Identifying AGN Host Galaxies by Machine Learning with HSC+WISE

Yu-Yen Chang1,2, Bau-Ching Hsieh2,6, Wei-Hao Wang2,6, Yen-Ting Lin2,6, Chen-Fatt Lim2,3,6, Yoshihiko Toba2,4,5, Yuxing Zhong5, and Siou-Yu Chang1

1 Department of Physics, National Chung Hsing University, 40227, Taichung, Taiwan; yurenchang.astro@gmail.com
2 Academia Sinica Institute of Astronomy and Astrophysics, P.O. Box 23–141, Taipei 10617, Taiwan
3 Graduate Institute of Astrophysics, National Taiwan University, Taipei 10617, Taiwan
4 Department of Astronomy, Kyoto University, Kitashirakawa-Oiwake-cho, Sakyo-ku, Kyoto 606–8502, Japan
5 Research Center for Space and Cosmic Evolution, Ehime University, 2–5 Bunkyo-cho, Matsuyama, Ehime 790–8577, Japan
6 Department of Physics, Waseda University, 1-6-1 Nishiwaseda Shinjuku-ku, Tokyo 169–0850, Japan

Received 2020 December 7; revised 2021 July 7; accepted 2021 July 19; published 2021 October 15

Abstract

We investigate the performance of machine-learning techniques in classifying active galactic nuclei (AGNs), including X-ray-selected AGNs (XAGNs), infrared-selected AGNs (IRAGNs), and radio-selected AGNs (RAGNs). Using the known physical parameters in the Cosmic Evolution Survey (COSMOS) field, we are able to create quality training samples in the region of the Hyper Suprime-Cam (HSC) survey. We compare several Python packages (e.g., scikit-learn, Keras, and XGBoost) and use XGBoost to identify AGNs and show the performance (e.g., accuracy, precision, recall, F1 score, and AUROC). Our results indicate that the performance is high for bright XAGN and IRAGN host galaxies. The combination of the HSC (optical) information with the Wide-field Infrared Survey Explorer band 1 and band 2 (near-infrared) information performs well to identify AGN hosts. For both type 1 (broad-line) XAGNs and type 1 (unobscured) IRAGNs, the performance is very good by using optical-to-infrared information. These results can apply to the five-band data from the wide regions of the HSC survey and future all-sky surveys.

Unified Astronomy Thesaurus concepts: Astronomy data analysis (1858); Active galaxies (17); Galaxies (573); Surveys (1671)

Supporting material: machine-readable table

1. Introduction

The processes driving the coevolution of galaxies, active galactic nuclei (AGNs), and their supermassive black holes remain a largely debated issue in extragalactic astrophysics. Specifically, the connection between the formation of stars in galaxies and the fueling of their central black holes is still not fully understood. Different techniques have been investigated to select and identify various AGN samples (e.g., Baldwin et al. 1981; Urry & Padovani 1995; Kauffmann et al. 2003; Lacy et al. 2004; Stern et al. 2005; Alonso-Herrero et al. 2006; Kewley et al. 2006; Tozzi et al. 2006; Ho 2008; Burlon et al. 2011; Juneau et al. 2011; Fabian 2012; Padovani et al. 2017; Hickox & Alexander 2018, and references therein). However, it is still difficult to properly select AGNs and avoid sample selection bias.

The increasing amount of astronomical data has led to a need for machine-learning (ML) methods. For example, decision trees (e.g., Ball et al. 2006; Vasconcellos et al. 2011; Sevilla-Noarbe & Etayo-Sotos 2015), support vector machines (e.g., Kovács & Szapudi 2015), neural networks (NNs; e.g., Odewahn et al. 1992; Bertin &Arnouts 1996), convolutional neural networks (CNNs; e.g., Dieleman et al. 2015; Huertas-Company et al. 2015; Krakowski et al. 2016; Kim & Brunner 2017; Domínguez-Sánchez et al. 2018), and XGBoost (e.g., Liu et al. 2019; Calderon & Berlind 2019; Lim et al. 2020; Golob et al. 2021) have been used to deal with data in astronomy and astrophysics.

Recently, ML has been applied to derive various physical parameters of galaxies (e.g., Masters et al. 2015; Krakowski et al. 2016; D’Isanto & Polsterer 2018; Hemmati et al. 2019; Bonjean et al. 2019; Davidzon et al. 2019). In particular, classification is one of the important issues of galaxy and AGN properties (e.g., Padovani et al. 2017; Hickox & Alexander 2018, and references therein). Therefore, ML can be used to efficiently classify AGNs (e.g., Carballo et al. 2008; Doert & Errando 2014; Kim et al. 2011; Cavuoti et al. 2014; Saz Parkinson et al. 2016; Fotopoulou & Paltani 2018; Bai et al. 2019; Faist et al. 2019; Chen et al. 2021; Poliszczuk et al. 2021) and will be helpful for us to identify different kinds of AGN hosts from non-AGN galaxies.

In this paper, we will compare several algorithms from scikit-learn, Keras, and XGBoost and show our results by using the state-of-the-art ML methods from XGBoost to identify X-ray-selected AGN (XAGN), infrared-selected AGN (IRAGN), and radio-selected AGN (RAGN) host galaxies with Hyper Suprime-Cam (HSC) and Wide-field Infrared Survey Explorer (WISE) data. The structure of this paper is as follows. We describe the data and sample selection in Section 2. We analyze the properties in Section 3. We discuss the results in Section 4 and summarize in Section 5. Throughout the paper, we use AB magnitudes, adopt the cosmological parameters ($\Omega_M, \Omega_L, h = (0.30, 0.70, 0.70)$), and assume the stellar initial mass function of Chabrier (2003).

