A&A, 697, A204 (2025)
https://doi.org/10.1051/0004-6361/202453630

Astronomy & Astrophysics

Selection of optically variable active galactic nuclei via a random forest algorithm⋆

D. De Cicco[1,2,3,⋆⋆], G. Zazzaro[4], S. Cavuoti[3,5], M. Paolillo[1,3,5], G. Longo[1], V. Petrecca[1,3], I. Saccheo[6,7], and P. Sánchez-Sáez[8,2]

[1] Department of Physics, University of Napoli "Federico II", via Cinthia 9, 80126 Napoli, Italy
[2] Millennium Institute of Astrophysics (MAS), Nuncio Monseñor Sotero Sanz 100, Providencia, Santiago, Chile
[3] INAF – Osservatorio Astronomico di Capodimonte, via Moiariello 16, 80131 Napoli, Italy
[4] CIRA – Centro Italiano di Ricerche Aerospaziali, via Maiorise s.n.c., 81043 Capua, Italy
[5] INFN – Sezione di Napoli, via Cinthia 9, 80126 Napoli, Italy
[6] Dipartimento di Matematica e Fisica, Università Roma Tre, Via della Vasca Navale 84, 00146 Roma, Italy
[7] INAF – Osservatorio astronomico di Roma, Via Frascati 33, I-00040 Monte Porzio Catone, Italy
[8] European Southern Observatory, Karl-Schwarzschild-Strasse 2, 85748 Garching bei München, Germany

Received 30 December 2024 / Accepted 21 March 2025

ABSTRACT

Context. A defining characteristic of active galactic nuclei (AGN) that distinguishes them from other astronomical sources is their stochastic variability, which is observable across the entire electromagnetic spectrum. Upcoming optical wide-field surveys, such as the Vera C. Rubin Observatory's Legacy Survey of Space and Time, are set to transform astronomy by delivering unprecedented volumes of data for time domain studies. This data influx will require the development of the expertise and methodologies necessary to manage and analyze it effectively.

Aims. This project focuses on optimizing AGN selection through optical variability in wide-field surveys and aims to reduce the bias against obscured AGN. We tested a random forest (RF) algorithm trained on various feature sets to select AGN. The initial dataset consisted of 54 observations in the r-band and 25 in the g-band of the COSMOS field, captured with the VLT Survey Telescope over a 3.3-year baseline.

Methods. Our analysis relies on feature sets derived separately from either band plus a set of features combining data from both bands, mostly characterizing AGN on the basis of their variability properties and obtained from their light curves. We trained multiple RF classifiers using different subsets of selected features and assessed their performance via targeted metrics.

Results. Our tests provide valuable insights into the use of multiband and multivisit data for AGN identification. We compared our findings with previous studies and dedicated part of the analysis to potential enhancements in selecting obscured AGN. The expertise gained and the methodologies developed here are readily applicable to datasets from other ground- and space-based missions.

Key words. methods: statistical – surveys – galaxies: active

1. Introduction

Active galactic nuclei (AGN) are undoubtedly among the most fascinating sources in the Universe. They are powered by central supermassive black holes (e.g., Burbidge et al. 1963; Salpeter 1964; Rees 1984; Kormendy & Richstone 1995, and references therein) and display diverse characteristics across different wavebands due to both observational perspectives and intrinsic properties. A straightforward consequence of such differences is that no identification technique is capable of returning a complete sample of AGN, as almost a century of investigation has made clear. Nevertheless, AGN do share a key property, this being variability in both their continuum and line emission (though the amplitude and timescales of this variability differ depending, once again, on the waveband). In virtue of this, AGN selection through variability in optical and infrared (IR) bands offers a unique advantage, as it can leverage the multi-visit surveys from ground-based observatories, which have accumulated over the past few decades (e.g., Ulrich et al. 1993; Trevese et al. 1989, 1994, 2008; Klesman & Sarajedini 2007; Sarajedini et al. 2011; De Cicco et al. 2015, 2019; Sánchez-Sáez et al. 2019, 2023). New-generation telescopes – such as the Vera C. Rubin Observatory (Ivezić et al. 2019), expected to begin operations in mid-2025 – promise to revolutionize time-domain astronomy by providing the astronomical community with a groundbreaking data volume that will reshape our understanding of the Universe.

This work is part of a series dedicated to the search for AGN in the COSMOS field based on their optical variability and centered on the analysis of time series from the VLT Survey Telescope (VST). The VST is a 2.6-meter optical telescope located at the Paranal Observatory in Chile and designed for wide-field imaging surveys of the southern sky, with a field of view of 1 sq. deg and a pixel scale of 0.214″ (Capaccioli & Schipani 2011). Specifically, here we test the use of a random forest (RF; Breiman 2001) algorithm to select AGN on the basis of different sets of features quantifying their variability. Over the past decade, the VST time series have been extensively utilized to explore a range of topics, such as variable stars, transient events, and cosmology (e.g., Cappellaro et al. 2015; Falocco et al. 2015; De Cicco et al. 2015, 2019, 2021; Botticella et al. 2017; Fu et al. 2018; Liu et al. 2018, 2020; Poulain et al. 2020), as well as for

