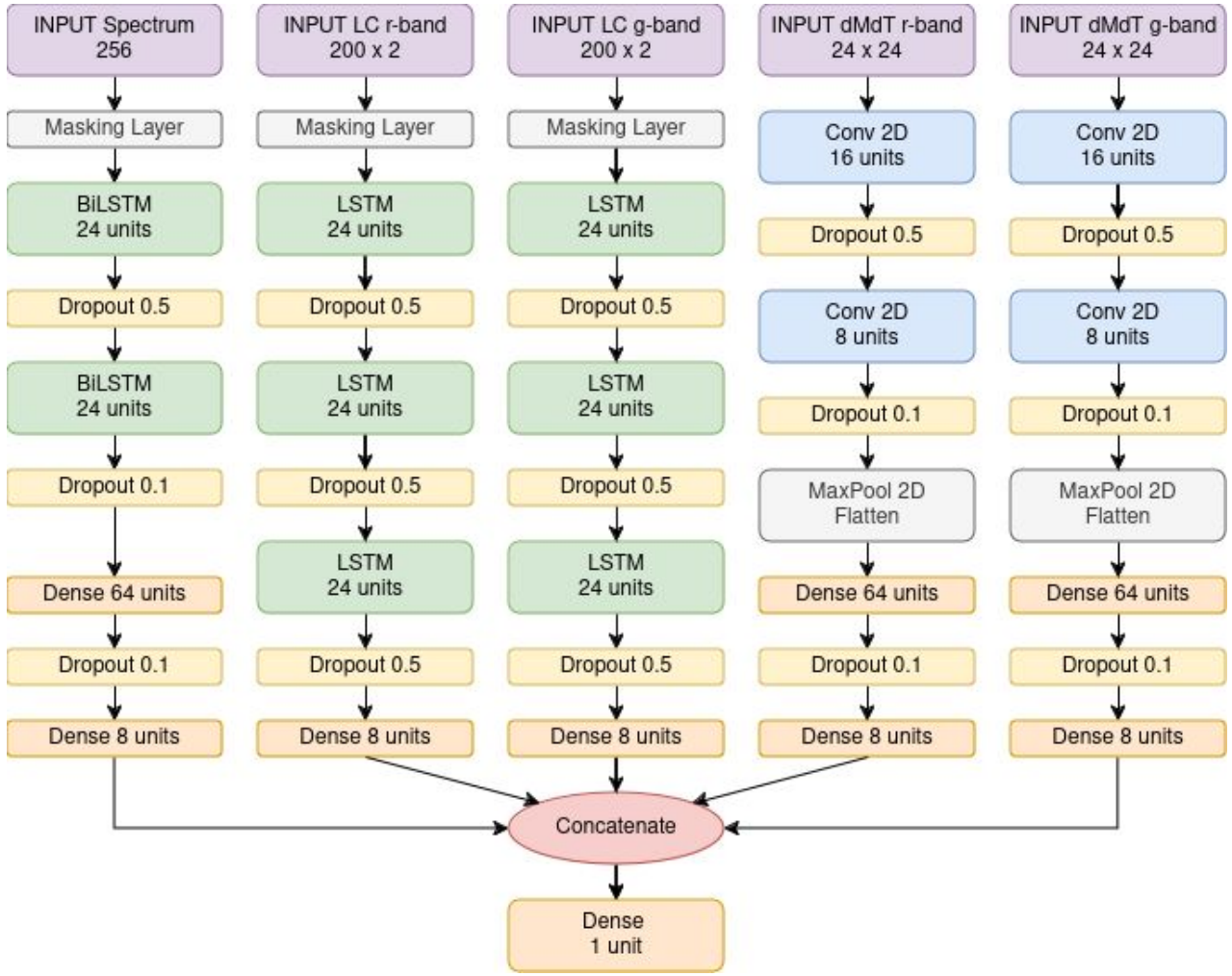
Multi-input: spectra + ZTF g/r light curves + DMDT.

Hierarchical: H-rich vs. H-poor, then II / IIb / IIn or Ib / Ic / Ic-BL.

Bi-LSTM on spectra, LSTM on photometry.

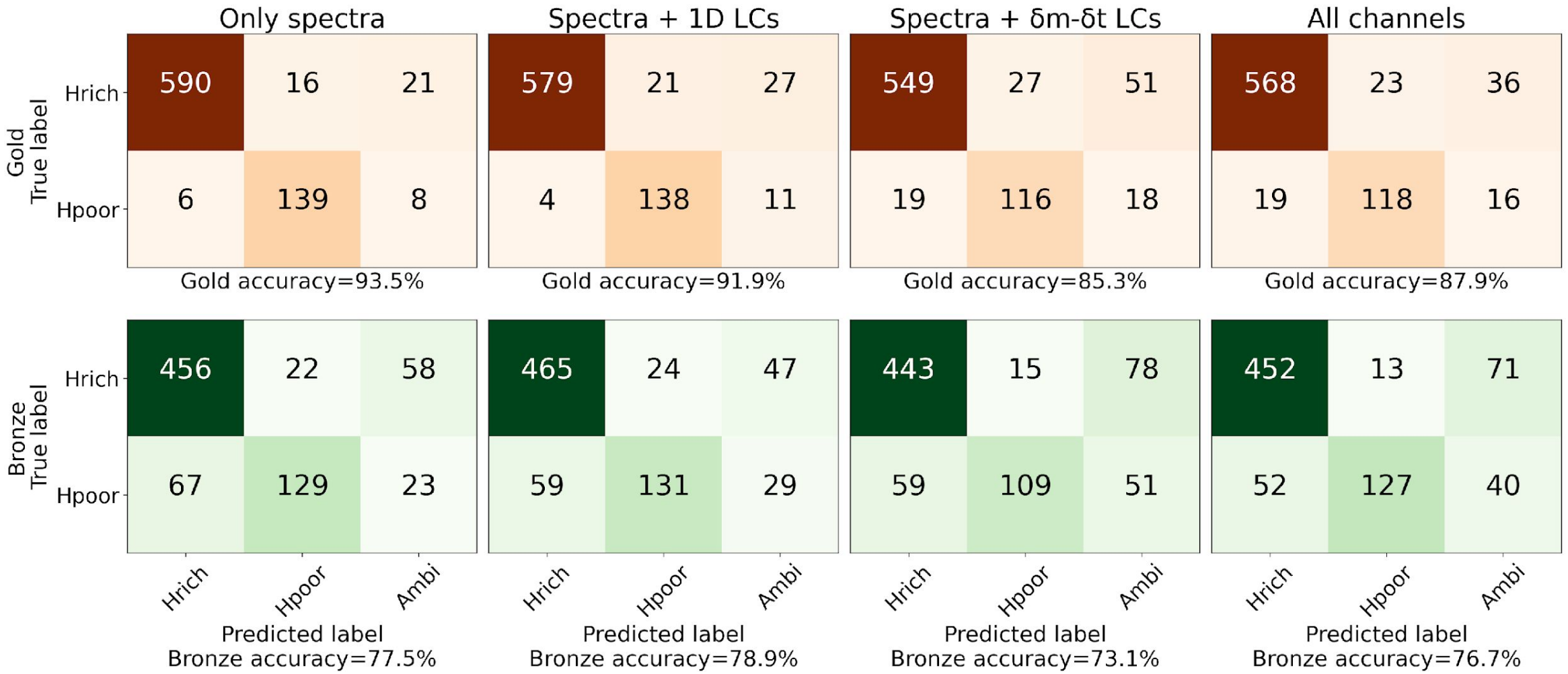
MC dropout (×100) ⟶ per-prediction uncertainty.