2. Data

We use the Wide layer of optical photometry from the HSC (Miyazaki et al. 2012) Subaru Strategic Program7 (SSP). The HSC-SSP is an optical imaging survey with five broadband filters (grizy-band) and four narrowband filters (see Aihara et al. 2018; Bosch et al. 2018). The HSC-SSP consists of three

7 http://hsc.mtk.nao.ac.jp/
layers: Wide, Deep, and UltraDeep. This work uses S19a wide-layer data (s19a_wide.forced; e.g., Miyazaki et al. 2018; Aihara et al. 2018, 2019; Toba et al. 2019). The HSC-SSP wide layer covers six fields (XMM-LSS, GAMA09H, WIDE12H, GAMA15H, HECTOMAP, and VVDS). The typical seeing is about 0′′.6 in the i band, and the astrometric uncertainty is about 40 mas in rms. There are 2,802,212 objects in the S19a Wide (HSC-Wide) catalog. Among them, 705,161 objects have photometric redshift (photoz_best in s19a_wide.specz) information, and 62,567 objects have spectroscopic redshift (specz_redshift in s19a_wide.specz) information.

We match the HSC-Wide catalog with the 1,182,108 objects in the COSMOS 2015 catalog (Laigle et al. 2016) in 1′′, and 660,352 objects remain. The spectral energy distribution (SED) fitting properties are from Chang et al. (2017a), which were derived by the MAGPHYS+AGN fitting technique (da Cunha et al. 2008, 2015) in the Cosmic Evolution Survey (COSMOS; Scoville et al. 2007).

We include infrared information by using ALLWISE data (Wright et al. 2010; Mainzer et al. 2011) at 3.4, 4.6, 12, and 22 μm (W1, W2, W3, and W4), identified as the brightest object within a search radius of 6′′, similar to the WISE point-spread function and the Sloan Digital Sky Survey (SDSS) +WISE sample in Chang et al. (2015). In order to avoid missing value problem, we focus on 32,204 galaxies with available photometric redshifts (e.g., Hsieh & Yee 2014; Tanaka 2015; we adopted the MIZUKI photometric redshift in this paper), as well as detections in the HSC grizy bands, W1, and W2. To reach better performance, we also investigated a subsample of 8193 bright galaxies with $R_{\text{mag}} < 23$ (see Section 3.3 for more details). The redshift distribution is shown in Figure 1.

The matched catalog contains optical information from HSC-Wide and infrared information from ALLWISE in the COSMOS field, which provides the training and test sets in our sample. We separate the sample into XAGNs (625 XAGNs; $L_X(2-10 \text{ keV}) > 10^{42} \text{ erg s}^{-1}$, absorption-corrected), IRAGNs (404 IRAGNs; using the mid-infrared colors from the IRAC 3.6, 5.8, 4.5, and 8.0 μm as defined in Chang et al. 2017), RAGNs (2225 RAGNs; $>3\sigma$ radio excess in log($L_{1.4\text{GHz}}$/SFR$_{\text{IR}}$) as defined in Smolčić et al. 2017; Delvecchio et al. 2017), and non-AGN galaxies (29,502 GALs) that are identified by previous work (Civano et al. 2016; Marchesi et al. 2016; Chang et al. 2017a, 2017b; Smolčić et al. 2017; Delvecchio et al. 2017). As shown in Figure 2, the total number of AGNs is 2702, and some AGNs belong to two or three AGN selection methods.Magnitude distributions of the 625 XAGNs, 404 IRAGNs, 2225 AGNs, 29,502 non-AGN galaxies, and 32,204 parent sample objects are shown in Figure 3.

In this paper, photometric redshifts from HSC-Wide are adopted for the 32,204 parent sample objects, including the 2702 AGNs. We check the photometric redshift quality of our sample by calculating the catastrophic error ($\eta = |z_2 - z_1|/(1 + z_1) > 0.15$) and the redshift accuracy ($\sigma_{\Delta z/\langle 1+z \rangle} = |z_2 - z_1|/\langle 1+z \rangle$) similar to Ilbert et al. (2009), where $z_2$ is the photometric redshift from the HSC-Wide catalog and $z_1$ is the redshift from the compared sample. Comparing to the available photometric redshift in the COSMOS 2015 catalog, the catastrophic error and redshift accuracy in the AGN sample (625 XAGNs, $\eta = 30.1\%$ and $\sigma_{\Delta z/\langle 1+z \rangle} = 0.0847$; 400 IRAGNs, $\eta = 39.2\%$ and $\sigma_{\Delta z/\langle 1+z \rangle} = 0.1082$; 1942 RAGNs, $\eta = 15.4\%$ and $\sigma_{\Delta z/\langle 1+z \rangle} = 0.0660$) are not small but not far from the whole parent sample (22,721 objects, $\eta = 23.8\%$ and $\sigma_{\Delta z/\langle 1+z \rangle} = 0.0744$). The photometric redshift errors show a dependence on magnitude, so both catastrophic error and
redshift accuracy can be a function of depth. For bright sources \((r_{\text{mag}} < 23)\), the catastrophic error and the redshift accuracy in the AGN sample (312 XAGNs, \(\eta = 22.4\%\) and \(\sigma_{\Delta z/(1+z)} = 0.0847\); 172 IRAGNs, \(\eta = 29.5\%\) and \(\sigma_{\Delta z/(1+z)} = 0.08\); 1024 RAGNs, \(\eta = 7.3\%\) and \(\sigma_{\Delta z/(1+z)} = 0.0474\) and the parent sample (6148 objects, \(\eta = 7.8\%\) and \(\sigma_{\Delta z/(1+z)} = 0.0432\) are smaller. Besides, the photometric redshifts from COSMOS themselves have associated errors (e.g., Salvato et al. 2019), so it is possible to have some redshift differences between the HSC-Wide and COSMOS catalogs. Comparing to the available spectroscopic redshift in HSC-Wide, the catastrophic error and redshift accuracy in the AGN sample (442 XAGNs, \(\eta = 4.8\%\) and \(\sigma_{\Delta z/(1+z)} = 0.0005\); 227 IRAGNs, \(\eta = 7.5\%\) and \(\sigma_{\Delta z/(1+z)} = 0.0005\); 1254 RAGNs, \(\eta = 2.7\%\) and \(\sigma_{\Delta z/(1+z)} = 0.0005\)) is also close to the whole parent sample (7701 objects, \(\eta = 4.1\%\) and \(\sigma_{\Delta z/(1+z)} = 0.0005\)). A possible explanation for the better results with the comparison to available spectroscopic redshift can be that most of the matched sources are bright (60\% of them are \(r_{\text{mag}} < 23\) sources and 97\% of them are \(r_{\text{mag}} < 25\) sources). As mentioned earlier, brighter sources show smaller photometric redshift errors.

