Investigating Machine Learning Models for Maser Candidate Selection

Cameron Petrey[1], Claire Sprague[2], Dr. Nathan Sprague[1], Dr. Anca Constantin[3]

(1) Department of Computer Science, JMU ; (2) Department of Data Analytics, Dickinson College;
(3) Department of Physics & Astronomy, JMU · JMU REU Symposium, July 30 2026

Abstract

Water megamasers enable precise measurements of the Hubble constant and supermassive black hole masses, yet surveys detect them in only ~3% of galaxies, making efficient target selection critical when telescope time is scarce. We test whether machine learning can outperform current galaxy selection in prioritizing megamaser candidates via host galaxy traits that can be observed and measured in large sky surveys. The most current availability of multiwavelength observations of galaxies with and without maser emission provides only small and highly imbalanced samples, which makes overfitting during model selection a central concern. To address this, we evaluate logistic regression and random forest classifiers, using synthetic datasets with feature distributions resembling the real observations to guide model selection before applying the models to the real data. We compare classifiers by their precision on top-ranked candidates, measuring the degree to which each model concentrates true megamasers among its highest-ranked candidates. These results are crucial for planning future surveys for megamaser detections with the Green Bank Telescope.

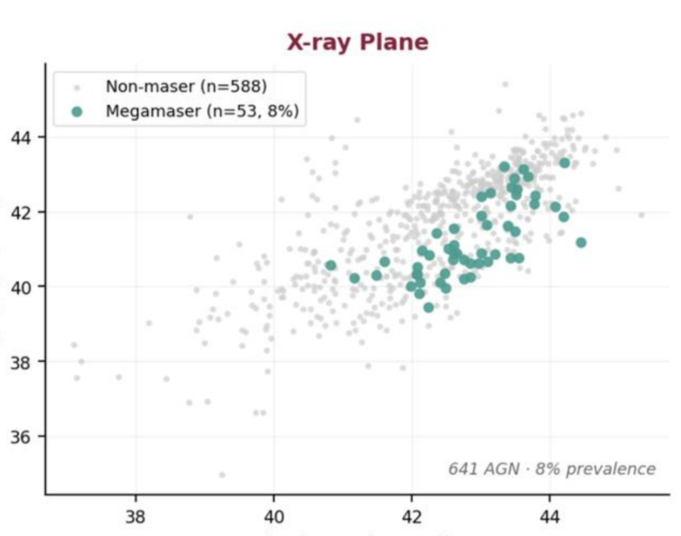
The Importance of Masers

- **Geometric distances:** orbital velocity + angular size
  → calculate distance without a luminosity calibration ladder
  → allow for the only fully geometric, assumption-free measurement of H_0, the Hubble constant, and thus the age of the universe
- **Black hole masses:** direct dynamical measurement from Keplerian disk kinematics

The Data: Multiwavelength Galaxy Features

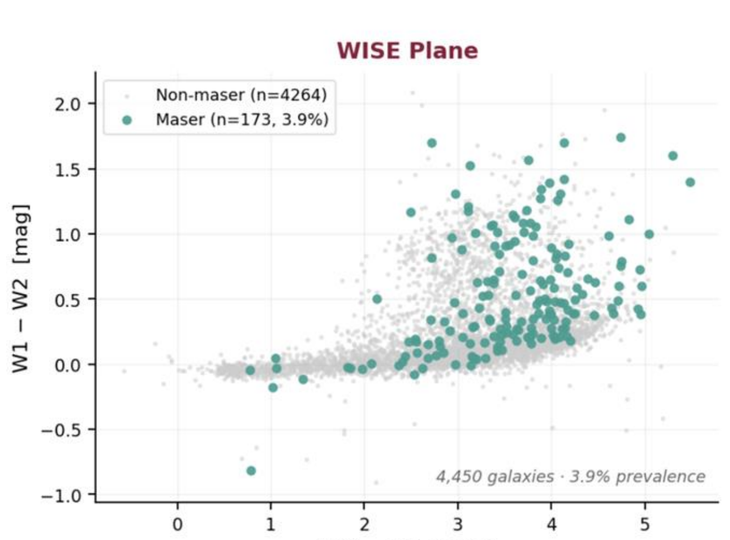
X-ray Plane
L_{12} — mid-IR luminosity
L_{ob} — observed X-ray luminosity

- 641 AGN with archival X-ray and mid-infrared observations (Kuo et al. 2020)
- ~8% maser prevalence
- feature correlation r = 0.74



WISE Plane
W1−W2 vs. W2−W3 (color-color diagram)

- mid-Infrared detections from the Wide-field Infrared Survey Explorer (WISE; Wright et al. 2010)
- ~4,400 galaxies from the AllWISE catalog
- ~3.3% maser prevalence
- larger but noisier sample