98% accuracy on gold-quality spectra.



Multi-channel network — arXiv:2412.08601

CCSNscore — Layer-2 confusion matrices



Subtype performance across input modalities — arXiv:2412.08601

4. Gravitational waves

GW data 101

Strain time series, kHz sampling, multiple detectors.

Q-transform $\longrightarrow$ 2-D spectrogram (now an image, CNN-friendly).

Noise is dominated by transient glitches — the real adversary.

DL tasks: glitch ID, real/bogus, source classification, parameter inference, and… predicting lock loss.

DMDT goes to LIGO (Biswas+ 2019)

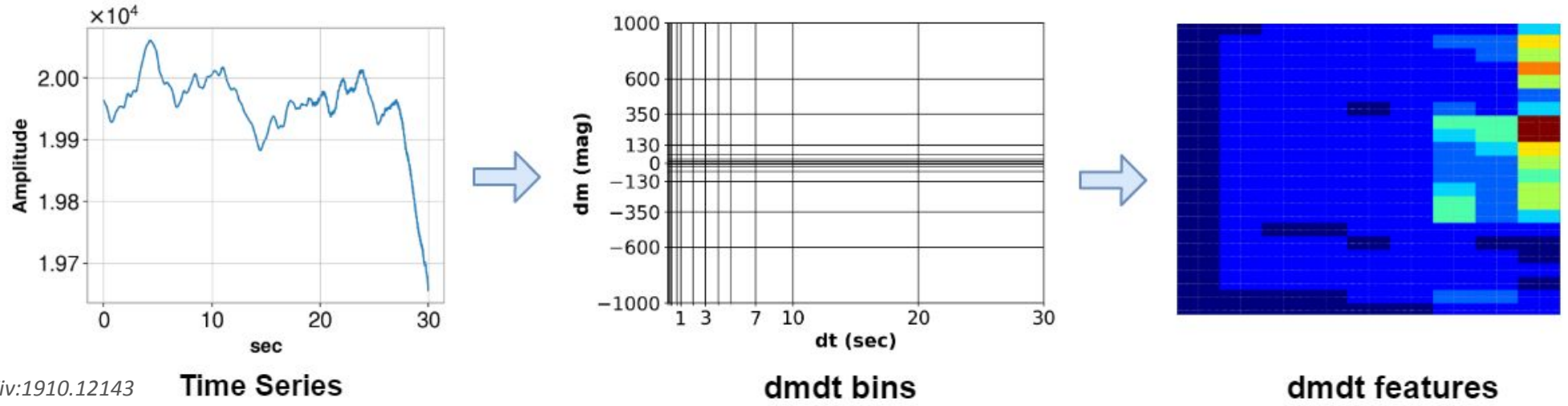
The DMDT representation applied to detector control channels.

22 auxiliary channels ⟶ DMDT images ⟶ CNN.

98% accuracy predicting lock loss at t = 0, 90% at t = −30 s.

3-channel subset (SRCL, MICH, REFL) captures most events.

Same representation trick — wildly different physics.



DMDT pipeline for IFO channels — arXiv:1910.12143

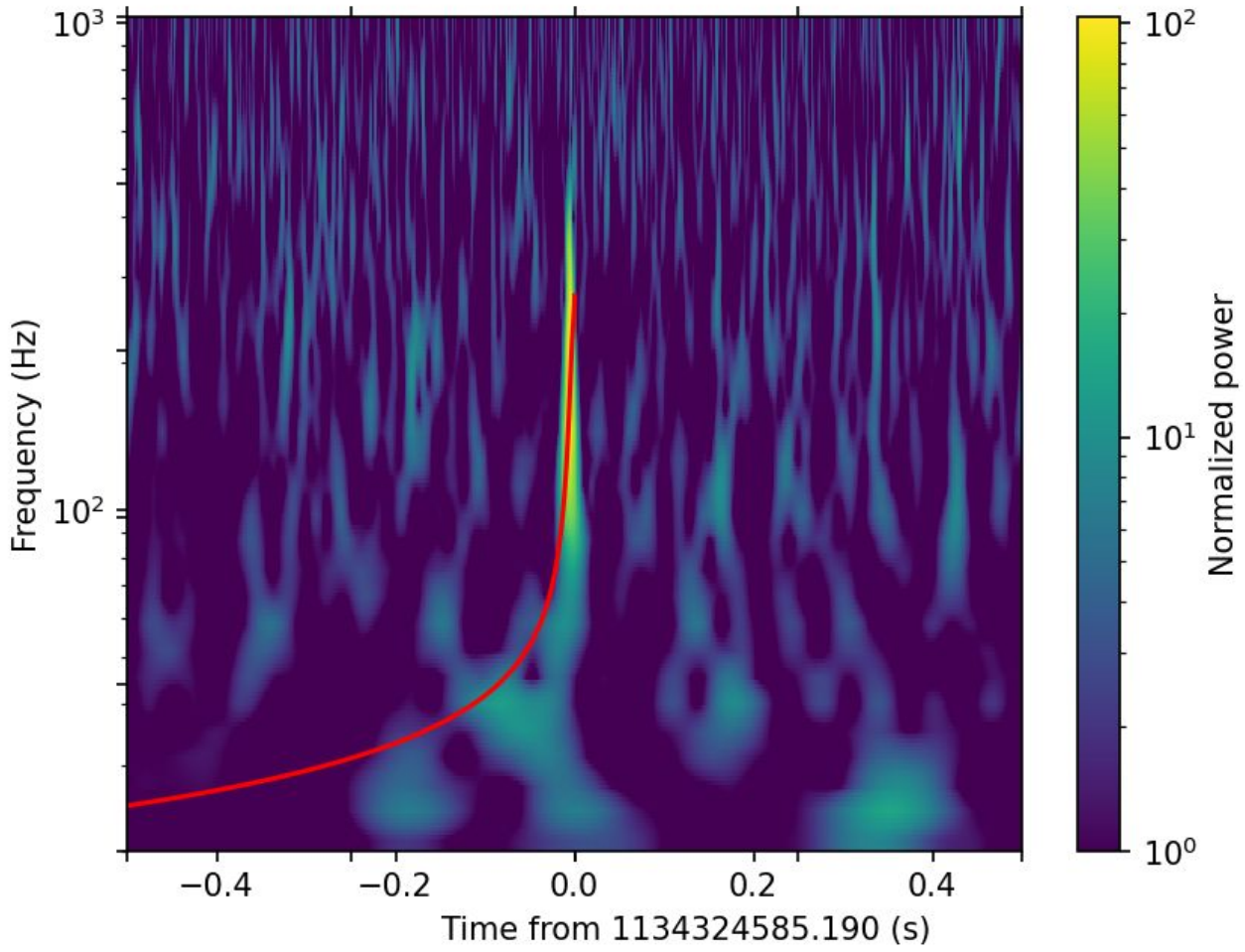
GWSkyNet — first GW real/bogus (Cabero, Mahabal, McIver 2020)

Inputs: BAYESTAR sky-map + 3D volume + detector metadata.

Non-sequential CNN: residual image branches + MLP for numerics.