3. Analysis

3.1. Evaluations

We evaluate the quality of the classification schemes with their performance. First, we defined true positive (TP; an AGN source that is classified as an AGN), true negative (TN; a non-AGN source that is not classified as an AGN), false positive (FP; a non-AGN source that is classified as an AGN), and false negative (FN; an AGN source that is not classified as an AGN). Therefore, the true-positive rate (TPR), true-negative rate (TNR), false-positive rate (FPR), and false-negative rate (FNR) are 

\[ TPR = \frac{TP}{TP + FN}, \quad TNR = \frac{TN}{TN + FP}, \quad FPR = \frac{FP}{FP + TN}, \quad FNR = \frac{FN}{FN + TP}, \]

respectively. A good classification will classify an AGN as an AGN (high TPR) and a non-AGN as a non-AGN (high TNR), rather than a non-AGN as an AGN (low FPR) and an AGN as a non-AGN (low FNR). Therefore, high performance can be determined by high accuracy, precision, recall, F1 score, and area under the receiver operating characteristic (AUROC), as described below.

1. Accuracy: fraction of sources (AGN and non-AGN) that are classified correctly over all sources,

\[ ACC = \frac{TP + TN}{TP + TN + FP + FN}. \]  

2. Precision: AGN sources that are classified correctly as AGNs over all classified AGNs,

\[ P = \frac{TP}{TP + FP}. \]

3. Recall: AGN sources that are classified correctly as AGNs over all AGN sources,

\[ R = \frac{TP}{TP + FN}. \]

4. F1 score: harmonic mean of the precision and recall,

\[ F1 = 2 \times \frac{P \times R}{P + R}. \]

5. AUROC: shown in the receiver operating characteristic (ROC) curve (TP versus FP) of Figure 4.

3.2. Techniques and Parameters

We use several algorithms from Python packages scikit-learn,\(^8\) Keras,\(^9\) and XGBoost\(^10\) to identify XAGNs, IRAGNs, RAGNs, and non-AGN galaxies (GALs).

First, we use the logistic regression and random forest classifier algorithms in scikit-learn. The logistic regression provides basic logistic functions to model a binary dependent variable. The performance (e.g., accuracy, precision, recall, F1 score, and AUROC) has no significant changes after about 100 iterations. We choose 1000 as the maximum number of iterations (max_iter = 1000) in Table 1.

The random forest classifier in scikit-learn provides an estimator that fits a number of decision tree classifiers on various subsamples of the data and uses averaging to improve the performance and control overfitting. The performance has no significant changes after about 100 trees. We choose 1000 as the number of trees in the forest (n_estimators = 1000) in Table 1.

We use the sequential model in Keras, which is a deep-learning application programming interface written in Python, running on top of the ML platform TensorFlow. A sequential model can stack pain layers, where each layer has

---

\(^8\) [http://scikit-learn.org/](http://scikit-learn.org/)

\(^9\) [https://keras.io/](https://keras.io/)

\(^10\) [https://xgboost.readthedocs.io/](https://xgboost.readthedocs.io/)
Table 1
Numbers of TNR, FPR, FNR, TPR, Accuracy, Precision, Recall, F1 Score, and AUROC of Different Algorithms (Logistic Regression, Random Forest, Keras, and XGBoost) for XAGNs, IRAGNs, and RAGNs (All Available Objects; N = 32, 204)

| Type   | TNR   | FPR   | FNR   | TPR   | ACC   | P     | R     | F1    | AUROC |
|--------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Logistic Regression: HSC+W12 |
| XAGN   | 0.76±0.03 | 0.24±0.03 | 0.39±0.05 | 0.61±0.05 | 0.78±0.03 | 0.52±0.00 | 0.68±0.02 | 0.48±0.01 | 0.68±0.02 |
| IRAGN  | 0.55±0.07 | 0.45±0.07 | 0.33±0.08 | 0.67±0.08 | 0.54±0.05 | 0.51±0.00 | 0.62±0.02 | 0.39±0.02 | 0.62±0.02 |
| RAGN   | 0.79±0.02 | 0.21±0.02 | 0.39±0.03 | 0.61±0.03 | 0.78±0.01 | 0.58±0.00 | 0.70±0.01 | 0.58±0.01 | 0.70±0.01 |

Random Forest: HSC+W12

| XAGN   | 1.00±0.00 | 0.00±0.00 | 0.89±0.04 | 0.11±0.04 | 0.98±0.00 | 0.93±0.01 | 0.56±0.02 | 0.60±0.02 | 0.56±0.02 |
| IRAGN  | 1.00±0.00 | 0.00±0.00 | 0.90±0.05 | 0.10±0.05 | 0.99±0.00 | 0.83±0.01 | 0.55±0.02 | 0.84±0.02 | 0.55±0.02 |
| RAGN   | 0.58±0.01 | 0.00±0.00 | 0.90±0.02 | 0.10±0.02 | 0.92±0.00 | 0.68±0.01 | 0.54±0.01 | 0.55±0.01 | 0.55±0.01 |