The Problem: Unbalanced Data

- **Class imbalance:** only ~55 positive labels from 641 galaxies (~8%)
- **Overfitting risk:** with fewer than 60 masers, standard cross-validation alone can be misleading for model selection (Cawley & Talbot 2010)

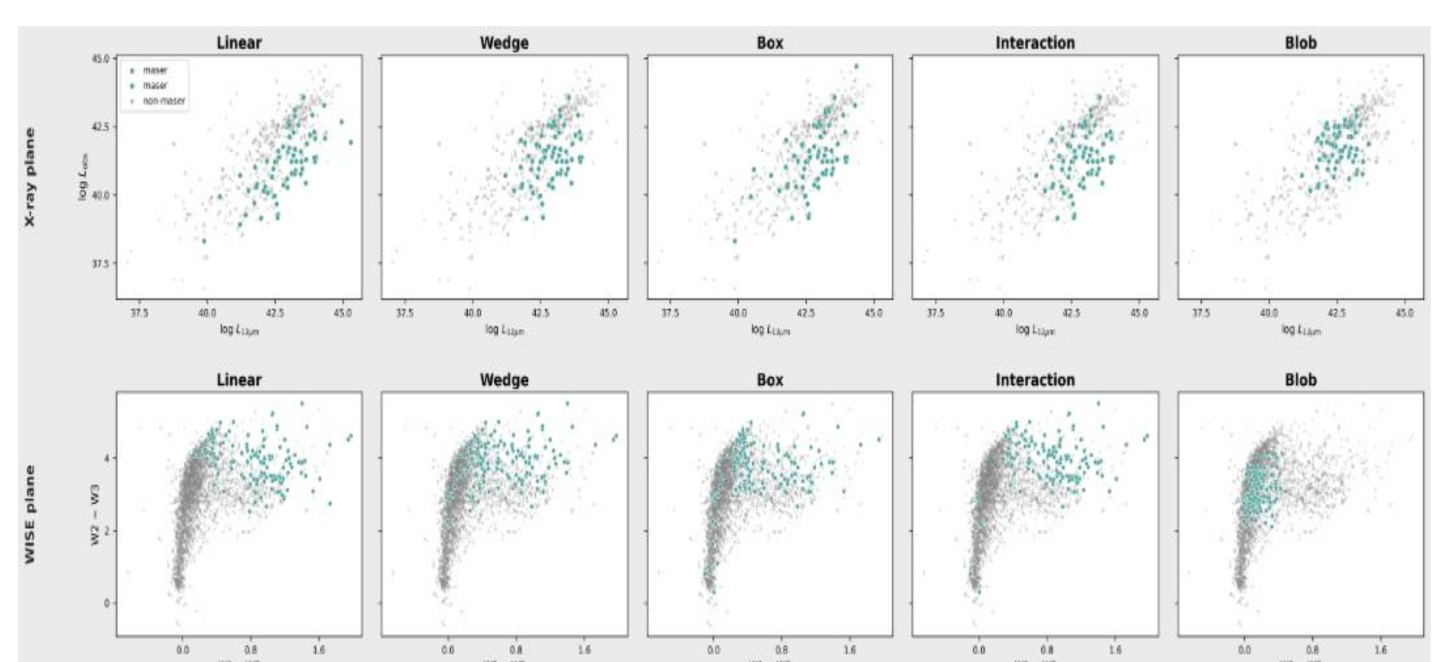
Key constraint: We cannot safely validate model selection directly on real data. A rigorous pre-observation evaluation is required.

Our Solution: Synthetic Data

We generate synthetic maser labels on real galaxy features resampled from the real columns with jitter, preserving distributions, correlations, sample sizes, and maser rates, under five plausible boundary geometries (Morris et al. 2019), evaluate classifiers against the known ground truth, then carry the winning model to real data.

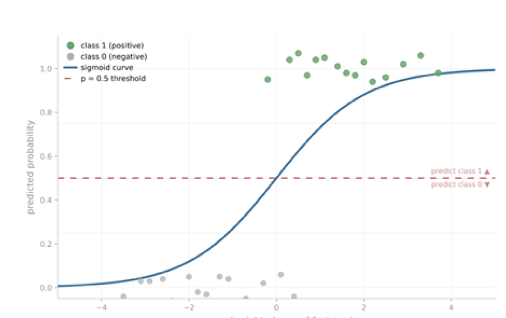
- Real galaxy measurements provide feature distributions; labels are synthetic
- Five synthetic boundary shapes are tested to ensure robustness
- 5-fold cross validation × 40 seeds per scenario for stable estimates