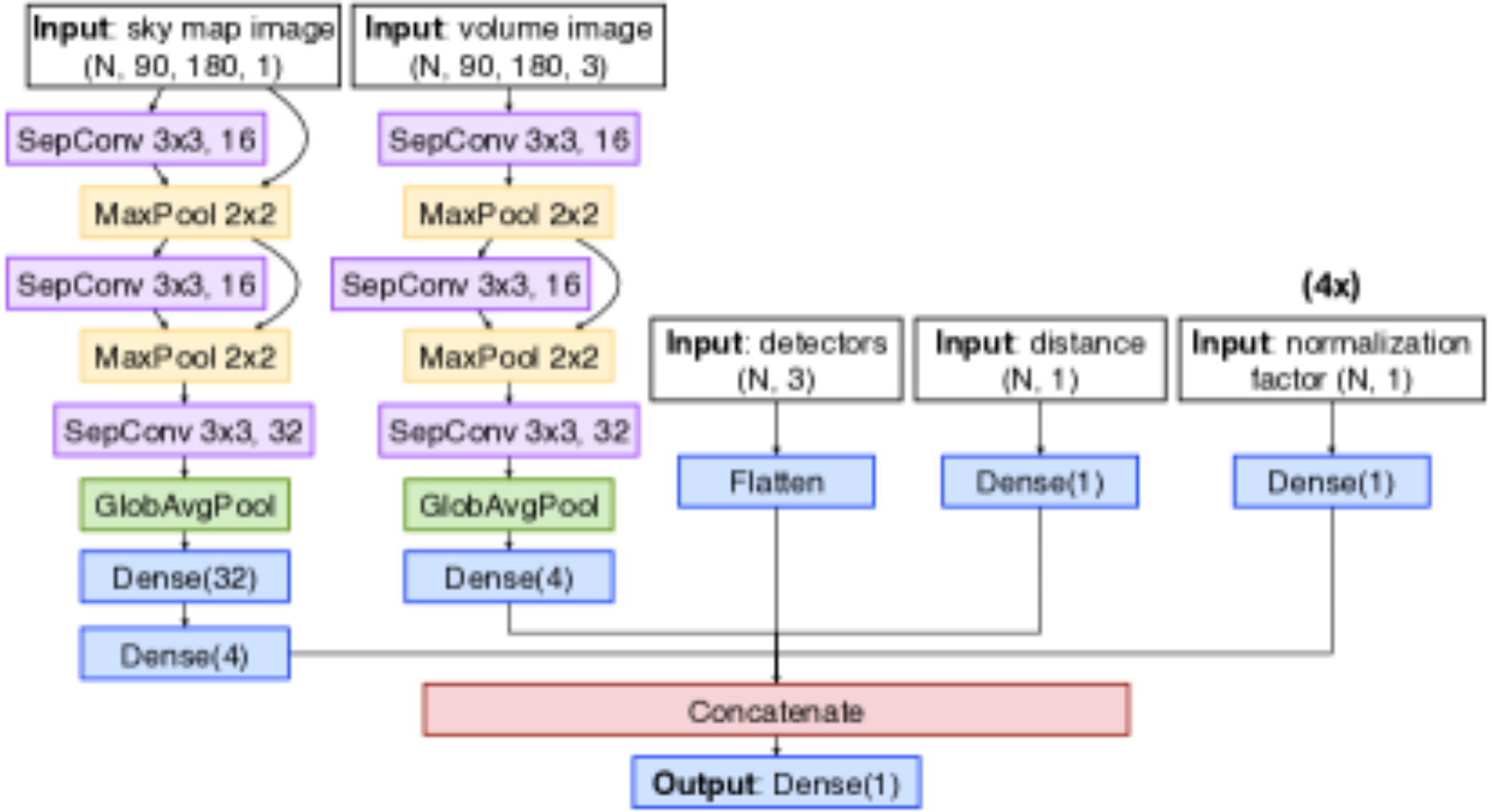
Trained on ~2.9k simulated CBCs + 1.3k real glitches.

93.5% accuracy; 97.7% precision @ 92.4% recall.



Blip-glitch spectrogram vs. template — arXiv:2010.11829

GWSkyNet — multi-branch architecture



CNN branches on sky-map images + MLP on metadata — arXiv:2010.11829

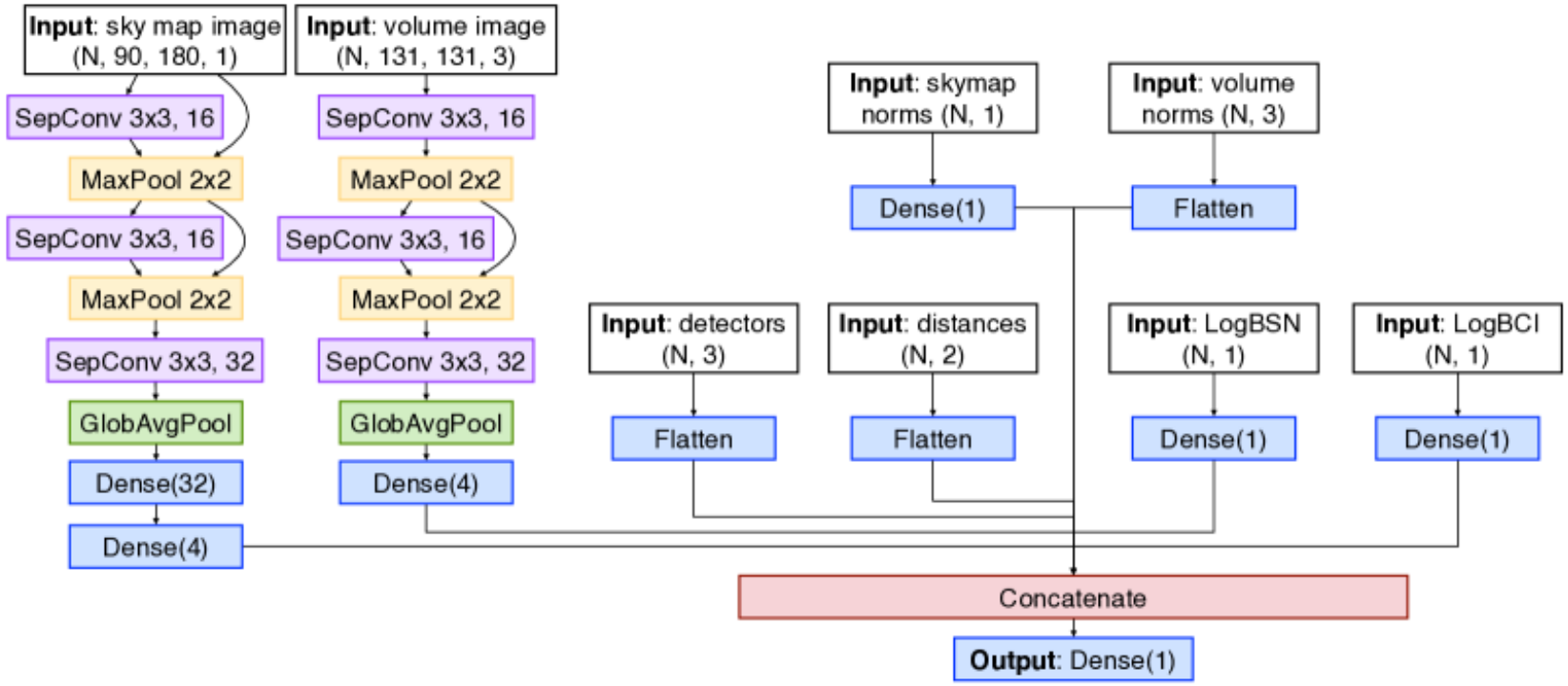
GWSkyNet-Multi — three classes (Abbott+ 2021)

LVC paper: BBH vs. NS-bearing vs. glitch.

Three one-vs-all CNNs in a hierarchical decision tree.