Keras: HSC-W12

| XAGN   | 0.98±0.03 | 0.02±0.03 | 0.86±0.09 | 0.14±0.09 | 0.94±0.02 | 0.53±0.06 | 0.59±0.04 | 0.54±0.02 | 0.78±0.02 |
| IRAGN  | 0.99±0.01 | 0.01±0.01 | 0.86±0.04 | 0.14±0.04 | 0.98±0.01 | 0.57±0.03 | 0.57±0.02 | 0.57±0.02 | 0.65±0.03 |
| RAGN   | 0.86±0.03 | 0.14±0.03 | 0.53±0.08 | 0.47±0.08 | 0.84±0.03 | 0.58±0.01 | 0.60±0.03 | 0.60±0.01 | 0.80±0.01 |

XGBoost: HSC+W12

| XAGN   | 0.98±0.03 | 0.02±0.03 | 0.75±0.05 | 0.25±0.05 | 0.97±0.05 | 0.60±0.08 | 0.61±0.02 | 0.60±0.08 | 0.80±0.02 |
| IRAGN  | 1.00±0.01 | 0.01±0.02 | 0.78±0.05 | 0.22±0.05 | 0.99±0.02 | 0.64±0.08 | 0.62±0.03 | 0.62±0.06 | 0.72±0.03 |
| RAGN   | 0.84±0.03 | 0.16±0.03 | 0.40±0.04 | 0.60±0.04 | 0.88±0.01 | 0.61±0.03 | 0.63±0.01 | 0.63±0.02 | 0.83±0.01 |

Note. The uncertainties are derived by bootstrapping.

For the above algorithms, we split our data into training (67%) and test (33%) samples. We test how the algorithms reach their best performance by choosing the input parameters as mentioned above. Because AGNs only represent a small fraction (<10%) of the whole sample, we randomly chose the same number of selected galaxies for the selected AGNs in the training sample. In other words, we used the oversampling technique to deal with the imbalance problem.

3.3. Performances

In Table 1, we compare TNR, FPR, FNR, TPR, accuracy, precision, recall, F1 score, and AUROC for logistic regression, random forest, sequential models in Keras, and XGBoost algorithms. To investigate the errors, we bootstrap the sample 100 times for each algorithm and estimate their uncertainties. To achieve high performance, we expect to have high TNR, low FPR, low FNR, high TPR, high accuracy, high precision, high recall, high F1 score, and high AUROC. In general, the random forest, the sequential model in Keras, and the XGBoost algorithms provide higher TNR and lower FPR than the logistic regression algorithm, which is kind of a benchmark in binary classification. The random forest can reach the highest precision, but XGBoost provides the lower FNR, higher TPR, higher recall, higher F1, and highest AUROC and still has high TNR, low FPR, and high precision. Considering the purpose is to select enough AGN sources correctly, we chose XGBoost to identify AGNs and discuss the accuracy, precision, recall, F1 score, and AUROC for different samples in the following.

We perform grid searching to test the performance of the input parameters with XGBoost as shown in Figure 5. We find that performance does not improve significantly beyond 100 iterations, so it is reasonable that our training stops before 300 rounds of iteration, as mentioned in the previous subsection. In Figure 5, the brighter sample shows a better AUROC and F1 score for XAGNs and IRAGNs in XGBoost. By considering one input tensor and one output tensor. By considering the computing speed and performance, we use early stopping (patience = 5; stop with lack of improvement after five epochs) with the callback function and 1000 epochs of model fitting. As a result, all of the training stops before 100 epochs of model fitting in Table 1. We use flatten, dense, and dropout layers for this work. The setting of 2 in the last dense layer sets the whole sequential model as a binary classifier.

```python
model = Sequential()
model.add(Convolution1D(32, 3, border_mode='same', input_shape = (inp, 1)))
model.add(Convolution1D(32, 3, border_mode='same'))
model.add(Flatten())
model.add(Dense(128, activation = 'relu'))
model.add(Dropout(0.5))
model.add(Dense(2, activation = 'softmax'))
```

We also use XGBoost, which is an ML algorithm to optimize the distributed gradient boosting library and provides a parallel tree boosting in a fast and accurate way. We use early stopping (early_stopping_rounds = 5; stop with lack of improvement after five rounds) and 1000 rounds of iteration. As a result, all of the training stops before 300 rounds of iteration in Table 1. We choose the following parameters by considering the computing speed and performance with the grid-searching technique.

```python
param = {
    'max_depth': 5,
    'eta': 0.3,
    'objective': 'multi:softprob',
    'num_class': 2}
```
The uncertainties are derived by bootstrapping.

Both the sample size and the performance, we decided to use $r_{\text{mag}} < 23$ (cModel photometry is adopted in this paper) as a cut for our subsample of 8193 bright galaxies.

We show the performances for our parent sample (32,204 objects without a missing value of photometric redshift, HSC $grizy$, W1, and W2) in Table 2. We also show the performances of a bright sample (8193 objects with $r_{\text{mag}} < 23$) in Table 3. We list several choices of features: HSC only ($grizy$+photometric redshift), HSC+W12 ($grizy$+W1+W2+photometric redshift), HSC+WISE ($grizy$+W1+2W2+W3+W4+photometric redshift), and WISE only (W1+W2+W3+W4+photometric redshift). In general, the performances of the feature choices can be ranked by HSC+W12 $>$ HSC+WISE $>$ HSC only $>$ WISE only. It suggests that the additional infrared photometry (both HSC+W12 and HSC+WISE) is helpful for the classification. However, the performances cannot be improved further if we include all WISE bands, perhaps because of the missing data in the W3 and W4 bands, as well as the complexity of including an upper limit in the learning process. Therefore, we adopted HSC+W12 for further tests in this paper. Moreover, pure optical information (HSC) can still identify AGNs, but this is not the case for using infrared information (WISE) only.