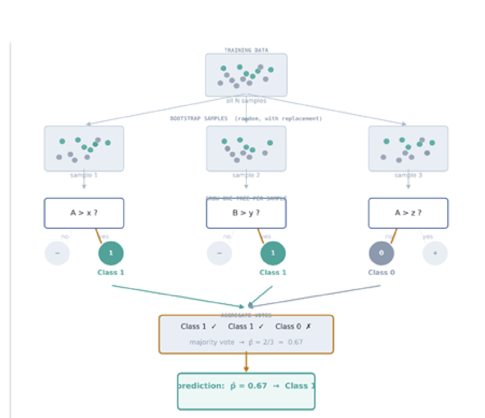
The Five Boundary Scenarios



Models Tested

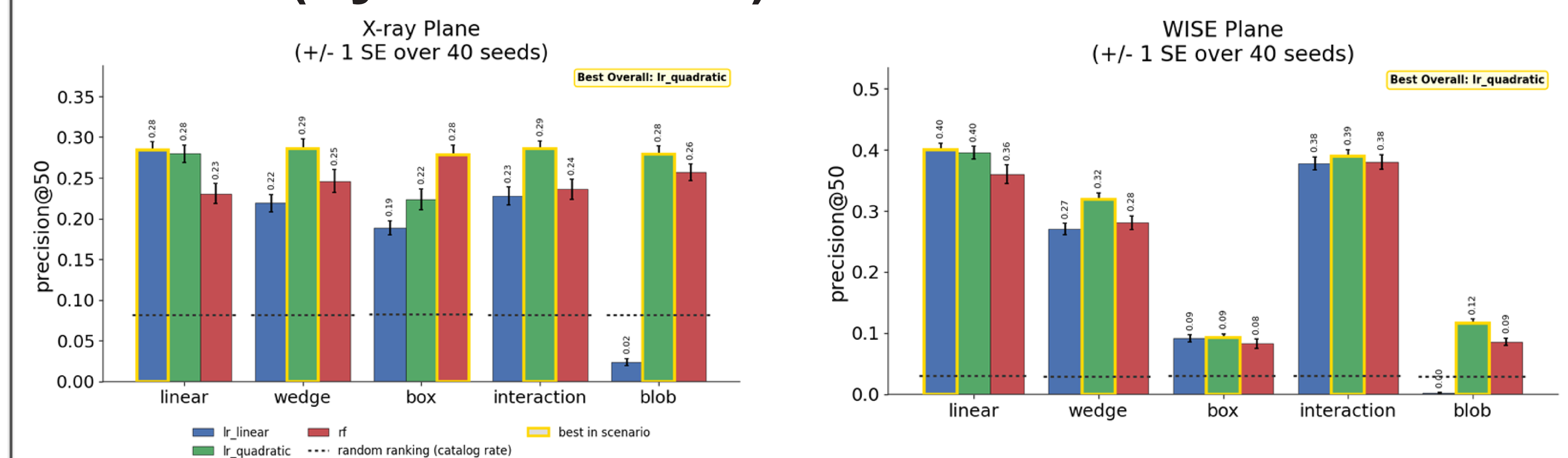
- **LR —** Logistic Regression: fits a single straight-line boundary separating masers from non-masers in feature space. Simple and interpretable, but fails whenever the true boundary is curved.
- **QLR —** Quadratic LR: same idea as LR, but first transforms the features into squared and cross-product combinations, allowing the boundary to be curved (elliptical or parabolic).
- **RF —** Random Forest: builds hundreds of decision trees — each one asking a sequence of yes/no questions about the features — then takes a majority vote. Individual trees are deliberately kept simple to avoid memorizing the small training set.





