Random Forest Classification of 22 GHz Maser Galaxies: A Database Analysis of GBT Observations

by
Emily McMahon

Instructor: Dr. Jim Braatz
National Radio Astronomy Observatory
University of Virginia

This thesis is submitted in partial completion of the requirements of the
BS Astronomy-Physics Major

Abstract

To date, approximately 180 extragalactic 22 GHz H_2O masers have been detected, with 115 of these discovered using the Green Bank Telescope (GBT). Of the GBT detections, 80 were identified by the Megamaser Cosmology Project (MCP), highlighting its significant contribution to the field. These discoveries come from a survey of approximately 4,600 galaxies using the GBT, putting the rate of discovery at 2.5%. These rare detections are of exceptional value due to their relevance to precision cosmology. They enable direct measurements of supermassive black hole (SMBH) masses and geometric distances to their host galaxies, both of which are crucial for refining estimates of the Hubble constant. In order to increase the likelihood of identifying these valuable galaxies, we developed a machine learning model designed to accurately classify galaxies based on their detection status. Specifically, we employed a random forest model that analyzes WISE band colors (W1-W2, W2-W3, W3-W4) and near-infrared magnitudes (j, h, k) to predict the presence of 22 GHz H_2O megamasers. The accuracy of our model after cross validation averages to approximately 74%.

1 Introduction

The detection of 22 GHz water vapor megamasers in the circumlunar accretion disks of active galactic nuclei (AGN) provides a method to measure the geometric distances to these galaxies. The MCP leverages this technique to estimate the Hubble constant (H_0) without relying on conventional distance measurement methods such as standard candles or distance ladders. These rare systems are of immense importance, offering both precise measurements of SMBH masses and geometric distances to their host galaxies–critical factors in refining H_0.

Currently, approximately 180 extragalactic 22 GHz H_2O masers have been detected. Of these, 115 were discovered using the GBT, and 80 of those were found through the efforts of the MCP. These detections stem from a GBT survey targeting about 4,600 galaxies, placing the overall detection rate at a mere 2.5%. This low yield underscores the difficulty of identifying suitable candidates using traditional selection methods and motivates the need for improved targeting strategies.

We find these megamasers by surveying nearby AGNs, with a specific lens on Seyfert 2 galaxies. Much of the motivation for our project stems from the findings of Kuo et al. (2018), which demonstrated that optically obscured active galactic nuclei (AGNs)–often missed in optical surveys due to dust extinction–can be effectively identified using mid-infrared data from the all-sky WISE survey (Wright et al., 2010). The study revealed that specific WISE color indices, along with optical u–r colors, are statistically significant predictors of 22 GHz H_2O megamaser detections, highlighting a promising approach for identifying maser-hosting galaxies. Figure 1 from this paper demonstrates the significance of maser emission as a function of colors, which theorizes a 9% megamaser detection rate.

We replicated this graph in Figure 2 for all extragalactic observations made by the GBT, highlighting the maser galaxies. In both of these figures, we see a distinct relationship between maser detection and color. However, the underlying relationship between these photometric properties and maser activity is not yet fully understood and remains a subject of ongoing investigation. Motivated by this gap, we sought to determine whether machine

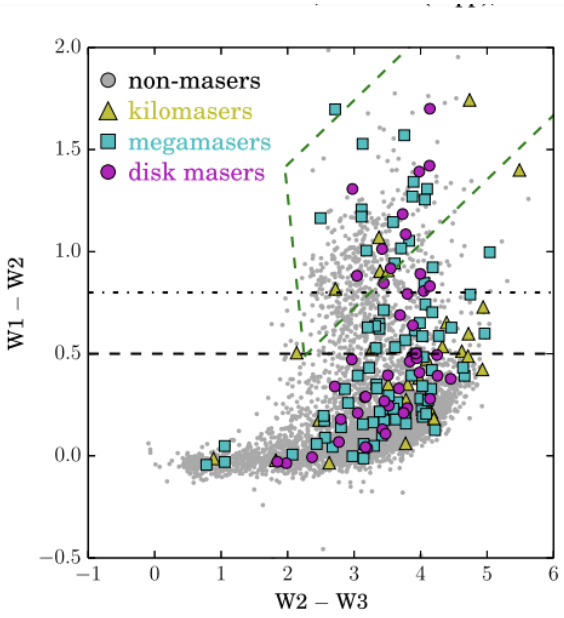

Figure 1: WISE color–color diagram of MCP galaxies. Distribution of maser (detections) and non-maser (non-detections) galaxies in the W2−W3 versus W1−W2 color space. The horizontal dashed and dot-dashed lines correspond to W1−W2 = 0.5 and 0.8, respectively. The green dashed line indicates the AGN wedge defined by Mateos et al. (2012). *Figure reproduced from Kuo et al. (2018).*

learning algorithms could leverage these relatively underutilized galactic features to classify galaxies with confirmed maser detections, thereby assessing their utility as predictive tools within a broader astrophysical framework.