Comparing Tables 2 and 3, the trained machine performs better for the sample of bright objects, especially for the XAGN and IRAGN sample. Therefore, we focus on this sample to discuss the identification of different AGN types. We find that the performance of bright RAGNs cannot be improved, which is consistent with the grid-searching results in Figure 5.

In Table 4, we separate the XAGN sample into 130 broad-line (type 1) and 241 non-broad-line (type 2) by their spectral type according to Civano et al. (2016) and Marchesi et al. (2016). We also test a classification of the XAGN sample according to their best-fitting SED templates and find similar results. We also separate the IRAGN sample into 110 unobscured (type 1) and 277 obscured (type 2) by their best-fitting SED templates according to Chang et al. (2017a). As a result, we find that both XAGN (type 1) and IRAGN (type 1)
Numbers for Accuracy, Precision, Recall, F1 Score, and AUROC for XAGNs, IRAGNs, and RAGNs ($r_{\text{mag}} < 23; N = 8, 193$)

| Type   | ACC   | P     | R     | F1    | AUROC  |
|--------|-------|-------|-------|-------|--------|
| XAGN   | 0.94 ± 0.04 | 0.64 ± 0.10 | 0.62 ± 0.03 | 0.63 ± 0.08 | 0.78 ± 0.03 |
| IRAGN  | 0.98 ± 0.01 | 0.64 ± 0.07 | 0.66 ± 0.04 | 0.65 ± 0.05 | 0.74 ± 0.04 |
| RAGN   | 0.54 ± 0.10 | 0.59 ± 0.06 | 0.64 ± 0.03 | 0.48 ± 0.08 | 0.70 ± 0.04 |

Table 3

The uncertainties are derived by bootstrapping.

Numbers for Accuracy, Precision, Recall, F1 Score, and AUROC for XAGNs, IRAGNs, and RAGNs (all+type1+type2), IRAGNs (all+type1+type2), and RAGNs (all) for All Available Objects in the 32,204 Parent Sample

| Type       | ACC   | P     | R     | F1    | AUROC  |
|------------|-------|-------|-------|-------|--------|
| XAGN(type 1)| 0.99 ± 0.00 | 0.78 ± 0.04 | 0.75 ± 0.05 | 0.76 ± 0.04 | 0.95 ± 0.02 |
| XAGN(type 2)| 0.99 ± 0.00 | 0.54 ± 0.04 | 0.52 ± 0.03 | 0.53 ± 0.03 | 0.78 ± 0.02 |
| IRAGN      | 0.99 ± 0.02 | 0.70 ± 0.08 | 0.60 ± 0.03 | 0.63 ± 0.06 | 0.73 ± 0.03 |
| IRAGN(type 1)| 0.99 ± 0.00 | 0.76 ± 0.04 | 0.70 ± 0.05 | 0.73 ± 0.04 | 0.81 ± 0.04 |
| IRAGN(type 2)| 0.99 ± 0.00 | 0.58 ± 0.52 | 0.75 ± 0.03 | 0.53 ± 0.03 | 0.66 ± 0.03 |
| RAGN       | 0.87 ± 0.01 | 0.61 ± 0.03 | 0.68 ± 0.01 | 0.63 ± 0.02 | 0.83 ± 0.01 |

Table 4

The uncertainties are derived by bootstrapping.

Numbers for Accuracy, Precision, Recall, F1 Score, and AUROC for XAGNs+IRAGNs, XAGNs+RAGNs, IRAGNs+RAGNs, and XAGNs+IRAGNs+RAGNs in the 32,204 Parent Sample

| Type       | ACC   | P     | R     | F1    | AUROC  |
|------------|-------|-------|-------|-------|--------|
| X+IRAGN    | 0.98 ± 0.00 | 0.68 ± 0.03 | 0.69 ± 0.03 | 0.68 ± 0.03 | 0.93 ± 0.02 |
| X+RAGN     | 0.98 ± 0.00 | 0.68 ± 0.03 | 0.69 ± 0.03 | 0.68 ± 0.03 | 0.92 ± 0.02 |
| IR+RAGN    | 0.99 ± 0.00 | 0.66 ± 0.04 | 0.64 ± 0.04 | 0.65 ± 0.03 | 0.86 ± 0.03 |
| X+IR+RAGN  | 0.99 ± 0.00 | 0.68 ± 0.05 | 0.63 ± 0.05 | 0.65 ± 0.05 | 0.92 ± 0.03 |

Table 5

The uncertainties are derived by bootstrapping.

have much better performance than XAGN (type 2) and IRAGN (type 2). There are 197 XAGNs and IRAGNs (X+IRGs) in common, 289 XAGNs and RAGNs (X+RGs) in common, 180 IRAGNs and RAGNs (IR+RGs) in common, and 114 AGNs belong to three AGN selection methods (X+IR+RGs). We test the performance for AGNs that belong to more than one sample as shown in Table 5. The performance is slightly better but close to the single AGN selection method.

Finally, we test the performance as a function of redshift by breaking the sample into three redshift bins (0.5 < z < 1.5, 1.5 < z < 2.5, and 2.5 < z < 3.5), as shown in Table 6. The
Performances are very similar at all redshift bins, as well as in the whole sample. It suggests that redshift ranges do not seem to be a dominant factor affecting the performances, and \textit{XGBoost} might be able to distinguish the sample with the redshift and photometry information.