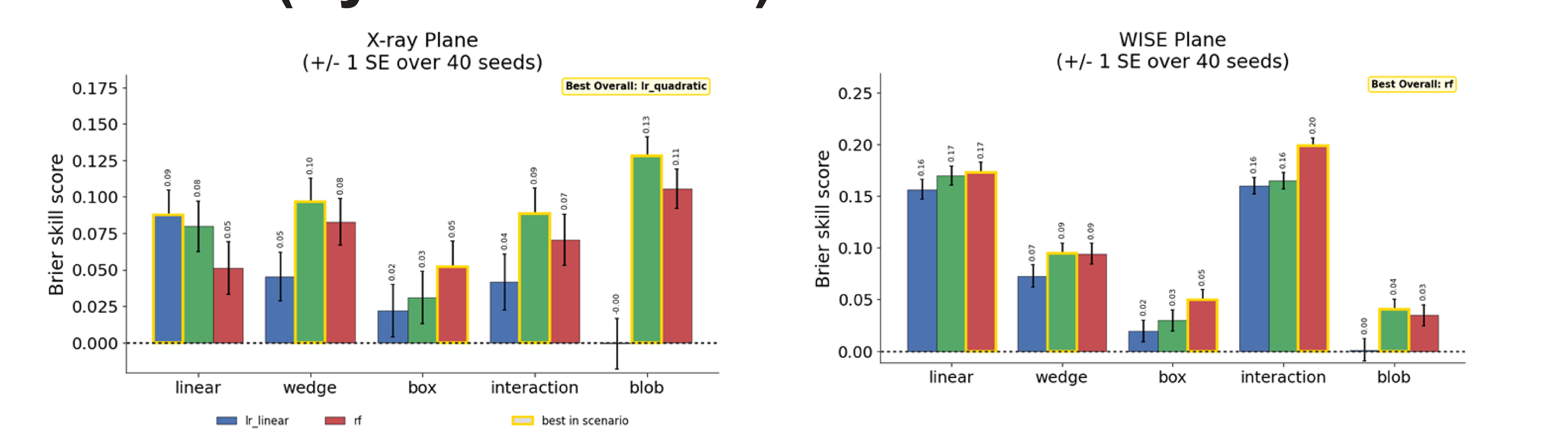
(Pedregosa et al. 2011)

Results (Synthetic Data) - Precision

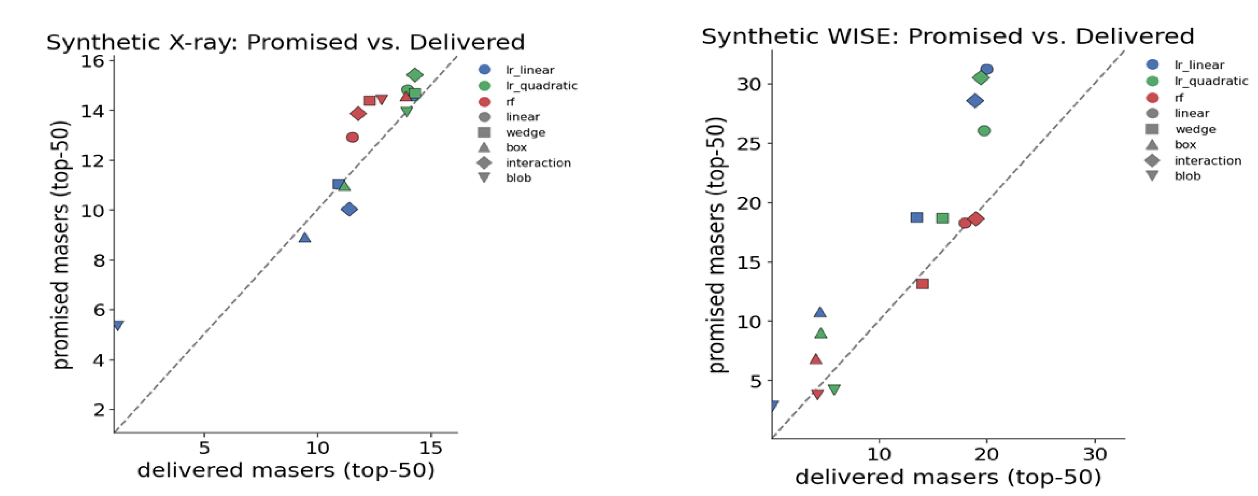


Precision @50: Fraction of true masers recovered in the top-50 highest-ranked candidates, averaged across 8 random seeds. Dashed line = random baseline (~8% X-ray, ~3.3% WISE).

Results (Synthetic Data) - Calibration



Brier Skill Score: how much better each model's probability estimates are compared to simply predicting the maser rate for every galaxy. Zero = no better than random; higher = more informative probabilities. QLR and LR are strongest across scenarios.



Promised vs Delivered: For each model, we sum predicted probabilities over the top-50 ranked candidates ("promised") and compare to the masers actually found there ("delivered"). A ratio of 1.00× means the model's confidence exactly matched reality. X-ray models range from overconfident to under-promising.

Conclusion

Quadratic logistic regression generally outperformed the other models in skill score and precision across both feature planes, though it tends to overpromise on the WISE plane. Its strong performance and simplicity make it the preferred candidate for ranking maser galaxies, and for applying it to the real data is the natural next step for this exploratory synthetic analysis.

References:

Cawley, G. C., & Talbot, N. L. C. 2010, Journal of Machine Learning Research, 11, 2079; **Kuo**, C. Y., et al. 2020, ApJ, 892, 18; **Morris**, T. P., White, I. R., & Crowther, M. J. 2019, Statistics in Medicine, 38, 2074; **Pedregosa**, F., et al. 2011, Journal of Machine Learning Research, 12, 2825; **Wright** et al. 2010, AJ, 140, 1868