⋆ Observations were provided by the ESO programs 088.D-4013, 092.D-0370, and 094.D-0417 (PI G. Pignata).
⋆⋆ Corresponding author: demetra.decicco@unina.it

more technical purposes, including the development of an outlier detection pipeline (Cavuoti et al. 2024). In particular, in the context of AGN selection, De Cicco et al. (2021) is a precursor work in our series dedicated to the VST-COSMOS field, where we initially tested the performance of a model based on an RF algorithm for AGN identification through optical variability, examining how different labeled sets (LSs) and sets of features affect the selection. Most features were selected because they characterize variability and were derived from the source light curves, but we also included six color features and a morphology indicator. That work confirmed the well-known reliability of optical variability as a tool for identifying unobscured AGN, yielding slightly better results than those of De Cicco et al. (2019), where a traditional approach had been adopted for the selection of AGN. Respectively, we obtained 91% versus 86% precision[1]; 69% versus 59% recall[2] for the identification of spectroscopically confirmed AGN; 94% versus 82% recall for the identification of spectroscopically confirmed Type I AGN; 21% versus 18% recall for the identification of spectroscopically confirmed Type II AGN (see Table 5 in De Cicco et al. 2021 for additional information). This last result might appear discouraging, yet it should not be surprising at all considering that the optical emission from Type II AGN is dominated by the host galaxy contribution. Indeed, shorter baselines had yielded even poorer results: when the baseline was limited to five months, De Cicco et al. (2015) found a 6% recall for Type II AGN, in agreement with what is usually found in the literature.

Identifying Type II AGN through optical variability has notoriously posed challenges for decades, classically attributed to their different orientation (e.g., Antonucci 1993; Urry & Padovani 1995; but see also LaMassa et al. 2015; MacLeod et al. 2016; Green et al. 2022, and references therein for a more detailed insight of this class of sources). Indeed, as they are observed roughly edge-on, the dust structure surrounding the accretion disk often prevents one from observing the disk emission, which primarily peaks in the UV/optical waveband.

In essence, this study is meant to expand on the AGN selection approach via RF presented in De Cicco et al. (2021), taking advantage of the quality and cadence of the VST-COSMOS survey but making use of a significantly larger feature set. Indeed, in our previous work, the variability features were derived solely from the *r*-band, which has the highest observing cadence among the VST-COSMOS bands. Here we start with those same *r*-band features but extend the set by adding the corresponding features computed from *g*-band data and by testing the inclusion of bivariate features, that is to say, features that combine data from both the *r* and *g* bands. Such features are expected to improve the selection, as many multiwavelength monitoring campaigns of AGN show correlations between variability properties in adjacent regions of the electromagnetic spectrum (e.g., Edelson et al. 1996; Vanden Berk et al. 2004). We test several RF classifiers that always operate on the same LS but use different feature sets for each test. Because our *r*- and *g*-band data were not always obtained simultaneously, we resort to data imputation in order to obtain, for each source, light curves covering the same baseline and with an equivalent number of simultaneous observations in the two bands used.

This work serves as a forecasting study in view of the highly anticipated Legacy Survey of Space and Time (e.g., LSST Science Collaboration 2009) from the already mentioned Rubin Observatory, and our series of tests are aimed at evaluating whether the use of multiband data enhances AGN selection and at examining the impact of the observing cadence as well as the inclusion of synthetic visits on selection accuracy. We also dedicate part of our analysis to the optimization of the selection of obscured AGN.

The structure of this paper is as follows: Section 2.1 introduces the dataset and the class imbalance problem affecting our LS. Section 3 describes the data imputation process and the set of features used and illustrates how we train our classifiers. Section 4 presents the various classifiers tested, with a focus on the selection of obscured AGN. Section 5 summarizes our main findings.

We note that throughout this work we tend to prefer the definitions "unobscured" and "obscured" – to stress the relative dominance of the nucleus on the host galaxy – rather than the corresponding terms "Type I" and "Type II". Indeed, while there is a correspondence between these two pairs of labels for AGN, we consider the distinction in types too strict and static, especially in light of various studies conducted in recent years, which have shown how AGN can "change" their type and exist as something between Type I and Type II (see references above). Nevertheless, in this work, we use catalogs of AGN selected by Marchesi et al. (2016) as Type I and Type II based on their spectral properties.

2. Data understanding

The data understanding phase involved a detailed exploration of the dataset. It emphasized the extraction of domain-relevant insights and the identification of patterns, anomalies, or limitations that may influence downstream analyses or model development.

2.1. The VST-COSMOS dataset

This work is based on a series of 54 *r*-band and 25 *g*-band visits of the COSMOS field captured by the VST, obtained during three observing seasons and spanning the same baseline of 3.3 yr. Each visit originally covered a ≈1 sq. deg area, but we masked ≈17% of it (mostly edges, defected areas, and saturated stars). The *r*-band dataset has been widely used in previous studies dedicated to AGN optical variability (De Cicco et al. 2015, 2019, 2021, 2022), while the *g*-band dataset is used for the first time for this kind of study. During the first observing seasons, corresponding to the first five months of observations, the planned observing cadence for the *r* and *g* band was of ≈3 and ≈10 days, respectively (with several observing constraints that affected both), which explains the differences in the sampling for the two bands. This led us to resort to data imputation, essentially consisting of replacing missing or incomplete data within a dataset to minimize the biases and errors arising from the lack of original data, as detailed in Sect. 3.1. This allowed us to use bivariate features at the expense of some arbitrariness in the imputation method.