3.4. Feature Importance

We show the feature importance of \textit{XGBoost} in Figure 6 for our bright sample ($r_{\text{mag}} < 23; N = 8193$), which indicates the importance of each input feature. For all of the contributions derived from photometry, photometric redshift is the important feature for XAGNs (the second, second, second, and first important features for the HSC, HSC+W12, HSC+WISE, and WISE samples, respectively), IRAGNs (the third, first, second, and first important features for the HSC, HSC+W12, HSC+WISE, and WISE samples, respectively), and RAGNs (the second, sixth, sixth, and first important features for the HSC, HSC+W12, HSC+WISE, and WISE samples, respectively). The differences between types can be explained by their different selection approaches.

If we only consider HSC photometry and the redshift in the first row (HSC only), the g band is the most important feature for all kinds of AGNs. If we add W1 and W2 photometry, as in the second row (HSC+W12), it is clear to see that they contribute to the identification for XAGNs (the fourth important feature), IRAGNs (the fourth important feature), and RAGNs (the third important feature). If we also add W3 and W4 photometry, as in the third row (HSC+WISE), W4 starts to be one of the important features. For identification with only WISE data (WISE only), the photometry redshift is the dominant feature, and W1, W3, and W4 are the second important features for XAGNs, IRAGNs, and RAGNs, respectively.

In general, optical bands (e.g., g, y, and their uncertainties) and photometric redshift are the most important features; the former might be because of the data quality, and the latter could be explained by the fact that the photometry redshift is already the combined information of the HSC photometry. However, it is difficult to solely focus on one band measurement because all photometry is correlated. If we remove the most important feature (e.g., g-band photometry), other correlated features (e.g., y-, r-, i-, and z-band photometry) would become the most important one, and the performances would not have significant differences. Moreover, if we remove the photometric redshift from the features, the ranking of the feature importance does not change much. To show the results with better performance, we keep the version with photometric redshift information.

We find that WISE photometry can provide important information for the classification. However, the performance would be as good as all WISE bands if we only consider W1/W2. Figure 6 shows that W3/W4 can dominate the features if we include them. This can be explained by the fact that there are still missing values for W3/W4 in our subsample, and the sensitivities of W3/W4 (0.86/5.4 mJy) are much worse than those of W1/W2 (0.068/0.098 mJy).

4. Discussion

4.1. Can Optical to Infrared Data Identify AGN Hosts?

We use the \textit{scikit-learn}, Keras, and \textit{XGBoost} packages to identify AGN hosts from the galaxy sample. We find that \textit{XGBoost} can provide a better performance (e.g., TNR, FPR, FNR, TPR, accuracy, precision, recall, F1 score, and ROC), perhaps because of the limit of the algorithms, the choice of our input parameters, and the characteristics of our data. In order to avoid the missing value problem due to nondetected data in the catalog, we consider 32,204 galaxies with available photometric redshifts, as well as detections in HSC grizy bands, W1, and W2. In fact, \textit{XGBoost} can also deal with a missing value and perform well, as shown in our HSC +WISE feature choice. In order to avoid the imbalance problem between AGNs and GALs, we use the oversampling technique. For this kind of data, accuracy (AGN and non-AGN, which are classified correctly over all sources) would be mainly
determined by the non-AGN sample. Therefore, we have to focus on the F1 (harmonic mean of the precision and recall) and AUROC (TP versus FP) values.

In general, optical HSC information with near-infrared information (HSC+W12 or HSC+WISE) shows the best performance (F1 score > 0.60; AUROC > 0.70) among all feature combinations, as shown in Tables 2 and 3. Pure optical (HSC only) or pure infrared (WISE only) may still provide some information, but it seems better to combine them. Both HSC+W12 and HSC+WISE show good performance (F1 score > 0.65; AUROC > 0.75) for our bright sample. This is consistent with the performance of traditional ML and deep-learning algorithms for galaxy morphology classification by Barchi et al. (2020). Moreover, the pure-infrared (WISE-only) performance can be improved for the bright sample, which shows that the optical bands are deeper than the infrared bands.

Besides, photometric redshift can be one of the important features and increase the performance by about 2%. It can be explained that the photometric redshift is derived by the optical photometry, so pure photometry information is enough to provide good performance. According to the results of feature importance, we notice that the input parameters are highly correlated. If we remove the most important feature (e.g., photometric redshift or g band), other correlated features will become the most important feature, and the performances will have no significant changes. Nevertheless, we show that the ML technique can have good performance in identifying AGN hosts with optical to infrared data.

We apply the same ML method to objects in the HSC-Wide layer for HSC-SSP Public Data Release 2 (PDR2; Aihara et al. 2019). Among the 3,238,247 objects, there are 112,609 with available photometric redshift, HSC grizy, W1, and W2 bands.
Table 7
AGN Candidates in the HSC-Wide Region for 112,609 Objects

| Bytes | Format | Units | Label | Explanations |
|-------|--------|-------|-------|--------------|
| 1-17  | I17    | ...   | objid | objid in HSC-SSP PDR2 |
| 19-27 | F9.5   | deg   | RAdeg | R.A. in HSC-SSP PDR2 (J2000) |
| 29-35 | F7.5   | deg   | DECdeg| decl. in HSC-SSP PDR2 (J2000) |
| 37-37 | f      | ...   | XAGN  | 1: XAGNs predicted by ML |
| 39-39 | I      | ...   | IRAGN | 1: IRAGNs predicted by ML |
| 41-41 | I      | ...   | RAGN  | 1: RAGNs predicted by ML |
| 43-49 | F7.5   | ...   | XAGNp | Probability of XAGNs |
| 51-57 | F7.5   | ...   | IRAGNp| Probability of IRAGNs |
| 59-65 | F7.5   | ...   | RAGNp | Probability of RAGNs |

43153636661927346 149.45994 1.73713 0 0 1 0.05203 0.12532 0.19787
43153636661946779 149.45624 1.71012 0 0 0 0.02193 0.06230 0.41638
43153636661947106 149.49619 1.71407 0 0 1 0.20267 0.12076 0.59937
43153636661948529 149.42880 1.73570 0 0 0 0.05203 0.12532 0.19787

Note. The input parameters are photometric redshift, ugriz, W1, and W2 bands. The training data are the 32,204 parent sample. We apply XGBoost as described in Section 3. The objects are predicted to be AGNs when the probabilities are larger than 0.5. Table 7 is published in its entirety in machine-readable format. A portion is shown here for guidance regarding its form and content (This table is available in its entirety in machine-readable form.)