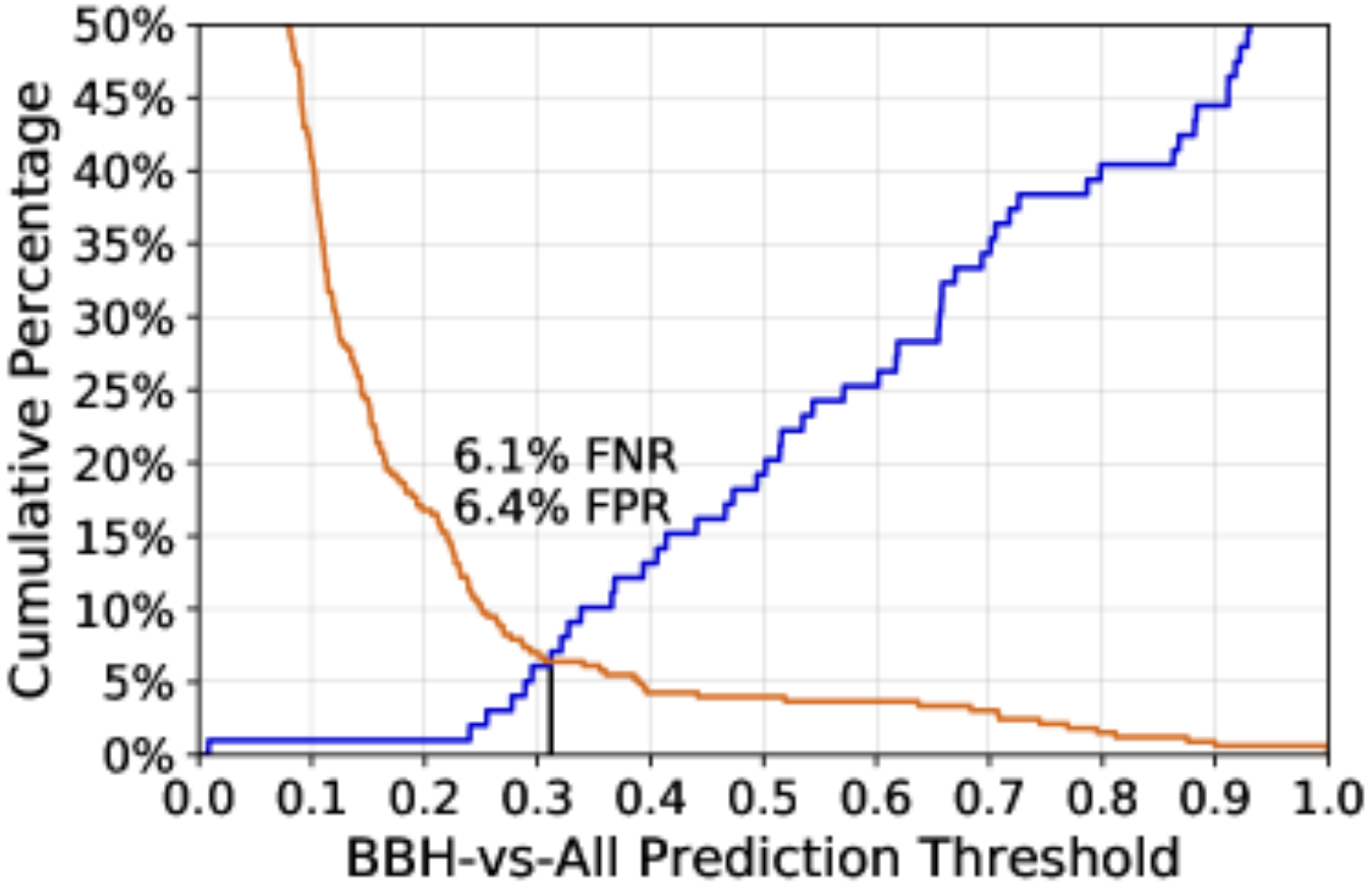
Same inputs as GWSkyNet, broader output space.

Correctly classified 36 / 40 O3a alerts.



Multi-branch architecture — arXiv:2111.04015

GWSkyNet-Multi — threshold tuning



FNR / FPR trade-off curves across thresholds — arXiv:2111.04015

Explainability — what is GWSkyNet looking at? (Raza+ 2023)

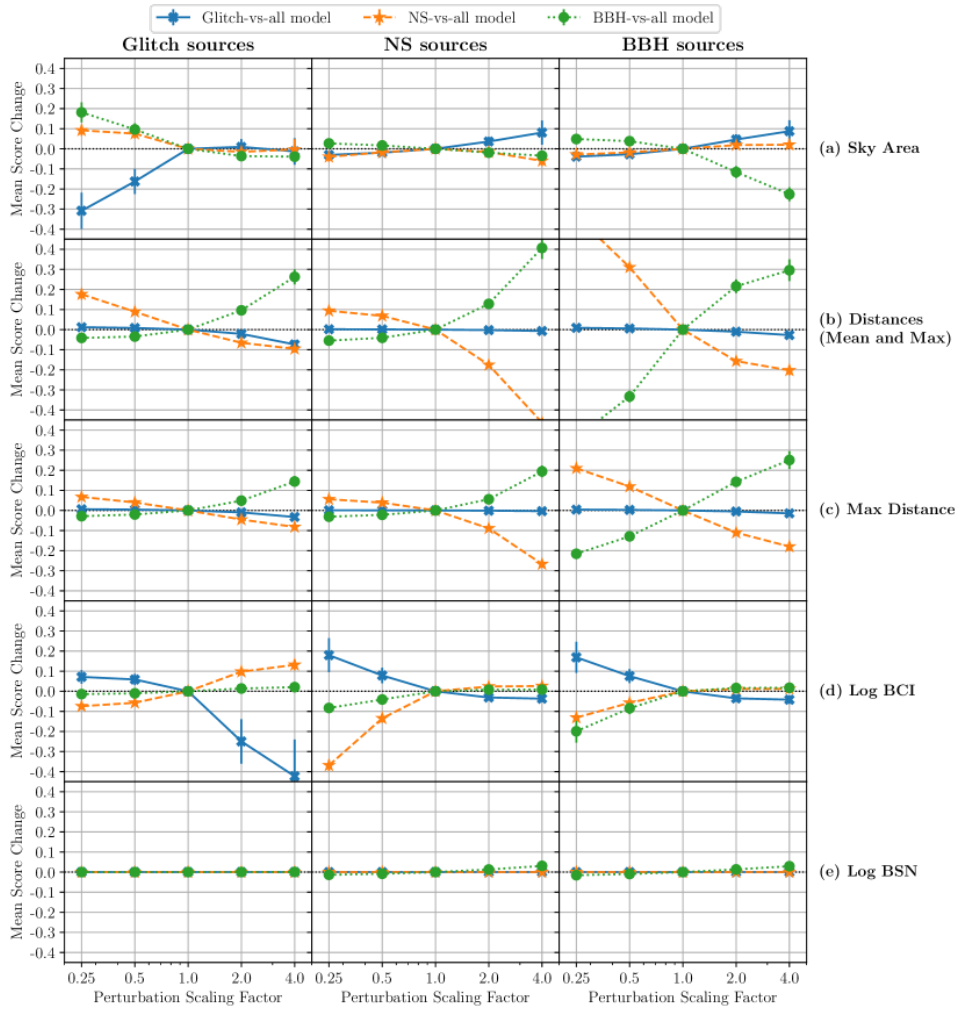
Systematic input perturbation (0.25× – 4×) to probe sensitivity.