The baseline of this dataset is currently being extended to more than 11 yr with two additional observing seasons, and it will be even longer, as more observations are ongoing. The visit depth is $r \lesssim 24.6$ and $g \lesssim 24.2$ mag for point sources

[1] Precision is defined as the ratio of the sources correctly classified as AGN (true positives) to all the sources classified as AGN regardless of whether the classification is correct or not (true positives+false positives). Hence, it gives the fraction of AGN correctly classified.

[2] Recall is defined as the ratio of the sources correctly classified as AGN (true positives) to all the known AGN in the LS regardless of whether they have been correctly classified or not (true positives + false negatives). Hence, it reveals how often known AGN are correctly classified.

Table 1. VST-COSMOS dataset for the r and g bands.

ID	obs. date	time (days)	r-band	g-band
1	2011-Dec.-18	0	y	n
2	2011-Dec.-22	4	y	n
3	2011-Dec.-27	9	y	y
4	2011-Dec.-31	13	y	n
5	2012-Jan.-02	15	y	n
6	2012-Jan.-06	19	y	n
7	2012-Jan.-18	31	y	n
8	2012-Jan.-20	33	y	n
9	2012-Jan.-22	35	y	y
10	2012-Jan.-24	37	y	n*
11	2012-Jan.-27	40	y	n*
12	2012-Jan.-29	42	y	n*
13	2012-Feb.-02	46	y	y
14	2012-Feb.-16	60	y	y
15	2012-Feb.-19	63	y	n*
16	2012-Feb.-21	65	y	n*
17	2012-Feb.-23	67	y	n*
18	2012-Feb.-26	70	y	y
19	2012-Feb.-29	73	y	n
20	2012-Mar.-03	76	y	n
21	2012-Mar.-13	86	y	n
22	2012-Mar.-15	88	y	n
23	2012-Mar.-17	90	y	y
24	2012-May-08	142	y	n
25	2012-May-09	143	n*	y
26	2012-May-11	145	y	n
27	2012-May-17	151	y	n
28	2013-Dec.-15	728	n	y
29	2013-Dec.-27	740	y	n
30	2013-Dec.-30	743	y	n
31	2014-Jan.-03	747	y	y
32	2014-Jan.-05	749	y	n
33	2014-Jan.-12	756	y	n
34	2014-Jan.-21	765	y	y
35	2014-Jan.-24	768	y	n
36	2014-Feb.-09	784	y	n
37	2014-Feb.-19	794	y	y
38	2014-Feb.-21	796	y	n*
39	2014-Feb.-23	798	y	n*
40	2014-Feb.-26	801	y	y
41	2014-Feb.-28	803	y	n*
42	2014-Mar.-04	807	n*	y
43	2014-Mar.-08	811	y	n
44	2014-Mar.-21	824	y	y
45	2014-Mar.-23	826	y	n*
46	2014-Mar.-25	828	y	n*
47	2014-Mar.-29	832	y	y
48	2014-Apr.-04	838	y	y
49	2014-Apr.-07	841	y	n
50	2014-Dec.-03	1081	y	n
51	2014-Dec.-16	1094	n	y
52	2014-Dec.-25	1103	n	y
53	2015-Jan.-03	1112	n	y
54	2015-Jan.-10	1119	y	n*
55	2015-Jan.-15	1124	n	y
56	2015-Jan.-23	1132	n	y
57	2015-Jan.-28	1137	y	n*
58	2015-Jan.-30	1139	n*	y
59	2015-Jan.-31	1140	y	n*
60	2-15-Feb.-14	1154	n	y
61	2015-Feb.-15	1155	y	n
62	2015-Mar.-10	1178	y	n
63	2015-Mar.-13	1181	n*	y
64	2015-Mar.-14	1182	y	n*
65	2015-Mar.-19	1187	y	n*
66	2015-Mar.-22	1190	n	y

Notes. The table reports the visit ID, date, time in days from the first observation, presence or absence (y/n) of a visit in the r and g bands. Different colors are associated to different y/n combinations for the two bands. The asterisk identifies the visits that were originally missing in either band, and that we included in the dataset via data imputation, as detailed in Sect. 3.1.

($\sim 5\sigma$ confidence level). This dataset can therefore be considered as a scaled-down version of what will be obtained from the LSST main survey, which is expected to cover a 10 yr baseline, with single-visit depths of 24.7 and 25.0 mag in the r and g bands, respectively. This is one of the reasons why in this series of works we have been extensively exploiting our dataset for LSST performance forecasting studies.

We report essential information about our dataset in Table 1, while we refer the reader to the above-mentioned papers for further details about the r-band dataset, and in particular we refer to De Cicco et al. (2015) for an overview of the reduction process. Throughout this work, magnitudes are in the AB system and time differences are expressed in the observer reference frame.