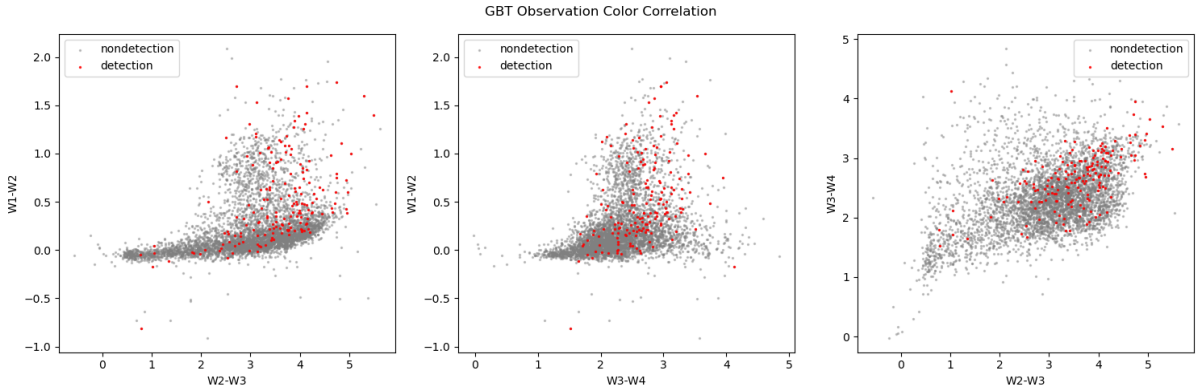

Figure 2: Three-panel plot showing the distribution of MCP maser and non-maser galaxies in different color-color diagrams: (left) W2-W3 vs. W1-W2, (middle) W3-W4 vs. W1-W2, and (right) W2-W3 vs. W3-W4. In all panels, gray points represent observations from the GBT, while red points indicate maser detections.

To select a robust machine learning model for our analysis, we drew inspiration from the work of Daoutis et al., (forthcoming), who demonstrated the effectiveness of random forest classifiers in distinguishing between star-forming, AGN, low-ionization nuclear emission-line region (LINER), composite, and passive galaxies, particularly within local and low-redshift samples. Their study utilized similar classification features, including infrared colors derived from the first three WISE bands (W1, W2, and W3) and optical colors from the u, g, and r

bands of the SDSS survey. Achieving an overall classification accuracy of 81%, their results motivated us to adopt a random forest model tailored to our own set of specified parameters.

2 Methods

We initiated our analysis by constructing a balanced dataset containing approximately equal numbers of galaxies with and without maser detections, as observed by the GBT. We then cross-matched these samples with the ALLWISE galaxy catalog to retrieve our parameters of interest: W1, W2, W3, W4, and the near-infrared bands j, h, and k. Of the 180 maser detection galaxies, 171 were within the ALLWISE survey database and contained IR data, allowing us to retain nearly all detection galaxies in our final dataset. To ensure comparability, we selected a matched sample of nondetection galaxies with similar photometric and physical characteristics. ALLWISE contained IR data for 135 of these matched nondetection galaxies. Given that apparent magnitude is distance-dependent, we further minimized potential biases by limiting our sample to galaxies within a specified distance range. Furthermore, to reduce the influence of distance-related effects, we converted the WISE band magnitudes into color indices—specifically W1–W2, W2–W3, and W3–W4—while retaining the magnitudes j, h, and k as individual features.

We began by splitting the dataset into training and testing sets using the `train_test_split` function from scikit-learn, allocating 80% of the data for training and 20% for testing. Stratified sampling was applied to ensure that the distribution of the binary target variable, detection, was maintained across both sets.

To estimate the model's performance and avoid overfitting, we employed 5-fold cross-validation on the training set using StratifiedKFold. This method ensured that each fold had a representative distribution of both detection and non-detection classes. The average cross-validation accuracy was computed to assess the generalization ability of the model on the training data.