Finding: sky-area + coherence Bayes factor drive real-vs-glitch.

Distance separates BH from NS classes.

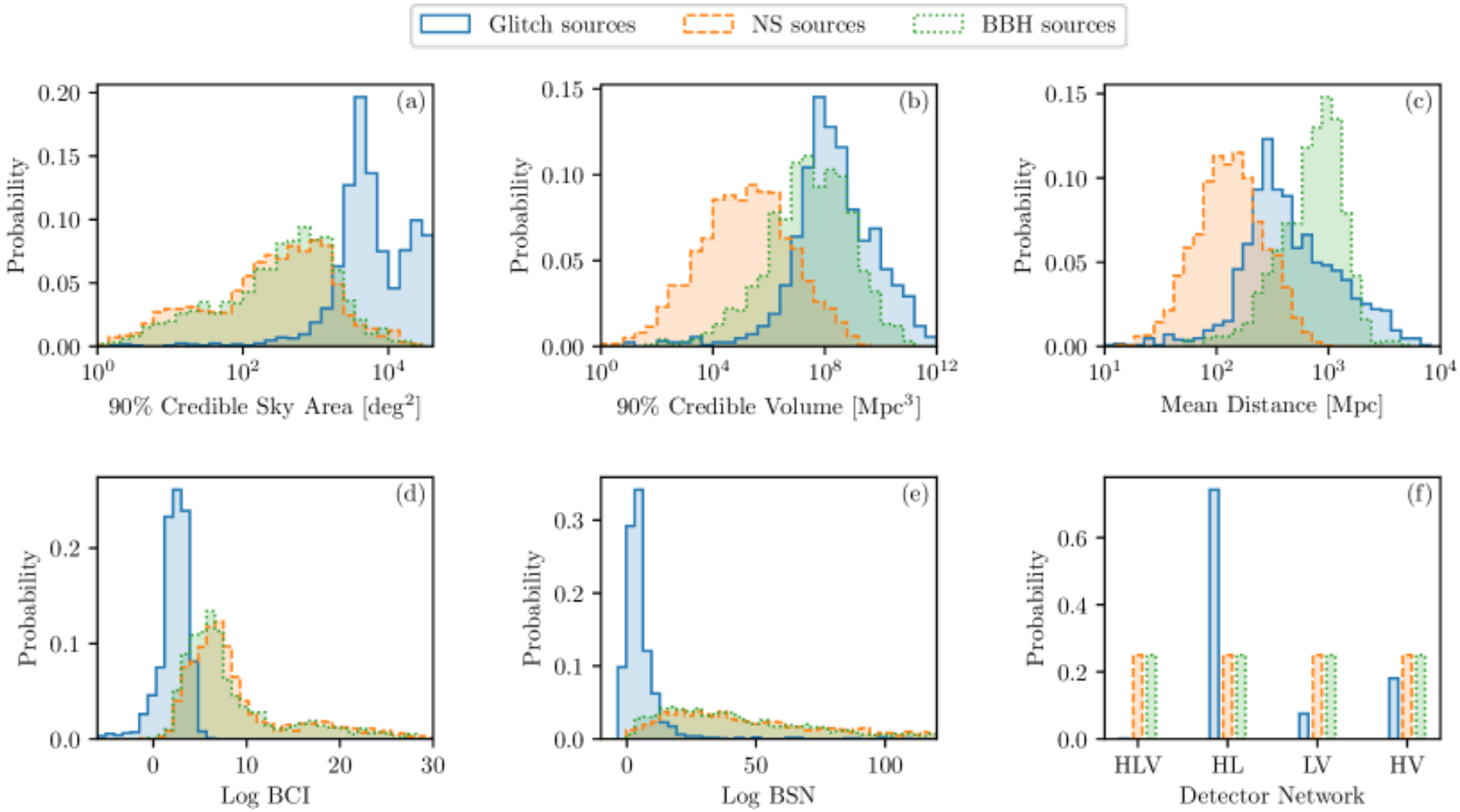
Explains 15 misclassifications in O3.

20 retrained models → genuine error bars.



Perturbation response curves — arXiv:2308.12357

Explainability — feature distributions by class



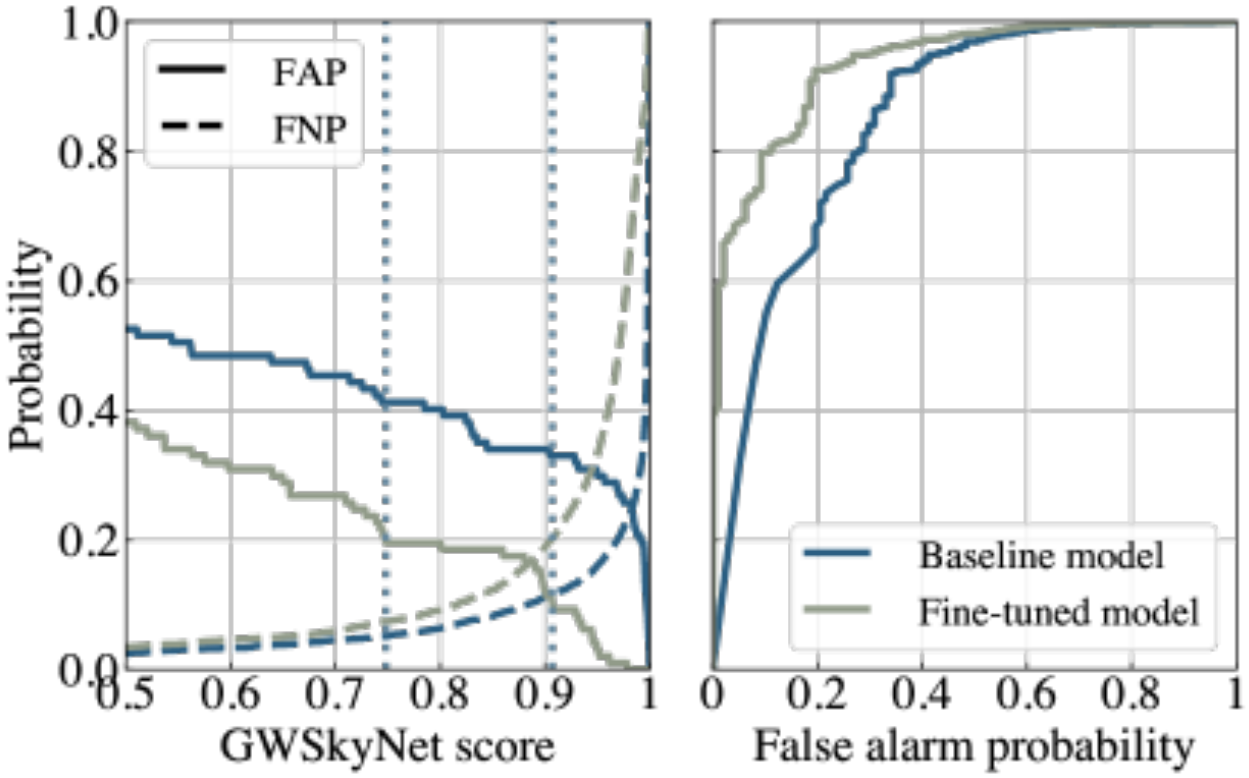
Sky area, distance, Bayes factors differ across glitch / NS / BH — arXiv:2308.12357

GWSkyNet II — operational in O4 (Chan+ 2024)

Retrained on O3 glitches; fine-tuned on mock-data alerts.

Transfer learning: freeze conv layers, retrain dense head.