2.2. The labeled set and the class imbalance problem

One of the issues one commonly has to face when handling real data is the imbalance in the number of objects of different types. Specifically, for this study we had to deal with an unbalanced LS, consisting of two main classes: 380 AGN and 2163 non-AGN, of which 1168 are stars and 995 are "inactive" galaxies, meaning galaxies where no nuclear activity is detected. The sources that make up our LS were selected from different catalogs from the literature, essentially adopting the same criteria described in De Cicco et al. (2021). In short, stars come from the COSMOS ACS catalog (Koekemoer et al. 2007; Scoville et al. 2007b) and the selection was refined by checking that these sources lie on the stellar locus on a $r-z$ versus $z-K$ diagram[3] (Nakos et al. 2009). Inactive galaxies were selected from the COSMOS2015 catalog (Laigle et al. 2016), which provides a classification based on the best-fit templates from Bruzual & Charlot (2003). We cross-matched our samples of stars and inactive galaxies with COSMOS catalogs available from the literature and excluded sources with conflicting classifications. Additional details can be found in Sect. 2.5 of De Cicco et al. (2021). For what concerns AGN, the LS here used is smaller than the one used in De Cicco et al. (2021), which was based on r-band data only, as here we had to

[3] As mentioned above, Sect. 2.5 of De Cicco et al. (2021) explains in detail how we selected a sample of 1000 stars. Here we simply did not set the size of the star LS to 1000 a priori; furthermore, contrary to what was done in De Cicco et al. (2021), here we included in the star LS six sources with $r-z > 1.5$ mag, as their classification as stars seems to be reliable.

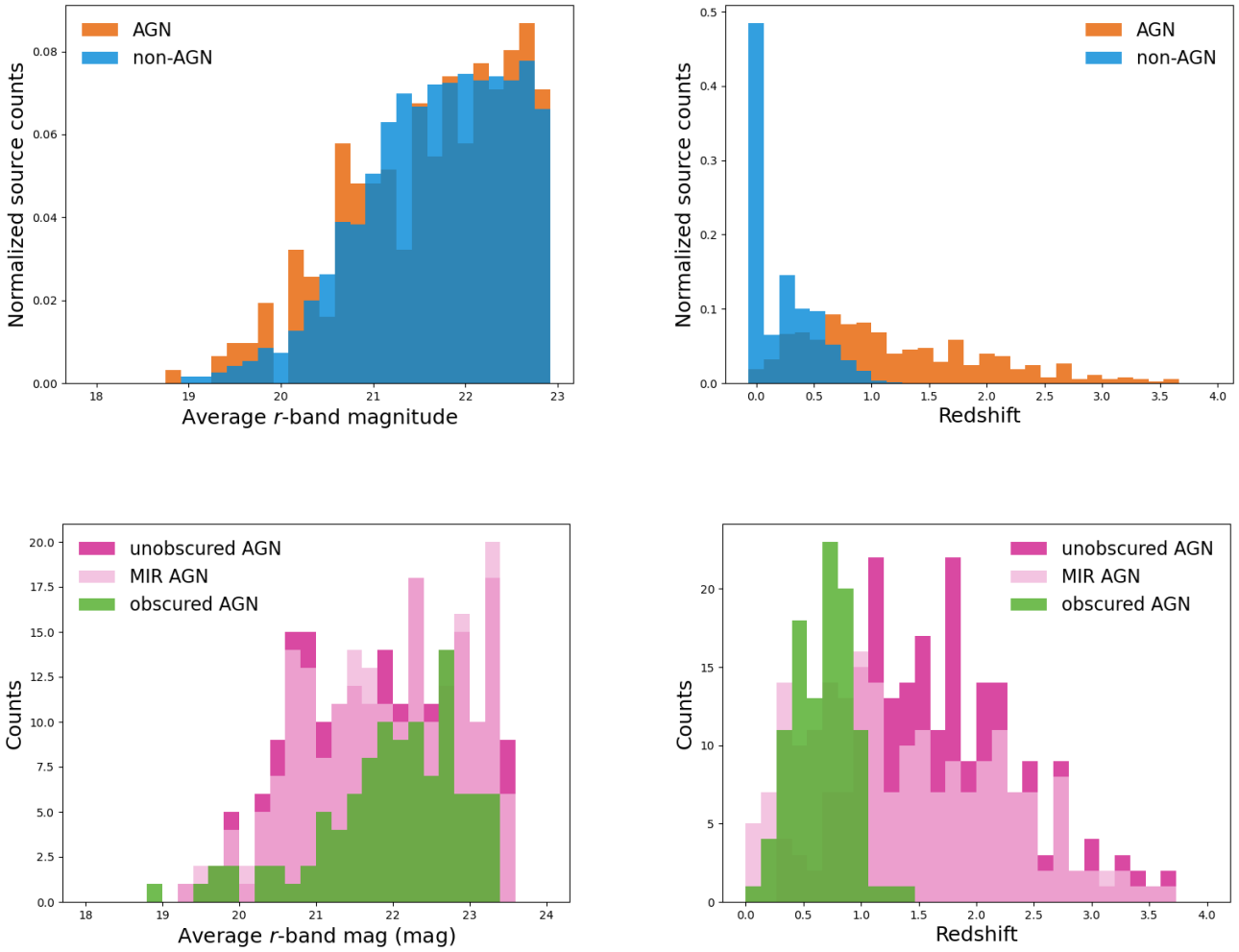

Fig. 1. Average r-band magnitude (*upper-left panel*) and redshift (*upper-right panel*) for the two classes (AGN and non-AGN) in our LS. Each histogram has been normalized to the total number of sources in the corresponding class. Average r-band magnitude (*lower-left panel*) and redshift (*lower-right panel*) for the AGN LS and the three subsamples of unobscured, obscured, and MIR AGN.