After cross-validation, the Random Forest classifier was retrained on the full training set. The model's final performance was evaluated using the held-out test set, which had not been used during training or cross-validation. The accuracy of the model on the test set was reported as the final evaluation metric.

The Random Forest classifier was used with default hyperparameters, and class balancing (`class_weight`='balanced') was applied to address our minor class imbalance. The model's final evaluation metrics included the average cross-validation accuracy, the accuracy on the test set, and a heatmap confusion matrix to visualize the specific areas in which our model failed.

3 Results

We began our model evaluation by calculating the accuracy score. As previously noted, we applied 5-fold cross-validation, which yielded an average accuracy of 74%. We produced a cumulative heatmap confusion matrix to visualize our model's performance metrics as can be see in Figure 3. A definition for each of these metrics can be found in Table 1. Figure 3

highlights our model's high sensitivity in detecting true positives, which in part may be due to the higher number of detection galaxies contained within our dataset.

Actual / Predicted	**Predicted Positive**	**Predicted Negative**
Actual Positive	**TP**: Correctly predicted detection	**FN**: Missed detection
Actual Negative	**FP**: Incorrect detection	**TN**: Correctly predicted non-detection

Table 1: Definitions of the four categories in a confusion matrix.

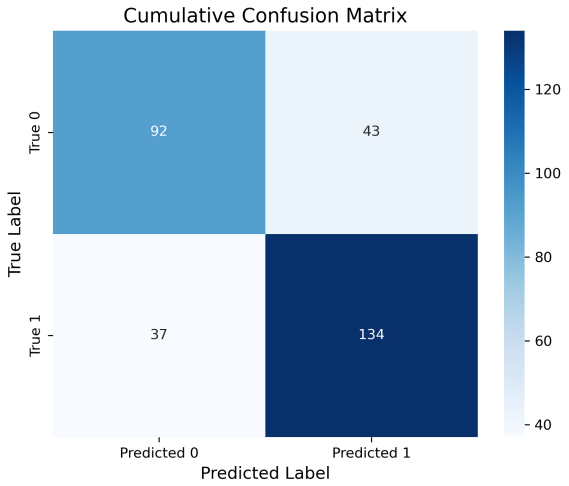

Figure 3: Heatmap of the confusion matrix showing the performance of the Random Forest classifier, with true versus predicted classifications of maser detection status. The matrix displays the true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) based on the model's predictions, with darker shades indicating higher values in each category.

In evaluating our model, we considered three important performance metrics: precision, recall, and the F1 score. While the scores vary slightly between folds, they remain relatively consistent, indicating that the model generalizes well across subsets of the data. On average, the model achieved a precision of 0.76, recall of 0.72, and an F1 score of 0.74, suggesting a strong balance between sensitivity and predictive accuracy.

Precision tells us how reliable our model's positive predictions are and can be calculated as follows:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

In the context of our project, the precision is the number of true detections out of all galaxies that our model classified as detections.

Recall measures how well the model classifies positives. More specifically, this metric tells us how many detections the model was able to correctly classify and can be calculated below:

Fold	Precision	Recall	F1 Score
1	0.75	0.70	0.72
2	0.78	0.73	0.75
3	0.76	0.72	0.74
4	0.74	0.69	0.71
5	0.77	0.75	0.76
Average	**0.76**	**0.72**	**0.74**

Table 2: Performance metrics (Precision, Recall, and F1 Score) across five cross-validation folds. Averages are reported in the final row.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Finally, the F1 score is perhaps the most important metric when evaluating a model's performance, as it finds the balance between precision and recall by combining the two:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1 score is calculated as the harmonic mean of precision and recall, meaning that if either precision or recall is low, the F1 score will also be low. This makes the F1 score particularly useful —- especially in cases of imbalanced datasets such as ours, where detections are more abundant than nondetections. With an average F1 score of 0.74, we can be confident that our model is striking a good balance between precision and recall.

In the context of our project, precision shows that when our model predicts a galaxy as a maser detection, it's usually correct. On the other hand, recall indicates that the model is catching most of the true detections, but there are still some that it misses. By combining the two metrics to obtain an F1 score of 0.74, we know that our model is not overly focused on one at the expense of the other. This is crucial, as it is not only important to make predictions, but to ensure they are reliable and comprehensive.