From research code to LVK production.



ROC: baseline vs. fine-tuned — arXiv:2408.06491

GWSkyNet-Multi II — radical simplification (Raza+ 2025)

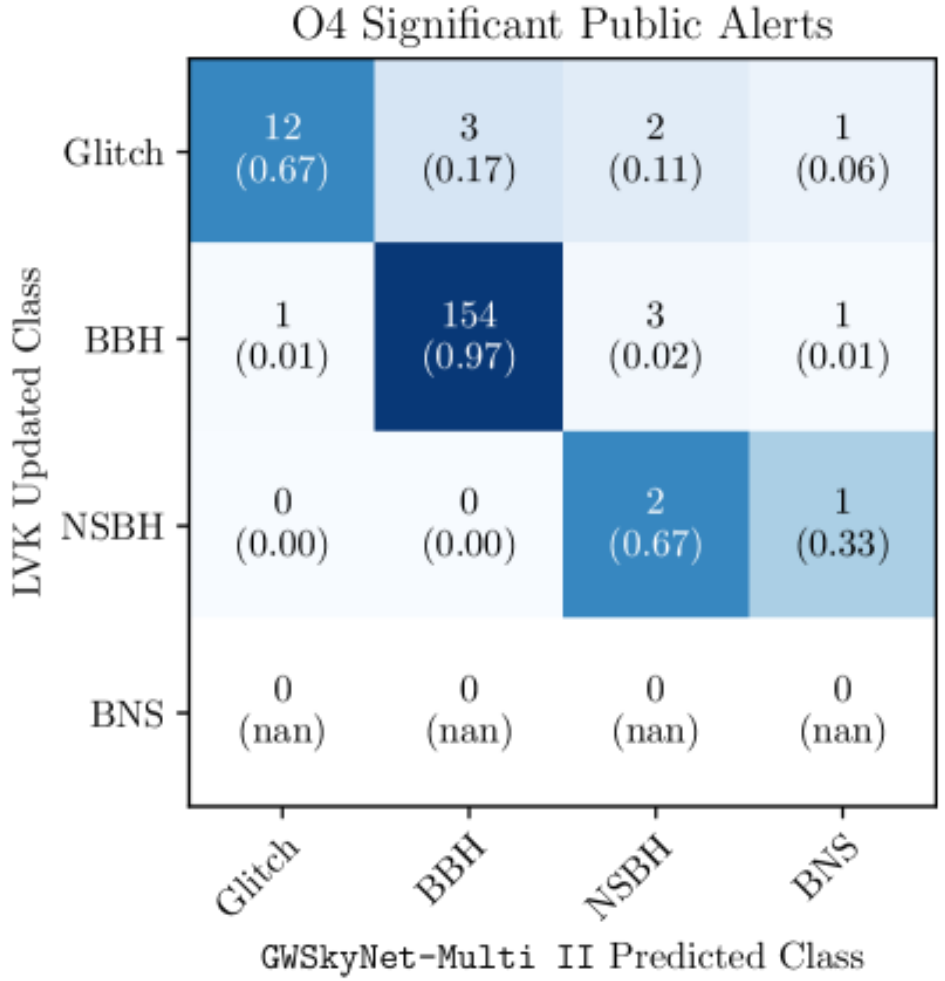
Drop image inputs; use 9 numeric summary statistics.

188 trainable parameters — 60× smaller than the predecessor.

Four classes now: glitch, BBH, NSBH, BNS.

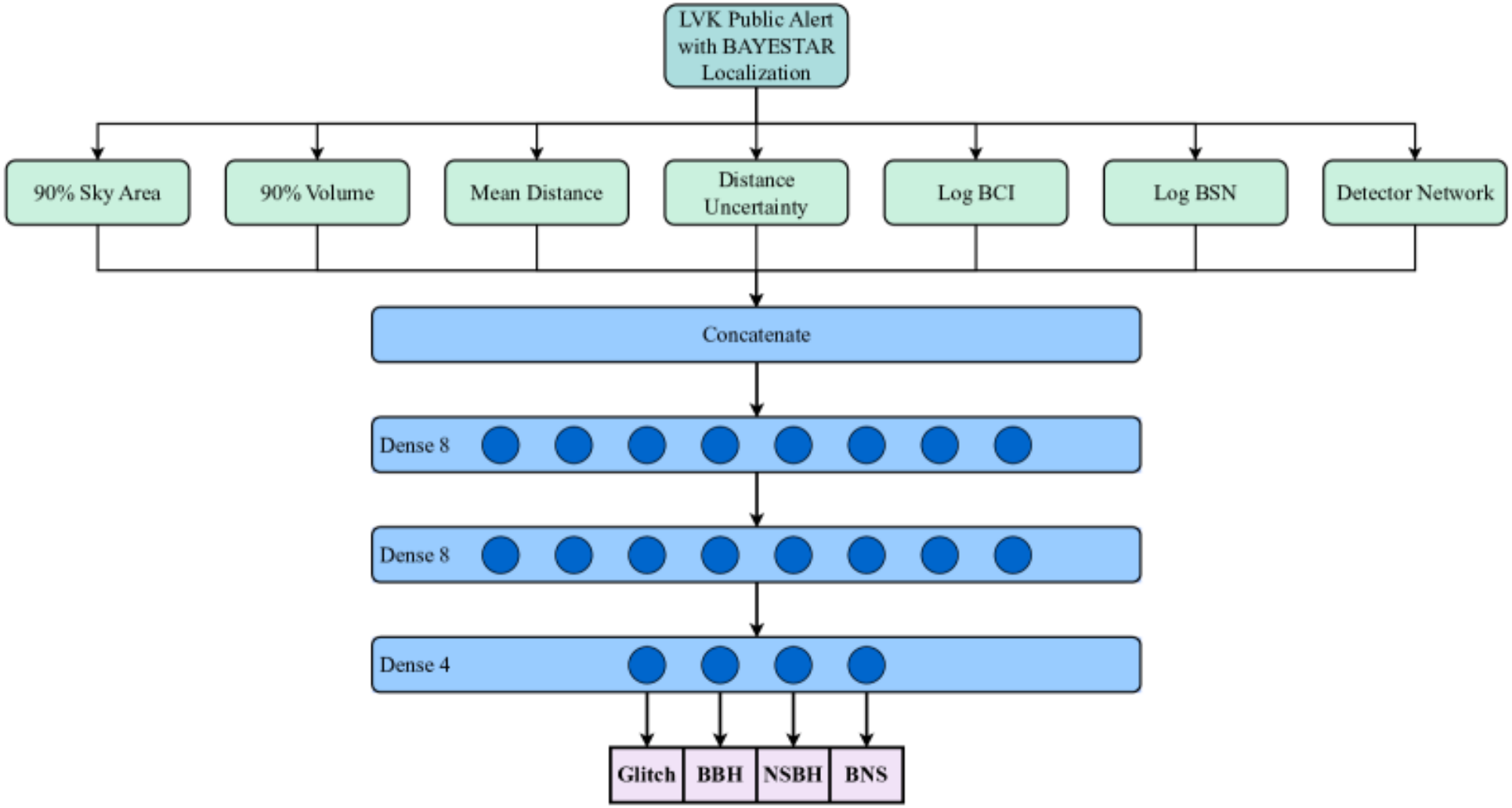
Ensemble of 20 models → uncertainty per alert.

93% accuracy on 180 O4 alerts (May 2023 – Dec 2024).