take into account the g band as well; as a consequence, here we excluded all the sources that were not detected in at least two visits in the g band, that is, the minimum required to compute the various variability features used. This requirement generally affects the non-AGN LS as well, but in the present work we compensated for the exclusion of some stars by including some others, while the size of the inactive galaxy LS is not significantly reduced by the above-mentioned requirement (only five sources are excluded from the present LS).

As in De Cicco et al. (2021), we identify AGN on the basis of different diagnostics, and classify them as spectroscopic Type I (i.e., unobscured, 217 sources) and Type II (i.e., obscured, 104), and mid-infrared-selected (MIR, 211, 59 of which lack spectroscopic classification). Specifically, the spectroscopic classification for Type I and Type II AGN comes from the *Chandra*-COSMOS Legacy Catalog (Marchesi et al. 2016), which means these are X-ray emitting sources for which an optical counterpart is available and that were classified as AGN via optical spectroscopy, based on the traditional criterion relying on the presence of broad ($\geq$2000 km s^{-1}) emission lines in a spectrum; the MIR classification is based on the criterion by Donley et al. (2012) which, in a diagram comparing the two MIR colors $\log(F[8.0]\mu m/F[4.5]\mu m)$ versus $\log(F[5.8]\mu m/F[3.6]\mu m)$, identifies a region where AGN typically place themselves, the MIR information coming from the already mentioned COSMOS2015 catalog; see also Sect. 4 of De Cicco et al. (2019) for further details. We note that a source can be classified as AGN by more than one criterion. In addition, we note that we also have information about the nature of the AGN in our LS, based on their X-ray emission (Marchesi et al. 2016; Brusa et al. 2010) and on their optical variability; nevertheless, we decided not to use such information in this work, as we are interested in a proper comparison with what we did in De Cicco et al. (2021), where we only consider spectroscopic- and MIR-selected AGN for our LS. We also note that while we always know which source in the non-AGN LS is a star and which one is a galaxy, for our purpose they all are simply considered as non-AGN. Hence, our classification will be binary, AGN being the minority (or positive) class, and non-AGN being the majority (or negative) class.

In Fig. 1 we show the magnitude and redshift distributions for the two classes of AGN and non-AGN in our LS, and also for the three subclasses that form our AGN LS. Consistent with previous works, we cut our LS to an average magnitude $r \leq 23.5$ mag, which is assumed as the completeness limit of our survey. For what concerns redshifts, the median value is 0.411 for the galaxies in the non-AGN class, being the redshift 0 for the stars, and 1.106 for AGN, which extend to higher redshifts, the highest being 3.715. If we focus on unobscured, obscured, and MIR-selected

AGN, we can see that their magnitudes roughly span the same range, while obscured AGN have lower redshifts than the other two subclasses, the median values being 1.657, 0.678, and 1.214 for unobscured, obscured, and MIR AGN, respectively. We stress that, while the first two subclasses are disjointed, the MIR subclass partly overlaps the other two.

The classifier training phase is often deeply conditioned by the majority class: though the models can have high general accuracy, at a closer look they show a low predictive accuracy for the minority class. Models trained with most learning algorithms on an unbalanced dataset frequently predict most records as negative. This is often regarded as a problem in learning from highly imbalanced datasets, especially when the minority class is the one we are interested in.

Imbalance is quantified by the imbalance ratio (e.g., Amin et al. 2016), defined as the ratio of minority class instances to majority ones (also referred to as skew; e.g., Jeni et al. 2013). Based on the above-listed numbers, the imbalance ratio of our LS, corresponding to the ratio between AGN and non-AGN, is $380/2163 = 0.1757$. Intuitively, the greater the imbalance in a dataset, the more complex the learning process. Hence the training of a classifier with satisfactory performance is increasingly challenging.

3. Data preparation

In the present section we illustrate how we dealt with the different sampling cadences of the two bands we used in this work. We also introduce the features here used, focusing on the ones that are new with respect to those used in De Cicco et al. (2021).