To further investigate the areas in which our model was failing, we produced a table of misclassified galaxies for a singular evaluation. Included in the table was the designation as referenced in the MCP catalog, the designation as referenced in ALLWISE, the input parameters used during testing and training, the true label indicating detection status, the predicted label, and a confidence level. This confidence level was constructed using the `predict_proba` method provided by the `RandomForestClassifier` classifier included in scikit-learn. The `predict_proba` function estimates class probabilities by aggregating the outputs of all decision trees in the random forest, calculating the proportion of trees that vote for each class. This proportion reflects the model's confidence in each possible outcome, providing a probabilistic interpretation of its predictions. This information is referenced in Table 3. The average confidence score for this subset of misclassified galaxies is 0.44. A confidence level below 50% for misclassifications indicates that our model was generally uncertain when making incorrect predictions. This suggests that the model is not confidently wrong, which is preferable to high-confidence misclassifications and may point to inherently ambiguous or borderline cases.

	source	designation	ra	dec	w1mpro	w2mpro	w3mpro	w4mpro	j_m	h_m	k_m	W1-W2	W2-W3	W3-W4	detection	Predicted Label	Confidence
60	J0912+2304	J091246.36+230427.3	138.19	23.07	13.35	13.19	9.9	8.02	15.02	14.35	14.06	0.16	3.28	1.88	1	0	0.09
33	J0350-0127	J035000.35-012757.5	57.5	-1.47	11.41	10.78	7.58	5.54	14.29	13.52	12.78	0.63	3.21	2.04	1	0	0.5
118	MCG+11-17-010	J135628.69+643743.9	209.12	64.63	12.08	11.73	8.4	6.62	14.62	13.85	13.44	0.35	3.34	1.78	1	0	0.39
147	NGC 6264	J165716.13+275058.5	254.32	27.85	12.16	12.05	8.57	5.44	14.24	13.49	13.3	0.11	3.47	3.14	1	0	0.38
76	J1028+1046	J102802.91+104630.4	157.01	10.78	12.05	11.87	8.25	5.79	13.92	13.85	13.4	0.18	3.63	2.46	1	0	0.45
24	J0253-0014	J025329.59-001405.4	43.37	-0.23	13.23	13.02	8.85	5.35	15.02	14.22	13.95	0.21	4.16	3.51	1	0	0.38
284	['J160707+220338']	J160707.00+220338.1	241.78	22.06	10.81	10.61	6.77	4.38	13.76	13.13	12.6	0.2	3.85	2.38	0	1	0.69
295	['Ark539']	J182848.07+502220.8	277.2	50.37	11.44	11.24	8.47	5.96	13.37	12.5	12.31	0.2	2.78	2.51	0	1	0.64

Table 3: Summary of Misclassified Galaxies Including Their Predicted Labels and Associated Confidence Scores from the Classifier

4 Discussion

The Random Forest Model achieved an average F1 score of 0.74, demonstrating a strong balance between precision and recall. This suggests that the model is fairly effective at distinguishing maser galaxies, or detections, from nondetections using photometric features alone. The relatively high precision indicates that most of the galaxies predicted to be detections do truly harbor masers within, while the recall value shows that a majority of actual maser galaxies are correctly identified.

We originally hypothesized that infrared colors and near-infrared magnitudes can serve as reliable predictors in the detection of maser emission in Active Galactic Nuclei. This hypothesis was supported by the model's ability to perform above chance–a 50% accuracy. The importance of features like W1-W2, W2-W3, and W3-W4 aligns with prior studies that emphasized the role of dust-obscured AGNs in maser emission.

The average confidence of 0.44 in misclassified galaxies suggests that the model is not confidently wrong. Conversely, the misclassifications appear to be made with a degree of uncertainty, which indicates that these galaxies may possess ambiguous or borderline photometric characteristics. This is rather encouraging, as it implies that the model is cautious in cases where the input data is less clearly separable, a preferable behavior in astrophysical classification instances.

Our classification accuracy of approximately 74% stands out among similar efforts, such as the 81% achieved by Daoutis et al. (forthcoming) using optical and infrared features. Differences in sample size, feature selection, and class distribution may explain this discrepancy. It should be noted that our model relied solely on photometric data without incorporating morphological or spectroscopic features, which could potentially enhance model performance.