O4 confusion matrix — arXiv:2502.00297

GWSkyNet-Multi II — model architecture



9 numeric inputs → 2 dense layers → 4 class probabilities (188 params) — arXiv:2502.00297

GWSkyNet-MassGap (Raza+ 2026)

Targets the 2–5 M⊙ "lower mass gap" between heaviest NS and lightest BH.

Two outputs: P(component in mass gap), P(an NS is involved).

Best performance for chirp mass ≥ 15 M⊙; harder at low mass (needs ratio).

Mean prediction error 9% (mass-gap), 6% (NS) on O4a candidates.

arXiv:2605.00391.

GW takeaways

- One toolbox, a growing family of tasks.
- From "is it real?" → "what is it?" → "where on the mass spectrum?"
- Real-time alerts + offline inference both matter.
- DMDT from variability shows up here too.
- Models get *simpler* as understanding improves (GWSkyNet-Multi II: 188 params).

5. Radio transients

FRBs and the candidate-rate problem

CHIME / DSA-110 generate thousands of dispersed candidates per day.

Dynamic spectra (time × frequency) are 2-D — natural CNN inputs.

More than real/bogus: morphology drives follow-up choices.

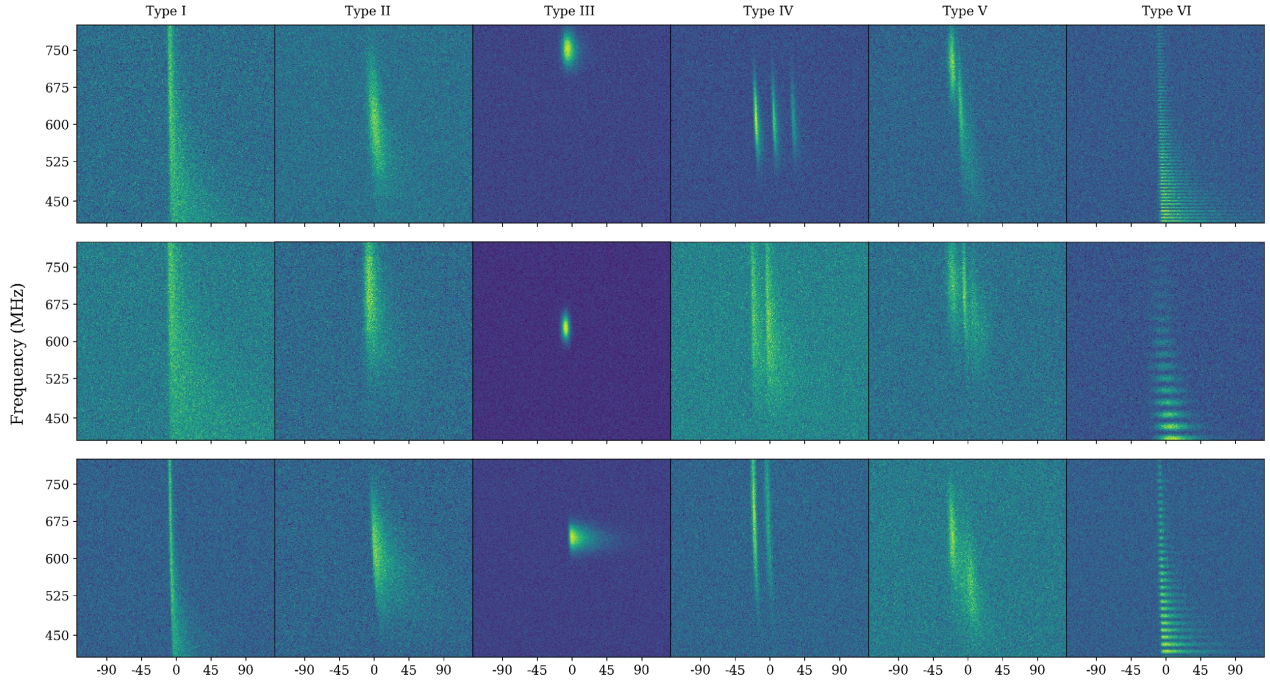
Repeaters look different from one-offs — and we want to know which is which fast.

Frabjous — DL for FRB morphology (Kumar+ 2025)

Six morphology archetypes (broadband, narrowband, multi-component, drifting, …).

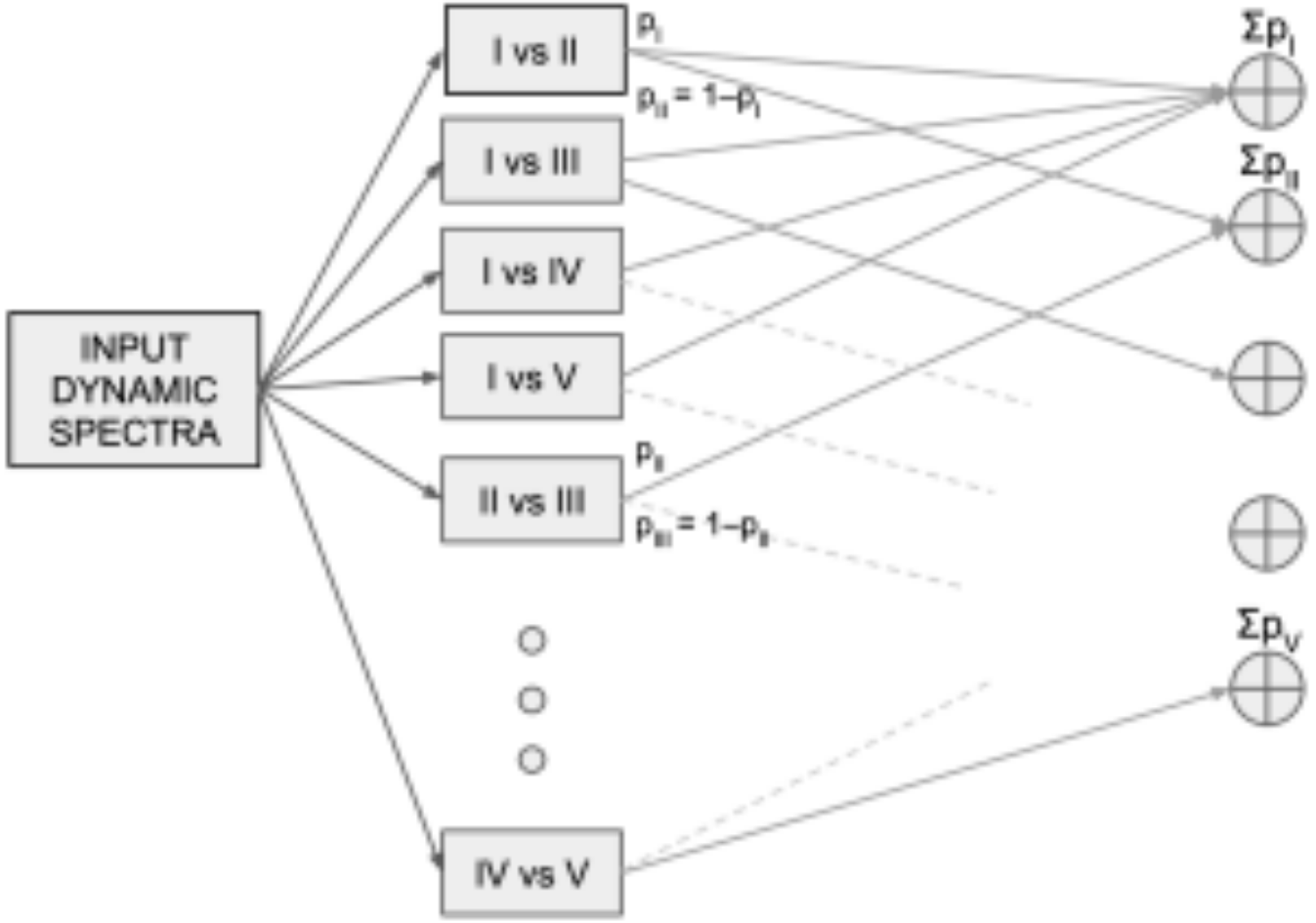
10 binary classifiers ⊕ confidence-voting matrix ⟶ 5-way label.