3.1. Missing data imputation

Ideally, in order to compute a feature based on data from two bands, we should have "simultaneous" – meaning as close as possible in time – visits in these two bands. Hence, we considered as good candidates for the computation of bivariate features only those pairs of visits that were obtained during the same night. This condition is fulfilled by 13 out of the 66 observing dates in our list; it is also worth noting that in four instances there is a one-night lag between observations in the r and g bands. Although ideally we would like to replicate the LSST observing cadence, where observations in different bands will be obtained in the same night, this shift of one night should not be a major issue for all the non-blazar objects. Based on this, with our dataset we should be able to build bivariate features using only 13 points per light curve in the best case scenario, that is, when a source is detected in each of the 13 corresponding visits. This would mean that most of our dataset would be wasted. Also, since the differences in the sampling of the two bands lead to a different number of visits for each of them, the various features that we would compute for each source would be obtained from a different number of visits. Hence, the weight of the various features in any rankings would be different as it would depend on the number of visits involved in their computation. We therefore resorted to data imputation in order to fill the gaps in the light curves when possible, as detailed in the following. We compared the data in the g and r bands, and considered all the cases where, on a given date, there is a visit in either band, but not in both. Hence we considered the closest visits before and after the given date in the band with the gap, and computed the time difference between the two, in days: if this is $\leq$15 days, we considered this as a good case for data imputation, that is, we assume that, knowing the properties of AGN variability, variations in such a time interval will not be too distant from a linear behavior. We can therefore fill the gap in the dataset via linear interpolation. We did this for both bands, filling four gaps in the r band, that is, the one with the denser sampling, and 16 gaps in the g band. We did not perform imputation when the time difference is >15 days. After this procedure we ended up with 33 visits per band. The error bars we associated to both real and synthetic magnitude values were computed from the whole sample of VST-COSMOS sources: we considered the average magnitude value for each source, and defined the error bar as the 95% uncertainty on that magnitude value.

3.2. Features used for AGN selection

The identification of AGN in this work is based on the use of a number of features. Most of them have been frequently used in the literature for variability studies, as they are suitable for variability analysis and can be computed from the source light curves; these features, which hereafter we refer to as "univariate", were computed independently for the r and the g band from the corresponding light curves. Specifically, we used the same univariate features used in De Cicco et al. (2021). Following that work, we also used the same set of color features as well as the only morphology feature there used; of course these were computed just once, being independent on our g- and r-band light curves. Details about these features can be found in Sects. 2.2, 2.3, and 2.4 of De Cicco et al. (2021), but we report here Table 1 from that work – Table 2 in this work – for the sake of convenience. In addition to the above-mentioned features, in this work we included a set of features that were computed combining the g and r bands together and that we therefore labeled "bivariate". For what concerns the bivariate features we basically resorted to a series of similarity measures, quantifying the distance between the two light curves in each pair. Some of these measures, such as the L_1-norm or L_2-norm, are very well-known and of immediate interpretation, while some others are more complex. In general, these features are grouped under different families (e.g., Tschopp & Hernandez-Rivera 2017), namely:

- The L_p Minkowski family, containing a series of measures corresponding to the generalized formula $\sqrt[p]{\sum_i |X_i - Y_i|^p}$ as the index p changes;
- The L_1 family, containing measures related to the absolute difference $\sum_i |X_i - Y_i|$ introduced in the L_1 distance;
- The intersection family, which contains measures related to the intersection of the g and r sets of points;
- The inner product family, where measures are defined on the basis of the inner product between the two sets of points in the two bands;
- The fidelity family, where measures are defined starting from the so-called Fidelity similarity, that is, the sum of the square root of the inner product;
- The χ^2 family, whose member features originate from the square of the Euclidean norm L_2;
- The Shannon entropy family, named after the Shannon entropy, which features in this group are based on;
- The combination family, where characteristics from different families are combined;
- The vicissitude family, which contains a number of features defined in Cha (2007).

The complete list of the bivariate features used in this work is reported in Table 3.

In total, we used a set of: 2×29 univariate features + one morphological feature + 6 color features + 39 bivariate features + 29 features defined as the differences between

Table 2. List of univariate variability features, morphology feature, and color features used in this work.