A key limitation within this study is the relatively small number of confirmed maser galaxies, which constrained the size of the balanced dataset. Additionally, the exclusion of certain potentially informative features–excluded simply because of their unknown relevance to maser activity–may have limited the model's discriminative power. The model also treats each galaxy independently and does not account for observational bias such as detection sensitivity limits or survey coverage. While our model may have broader potential across a wide range of parameters, this remains an open question requiring further investigation.

While our model demonstrates promising results using only WISE color indices and near-infrared fluxes, there is significant opportunity to improve its predictive power by incorporating additional multi wavelength data. Future work could expand on this analysis by integrating features from other domains known to correlate with AGN activity and maser emission. These include optical spectroscopic features such as emission line strengths and diagnostic line ratios seen in Figure 4, which have historically been effective in classifying galaxy nuclear activity Groves et al. (2006).

X-ray observations also hold considerable promise. Parameters such as X-ray luminosity and absorption column densities can provide insight into obscured AGN populations that might not be apparent in optical or infrared data alone (Brandt et al. (2005)). Similarly, radio continuum properties–including flux, spectral index, and morphology–may offer additional discriminative power, especially in identifying galaxies with compact radio cores indicative of AGN activity (Condon et. al (1992)).

Incorporating these additional features could potentially improve model performance by identifying maser-hosted galaxies that are missed by IR-based selection alone. A multi-model approach combining photometric, spectroscopic, and high-energy data may also reduce misclassifications of ambiguous cases and allow for more robust candidate selection in future maser surveys. Furthermore, applying more advanced models such as deep neural networks could be explored–particularly if the size of the labeled dataset increases through future detections.

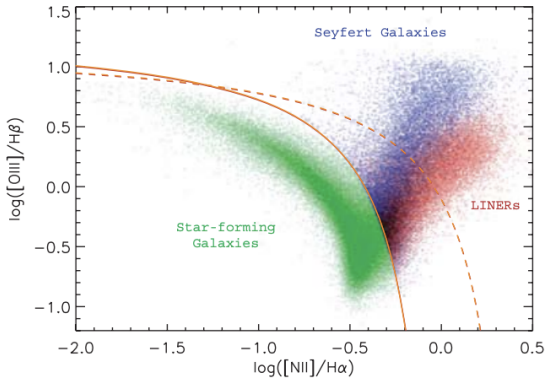

Figure 4: Diagnostic diagram of [N II] λ6584 Å / Hα versus [O III] λ5007 Å / Hβ for SDSS emission-line galaxies, illustrating the classification of galaxy nuclear activity. Star-forming galaxies are shown in green, Seyfert 2s in blue, and LINERs in red. The solid and dashed curves represent the Kauffmann and Kewley separators, respectively. These optical spectroscopic features, as shown here, have been historically useful in classifying AGN activity and maser emission. Reprinted from Groves et al. (2006).

5 Conclusion

This study represents a data-driven framework to improve the efficiency of future megamaser surveys by enhancing target selection strategies. Traditional approaches, which predominantly focus on Seyfert 2 galaxies, typically yield a low detection rate of around 2.5%. To address this limitation, we developed a machine learning-based method that integrates multi-wavelength observational data to identify galaxies with a higher likelihood of hosting H_2O megamasers. By training Random Forest classifiers on a labeled sample of known maser and non-maser galaxies, our algorithm learns to recognize complex, non-linear relationships across observational dimensions that may elude conventional selection methods.

Ultimately, this approach offers a pathway to significantly surpass the historical detection efficiency and to optimize the scientific return of future maser surveys. Beyond improving detection rates, this work also contributes to a deeper understanding of the physical conditions

and AGN environments that are conducive to maser emission, advancing both observational strategy and theoretical insight in the study of extragalactic water megamasers.

References

[1] Kuo, C. Y., Constantin, A., Braatz, J. A., et al. 2018, ApJ, 860, 169

[2] Wright, E. L., et al. 2010, *The Astrophysical Journal*, 140, 1868–1881.

[3] Daoutis, C., Kyritsis, E., Kouroumpatzakis, K., and Zezas, A. 2023, A versatile classification tool for galactic activity using optical and infrared colors, *Astronomy & Astrophysics*, (forthcoming).

[4] Groves B. A., Heckman T. M., Kauffmann G., 2006, Emission-line diagnostics of low-metallicity active galactic nuclei, MNRAS, 371, 1559

[5] Brandt W. N., Hasinger G., 2005, X-ray surveys and AGN, ARA&A, 43, 827

[6] Condon J. J., 1992, The extragalactic radio source population, ARA&A, 30, 575