The criterion is similar to the 32,204 parent sample as described in Section 2, and the sample size is around 3.5 times larger. As described in Section 3.3 and Table 1, the output catalog can reach a high accuracy and good performance (F1 score > 0.60; AUROC > 0.70) by XGBoost. According to the precision and recall scores in Table 1, the purity and completeness are larger than 60%. Higher purity and completeness can be achieved by subsamples as discussed in Section 4.2, but we keep the complete sample here. According to our prediction, there are 23,157 XAGN, 9541 IRAGN, and 38,876 RAGN candidates among the 112,609 objects. The predicted data provide a sample dozens of times larger than that of the training data, which require X-ray, infrared, or radio observations as described in Section 2. Table 7 shows the predicted catalog for AGN candidates in the entire HSC-Wide region. The catalog provides the probability (XAGNp, IRAGNp, and RAGNp) of being an AGN with each selection method, as well as the AGN flag (XAGN, IRAGN, and RAGN) when the probability is larger than 0.5.

### 4.2. Can We Classify Different Kinds of AGNs?

The performance is high (F1 score > 0.65; AUROC > 0.80) for bright XAGN and IRAGN host galaxies, as shown in Figure 5 and Table 3. However, the performance of bright RAGN host galaxies cannot be improved. As shown in Figure 5, a sample with bright galaxies shows a slightly worse AUROC and only slightly better F1 score for RAGN host galaxies. To avoid an imbalance between AGN types, we test the data with both over- and undersampling techniques and find that the performance from the XAGN and IRAGN samples is still better than that from the RAGN sample. This might be explained by the sample selection of different kinds of AGNs. For instance, IRAGNs are difficult to select if they are not bright enough in the infrared part, which can be highly correlated with the optical photometry. Because the RAGN sample size is about three to five times larger than the XAGN and IRAGN sample, it may suggest that the intermediate bright AGNs with >3σ radio excess in log(L4.1GHz/SFRIR) require more optical, IR, or radio observations to identify them. Besides, AGNs selected by multiple methods show a slightly better performance than single methods, as shown in Table 5. It might because these are objects with significant AGN features.

For both type 1 (broad-line) XAGNs and type 1 (unobscured) IRAGNs, the performance is very good (F1 score > 0.70; AUROC > 0.80) using the optical to infrared information as shown in Table 4. It can be explained by the point source–like features of type 1 AGNs. We test the structural parameters of type 1 and type 2 AGNs by using GAFLIT (Peng et al. 2010) results for HSC/ACS imaging in Chang et al. (2017a), which used single Sérsic profile fitting with HSC/ACS imaging. About 60% of the AGN sample can be fitted with a single Sérsic profile. We find that the fitting radii of type 1 AGNs are significant smaller than those of type 2 AGNs (the significant probability from the Kolmogorov–Smirnov test is smaller than 5%). And the Sérsic indices of type 1 AGNs are marginally larger than those of type 2 AGNs. This suggests that the point source–like features of type 1 AGNs are distinguishable. Moreover, the structural parameters of non-AGNs are also a significant difference from both type 1 and type 2 AGNs, but a detailed matched mass sample should be investigated. To avoid sample selection constraint by structural measurement, we do not include HSC morphology as input parameters. A detailed study of HSC imaging could also be explored with more sophisticated ML techniques. In this work, we focus on HSC photometry and show that we are able to classify type 1 AGNs with our technique with very high performance. On the other hand, the performance is still fine (F1 score > 0.50; AUROC > 0.60) for type 2 (non-broad-line) XAGNs and type 2 (obscured) IRAGNs. In Table 6, we separate the galaxies into different redshift bins, which show that redshift range is not a dominant factor to identify AGN host galaxies by ML. This is consistent with Pasquet-Itam & Pasquet (2018), who used CNNs to classify and predict the photometric redshift of quasars in SDSS Stripe 82 via light curves. For example, XAGNs show better performance at
1.5 < z < 2.5, but not RAGNs. This might be explained by the identification of the original AGN sample. The sizes of our various subsamples are not very large, especially for bright objects and individual bins. We estimate uncertainties by bootstrapping for the whole learning process, and it seems reasonable for most cases. Nevertheless, a larger AGN sample in future observations would be able to constrain the subsamples better.

The main bands used are in the optical, so we also consider optically selected AGNs. We match our catalog with the optical spectral catalog in the SDSS Data Release 16 (Ahumada et al. 2020). The classifications (CLASS and SUBCLASS) that are labeled as QSO, AGN, BROADLINE, and AGN BROADLINE in SDSS are adopted as optically selected AGNs. The sample size is quite small (38 AGNs out of 231 matched objects), but the performance is very good \((F1 \text{ score } = 0.86 \pm 0.05; \text{ AUROC } = 0.93 \pm 0.06)\). It suggests that our method can also apply to optically selected AGNs, and a larger sample of training data is required to provide general information. Besides, we test AGNs that are labeled by broadband SED fitting in Delvecchio et al. (2017) and Chang et al. (2017a). The performance is also fine \((F1 \text{ score } > 0.60; \text{ AUROC } > 0.65)\). In this work, we focus on XAGNs, IRAGNs, and RAGNs. Nevertheless, the ML technique can also apply to other AGN selection methods in the future.