	Feature	Description	Reference
classic var. features	A_{SF}	rms magnitude difference of the SF, computed over a 1 yr timescale	Schmidt et al. (2010)
	γ_{SF}	Logarithmic gradient of the mean change in magnitude	Schmidt et al. (2010)
	GP_DRW_τ	Relaxation time τ (i.e., time necessary for the time series to become uncorrelated), from a DRW model for the light curve	Graham et al. (2017)
	GP_DRW_σ	Variability of the time series at short timescales ($t \ll \tau$), from a DRW model for the light curve	Graham et al. (2017)
	ExcessVar	Measure of the intrinsic variability amplitude	Allevato et al. (2013)
	P_{var}	Probability that the source is intrinsically variable	McLaughlin et al. (1996)
	IAR_{ϕ}	Level of autocorrelation using a discrete-time representation of a DRW model	Eyheramendy et al. (2018)
FATS features	Amplitude	Half of the difference between the median of the maximum 5% and of the minimum 5% magnitudes	Richards et al. (2011)
	AndersonDarling	Test of whether a sample of data comes from a population with a specific distribution	Nun et al. (2015)
	Autocor_length	Lag value where the auto-correlation function becomes smaller than η^e	Kim et al. (2011)
	Beyond1Std	Percentage of points with photometric mag that lie beyond 1σ from the mean	Richards et al. (2011)
	η^e	Ratio of the mean of the squares of successive mag differences to the variance of the light curve	Kim et al. (2014)
	Gskew	Median-based measure of the skew	–
	LinearTrend	Slope of a linear fit to the light curve	Richards et al. (2011)
	MaxSlope	Maximum absolute magnitude slope between two consecutive observations	Richards et al. (2011)
	Meanvariance	Ratio of the standard deviation to the mean magnitude	Nun et al. (2015)
	MedianAbsDev	Median discrepancy of the data from the median data	Richards et al. (2011)
	MedianBRP	Fraction of photometric points within amplitude/10 of the median mag	Richards et al. (2011)
	MHAOV_Period	Periodo obtained via the Multi-Harmonic Analysis of Variability periodogram	Huijse et al. (2018)
	PairSlopeTrend	Fraction of increasing first differences minus fraction of decreasing first differences over the last 30 time-sorted mag measures	Richards et al. (2011)
	PercentAmplitude	Largest percentage difference between either max or min mag and median mag	Richards et al. (2011)
	Q31	Difference between the 3rd and the 1st quartile of the light curve	Kim et al. (2014)
	Period_fit	False-alarm probability of the largest periodogram value obtained with LS	Kim et al. (2011)
	Ψ_{CS}	Range of a cumulative sum applied to the phase-folded light curve	Kim et al. (2011)
	Ψ_{η}	η^e index calculated from the folded light curve	Kim et al. (2014)
	R_{cs}	Range of a cumulative sum	Kim et al. (2011)
	Skew	Skewness measure	Richards et al. (2011)
	Std	Standard deviation of the light curve	Nun et al. (2015)
	StetsonK	Robust kurtosis measure	Kim et al. (2011)
morph.	class_star	HST stellarity index	Koekemoer et al. (2007), Scoville et al. (2007a)
colors	u-B	CFHT u magnitude – Subaru B magnitude	Laigle et al. (2016)
	B-r	Subaru SuprimeCam B mag – Subaru SuprimeCam $r+$ mag	Laigle et al. (2016)
	r-i	Subaru SuprimeCam $r+$ mag – Subaru SuprimeCam $i+$ mag	Laigle et al. (2016)
	i-z	Subaru SuprimeCam $i+$ mag – Subaru SuprimeCam $z++$ mag	Laigle et al. (2016)
	z-y	Subaru SuprimeCam $z++$ mag – Subaru Hyper-SuprimeCam y mag	Laigle et al. (2016)
	ch21	*Spitzer* 4.5 µm (*channel2*) mag – 3.6 µm (*channel1*) mag	Laigle et al. (2016)

Notes. The first two sections of the table report variability features; class_star is the only morphology feature used; the bottom section of the table lists the color features used, where ch21 is the only MIR color used, while the others are optical/near-infrared (NIR) colors. This table corresponds to Table 1 in De Cicco et al. (2021).

homologous univariate g- and r-band features + 29 features obtained from the $g - r$ light curves, for a total of 162 features.

3.3. Training with a heterogeneous labeled set

It is common practice to split the LS into two disjoint subsets: a training set – usually the 70–75% of the LS, used for model fitting – and a validation set – usually 30–25% of the LS –, dedicated to the fine-tuning of the hyperparameters and to assessing the performance of the model during the learning process. Nevertheless, considering the heterogeneous nature of our sample of AGN that were selected on the basis of different properties, such a choice would lead to results strongly dependent on the type of AGN used in the training. De Cicco et al. (2021) already addressed this issue, and thus, consistent with that work, we resorted to the leave-one-out cross-validation (LOOCV; Sammut & Webb 2010). This essentially consists in treating each source in the LS as a single-unit validation set,

Table 3. List of the bivariate features used in this work grouped under the families introduced in Sect. 3.2 (Tschopp & Hernandez-Rivera 2017).

Minkowski family	
City_Block, L_1-norm	$d_{\rm City} = \sum \lvert g_i - r_i \rvert$
Euclidean, L_2-norm	$d_{\rm Eucl} = \sqrt{\sum (g_i - r_i)^2}$
Chebyshev, L_∞-norm	$d_{\rm Cv} = \max_i \lvert g_i - r_i \rvert$
L_1 family	
Sørensen	$d_{\rm S\o} = \frac{\sum \lvert g_i - r_i \rvert}{\sum (g_i + r_i)}$
Gower	$d_{\rm Gw} = \sum \lvert g_i - r_i \rvert / b$
Kulczynski	$d_{\rm Kul} = \frac{\sum \lvert g_i - r_i \rvert}{\sum \min(g_i, r_i)}$
Canberra	$d_{\rm Canb} = \sum \frac{\lvert g_i - r_i \rvert}{(g_i + r_i)}$
Lorentzian	$d_{\rm Lor} = \sum \ln(1 + \lvert g_i - r_i \rvert)$
Intersection family	
Intersection	$d_{\rm Is} = \sum \lvert g_i - r_i \rvert / 2$
Wave_Hedges	$d_{\rm WH} = \sum \frac{\lvert g_i - r_i \rvert}{\max(g_i, r_i)}$
Motyka	$s_{\rm Mo} = \frac{\sum \max(g_i, r_i)}{\sum (g_i + r_i)}$
Czekanowski	$d_{\rm Cz} = \frac{\sum \lvert g_i - r_i \rvert}{\sum (g_i + r_i)}$
Ruzicka	$s_{\rm Mo} = \frac{\sum \min(g_i, r_i)}{\sum \max(g_i, r_i)}$
Inner product family	
Inner_Product	$s_{\rm Ip} = \sum g_i r_i$
Harmonic_Mean	$s_{\rm Hm} = 2 \frac{\sum g_i r_i}{(g_i + r_i)}$
Cosine	$s_{\rm Cos} = \frac{\sum g_i r_i}{\sqrt{\sum g_i^2 \sum r_i^2}}$
Jaccard	$d_{\rm Ja} = \frac{\sum (g_i - r_i)^2}{\sum (g_i^2 + r_i^2 - g_i r_i)}$
Dice	$d_{\rm Di} = \frac{\sum (g_i - r_i)^2}{\sum (g_i^2 + r_i^2)}$
Fidelity family	
Fidelity	$s_{\rm Fid} = \sum \sqrt{g_i r_i}$
Bhattacharyya	$d_{\rm Ba} = -\ln \sum \sqrt{g_i r_i}$
Squared-chord	$d_{\rm SC} = \sum (\sqrt{g_i} - \sqrt{r_i})^2$