55% accuracy on real CHIME catalog (random baseline: 20%).



Six simulated FRB archetypes — arXiv:2507.14854

Frabjous — classifier ensemble + voting matrix



Binary classifier outputs aggregated via confidence-voting — arXiv:2507.14854

6. Language models

Why LLMs for astronomy?

Literature grows faster than humans can read.

Codebases, archives, manuals — all natural-language interfaces.

Multi-modal: text + light curves + images + spectra.

Agents that vet alerts, plan follow-up, draft reports.

But — can we trust them? Or, how much can we trust them?

AstroAlertBench (Chen+ 2026)

Benchmark of multimodal LLMs on real ZTF alerts (arXiv:2605.05573).

1,500 alerts × 13 frontier models. (1,500 is tiny)

Three evaluation stages:

metadata grounding → scientific reasoning → hierarchical classification.

High accuracy does **not** track "honesty" (self-evaluation of reasoning).

1. First-alert inputs

Image Triplet + Alert Metadata

Science Reference Difference

Representative Metadata

objectId	ZTF20example
fid	2 (r band)
magpsf	18.45
sgscore1	0.99

Three 63×63 cutouts centered on candidate position, plus prompt-facing fields extracted from the alert packet.

Serialize →

2. Prompt Construction

A Consistent Multimodal Prompt

Alert ID: ZTF20example

Images show a localized candidate residual.

Metadata: r band; magpsf = 18.45; sgscore1 = 0.99.

Task: Answer Parts A–C in the required format.

Includes:

- Image triplets
- Metadata
- Field definitions
- Output instructions

Evaluate →

3. Structured Model Response

Machine-parseable Parts A–C

VLM

Part A
filter_band: r
subtraction_sign: positive

Part B
leading_interpretation: supernova candidate

Part C
stage1: real_object
stage2: astrophysical
stage3: supernova

Scored on metadata grounding, rationale, and staged classification.

Cross-cutting lessons

Representation > architecture: DMDT, Q-transforms, image triplets.

Labels — not compute — are the real bottleneck. Active learning is everywhere.

Asymmetric error costs shape the loss (missed kilonova vs. extra alert).

Humans-in-the-loop are everywhere.

What's next — the LSST / multi-messenger era

LSST: ~10× ZTF rates for a decade, starting soon.

Foundation models across surveys and wavebands.

Joint EM + GW + neutrino inference, not separate pipelines.

Real-time decisions delegated to (audited) agents.

The toolbox built on ZTF + LIGO + TESS + CHIME generalises.

Interpretability and explainability

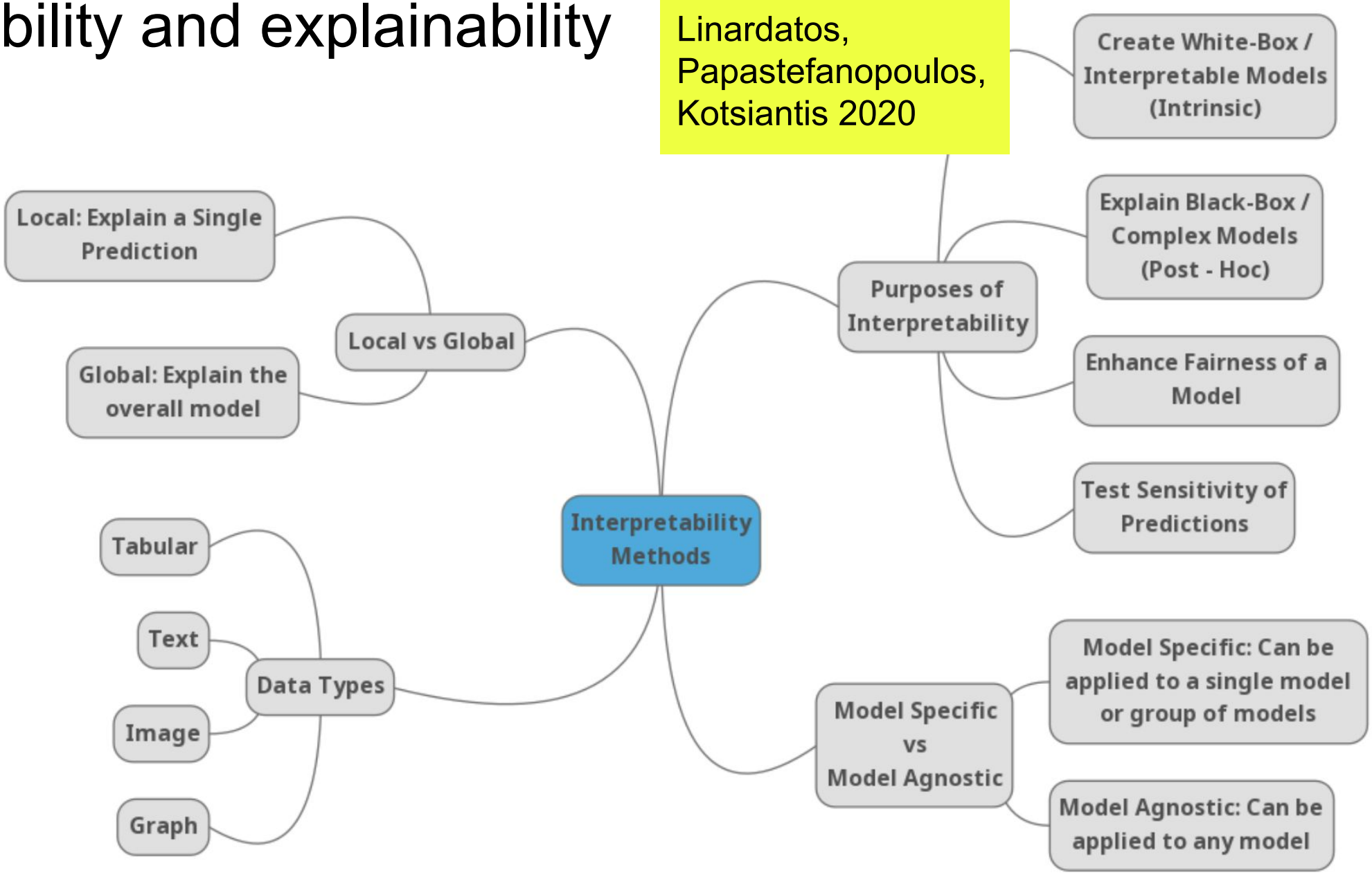
Linardatos, Papastefanopoulos, Kotsiantis 2020

Post-hoc

LIME: Local Interpretable Model-agnostic Explanations

SHAP: Shapley Additive Explanations

Fairness



References

Images: 1907.11259, 1904.05920, 1710.01422

Light curves: 1709.06257, 2009.14071, 2102.11304, 2312.00143, 2101.09227

Spectra: 2104.12980, 2412.08601

GW: 1910.12143, 2010.11829, 2111.04015, 2308.12357, 2408.06491, 2502.00297, 2605.00391

Radio: 2507.14854

LLMs: 2605.05573