We do a sanity check by validating data outside the COSMOS field in the HSC-Wide layer: the XXL AGN sample (Menzel et al. 2016), the SWIRE IRAGN sample (Lonsdale et al. 2003), and the RAGN sample in Best & Heckman (2012). Though the techniques of AGN selection are slightly different, the performances \((0.70 > F1 \text{ score } > 0.50; 0.80 > \text{ AUROC } > 0.60)\) show that our algorithms seem to work fine.

In general, ML techniques are able to classify different kinds of AGNs, but the performance will depend on the AGN selections, AGN types, magnitude choices, and sample size. In the future, it would also be helpful for us to investigate other physical properties of AGNs and their host galaxies.

5. Summary

In this paper, we have identified XAGNs, IRAGNs, RAGNs, and GALs with ML techniques. Our main findings are as follows.

1. The HSC (optical) information with WISE band 1 and band 2 (near-infrared) information performs well \((F1 \text{ score } > 0.60; \text{ AUROC } > 0.70)\) to identify AGN hosts.
2. The performance is good \((F1 \text{ score } > 0.65; \text{ AUROC } > 0.75)\) for bright XAGN and IRAGN host galaxies.
3. For both type 1 (broad-line) XAGNs and type 1 (unobscured) IRAGNs, the performance is very good \((F1 \text{ score } > 0.70; \text{ AUROC } > 0.80)\).
4. These results can apply to the five-band data from the wide regions of the HSC survey and future all-sky surveys.

We thank the referee, as well as Chien-Hsiu Lee for helpful comments and discussions. Y.Y.C. acknowledges financial support from Ministry of Science and Technology of Taiwan grants (109-2112-M-005-003-MY3, 108-2112-M-001-014-, and 105-2112-M-001-029-MY3). We gratefully acknowledge the contributions of the entire COSMOS collaboration.

The Hyper Suprime-Cam (HSC) collaboration includes the astronomical communities of Japan and Taiwan and Princeton University. The HSC instrumentation and software were developed by the National Astronomical Observatory of Japan (NAOJ), the Kavli Institute for the Physics and Mathematics of the Universe (Kavli IPMU), the University of Tokyo, the High Energy Accelerator Research Organization (KEK), the Academia Sinica Institute for Astronomy and Astrophysics in Taiwan (ASIAA), and Princeton University. Funding was contributed by the FIRST program from the Japanese Cabinet Office, the Ministry of Education, Culture, Sports, Science and Technology (MEXT), the Japan Society for the Promotion of Science (JSPS), the Japan Science and Technology Agency (JST), the Toray Science Foundation, NAOJ, Kavli IPMU, KEK, ASIAA, and Princeton University. This paper makes use of software developed for the Large Synoptic Survey Telescope. We thank the LSST Project for making their code available as free software at http://lsst.org. This paper is based on data collected at the Subaru Telescope and retrieved from the HSC data archive system, which is operated by the Subaru Telescope and Astronomy Data Center at the National Astronomical Observatory of Japan. Data analysis was in part carried out with the cooperation of the Center for Computational Astrophysics, National Astronomical Observatory of Japan.

This publication makes use of data products from the Wide-field Infrared Survey Explorer, which is a joint project of the University of California, Los Angeles, and the Jet Propulsion Laboratory/California Institute of Technology, and NEOWISE, which is a project of the Jet Propulsion Laboratory/California Institute of Technology. WISE and NEOWISE are funded by the National Aeronautics and Space Administration.

ORCID iDs

Yu-Yen Chang https://orcid.org/0000-0002-6720-8047
Bau-Ching Hsieh https://orcid.org/0000-0001-5615-4904
Wei-Hao Wang https://orcid.org/0000-0003-2588-1265
Yen-Ting Lin https://orcid.org/0000-0001-7146-4687
Chen-Fatt Lim https://orcid.org/0000-0003-1213-9360
Yoshiki Toba https://orcid.org/0000-0002-3531-7863

References

Ahumada, R., Prieto, C. A., Almeida, A., et al. 2020, ApJS, 249, 3
Aihara, H., Arimoto, N., Armstrong, R., et al. 2018, PASJ, 70, S4
Aihara, H., AlSayyad, Y., Ando, M., et al. 2019, PASJ, 71, 114
Alonso-Herrero, A., Pérez-González, P. G., Alexander, D. M., et al. 2006, ApJ, 640, 167
Bai, Y., Liu, J., Wang, S., & Yang, F. 2019, AJ, 157, 9
Baldwin, J. A., Phillips, M. M., & Terlevich, R. 1981, PASP, 93, 5
Ball, N. M., Brunner, R. J., Myers, A. D., & Tcheng, D. 2006, ApJ, 650, 497
Barchi, P. H., de Carvalho, R. R., Rosa, R. R., et al. 2020, A&C, 30, 100334
Best, P. N., & Heckman, T. M. 2012, MNras, 421, 1569
Bonjean, V., Aghanim, N., Salomé, P., et al. 2019, A&A, 622, A137
Bosch, J., Armstrong, R., Bickerton, S., et al. 2018, PASJ, 70, 55
Burlon, D., Ajello, M., Greiner, J., et al. 2011, ApJ, 728, 58
Calderon, V. F., & Benn, C. R., & Jiménez-Luján, F. 2008, MNras, 391, 369
Carballo, R., González-Serrano, J. I., Benn, C. R., & Jiménez-Luján, F. 2008, MNras, 391, 369
Cavuoti, S., Brescia, M., D’Abrusco, R., Longo, G., & Paolillo, M. 2014, MNras, 437, 968
Chabrier, G. 2003, PASP, 115, 763
Chang, Y.-Y., van der Wel, A., da Cunha, E., & Rix, H.-W. 2015, ApJS, 219, 8
Chang, Y.-Y., Le Floc’h, E., Juneau, S., et al. 2017a, ApJS, 233, 19