χ^2 family	
Squared_Euclidean	$d_{\rm SE} = \sum (g_i - r_i)^2$
Pearson χ^2	$d_{\rm Pea} = \sum (g_i - r_i)^2 / r_i$
Neyman χ^2	$d_{\rm Ney} = \sum (g_i - r_i)^2 / g_i$
χ^2	$d_{\rm Sq\chi} = \sum \frac{(g_i - r_i)^2}{g_i + r_i}$
Divergence	$d_{\rm Div} = 2 \sum \frac{(g_i - r_i)^2}{(g_i + r_i)^2}$
Clark	$d_{\rm Cl} = \sqrt{\sum \left[\frac{\lvert g_i - r_i \rvert}{(g_i + r_i)} \right]^2}$
Additive_Symmetric χ^2	$d_{\rm Ad\chi} = \sum \frac{(g_i - r_i)^2 (g_i + r_i)}{g_i r_i}$
Shannon's entropy family	
Kullback-Leibler	$d_{\rm KL} = \sum g_i \ln(g_i / r_i)$
Jeffreys	$d_{\rm Jef} = \sum (g_i - r_i) \ln(g_i / r_i)$
K-divergence	$d_{\rm Kdv} = \sum g_i \ln(2 g_i / (g_i + r_i))$
Topsøe	$d_{\rm Top} = \sum (g_i \ln \frac{2 g_i}{g_i + r_i} + r_i \ln \frac{2 r_i}{g_i + r_i})$ $-(g_i + r_i)/2 * ln((g_i + r_i)/2))$
Combination family	
Taneja	$d_{\rm Tan} = \sum (g_i + r_i)/2 \ln(\frac{g_i + r_i}{2\sqrt{g_i r_i}})$
Kumar-Johnson	$d_{\rm KJ} = \sum \frac{(g_i^2 - r_i^2)^2}{2 (g_i r_i)^{3/2}}$
Average(L_1-L_∞)	$d_{\rm avL} = \sum (\lvert g_i - r_i \rvert + \max_i \lvert g_i - r_i \rvert)/2$
Vicissitude family	
Vicis-Wave_Hedges	$d_{\rm V_{wh}} = \sum \lvert g_i - r_i \rvert / \min(g_i, r_i)$
Vicis-Symmetric χ^2_3	$d_{\rm vs\chi^2_3} = \sum (g_i - r_i)^2 / \max(g_i, r_i)$
Max-Symmetric χ^2	$d_{\rm MaxS} = \max(\sum (g_i - r_i)^2 / g_i, \sum (g_i - r_i)^2 / r_i)$
Min-Symmetric χ^2	$d_{\rm MinS} = \min(\sum (g_i - r_i)^2 / g_i, \sum (g_i - r_i)^2 / r_i)$

Notes. The functions used to compute distance or similarity measures require vectors X and Y (in this case, the g- and r-band light curves) and return the corresponding similarities or distances per the various measures given above.

while the remaining sources in the LS constitute the training set and are therefore used to classify the excluded source. Once this is done for each of the sources in the LS, a prediction is available for each of them. Hence, the LS serves as both the training and as the validation set, as each single-unit is used in the validation phase when it is not included in the training set.

4. Random forest-based tests using features in two bands

As we mentioned in Sect. 2.1, what is new in this work with respect to De Cicco et al. (2021) is the use of g-band observations in addition to and in combination with the r-band set of data. As a consequence, the first natural step is to compare the results obtained in De Cicco et al. (2021) with the ones here obtained. We stress once again that De Cicco et al. (2021) adopted for the validation the same approach used here, but in the present work we expanded the set of features compared to the one used in that work, consisting of the variability features computed from the r-band light curve, the morphology indicator, the five optical/NIR colors, and the MIR color, as detailed in Table 2. We also point out that the g- and r-band light curves used in this work have the same number of points per source due to the data imputation procedure described in Sect. 3.1, the maximum number of points being 33.

In order to compare the two works and make further tests, here we analyzed the performance of a model trained via an RF algorithm. The code here used is based on the use of the Python *scikit-learn* library. We took the imbalance in our classes into account by setting the parameter `class_weight = balanced_subsample` in the algorithm so that, for each bootstrap sample used to extract a subset of features to build a tree, the weights of each class were adjusted dynamically based on the class distribution in that bootstrap sample used to train that specific tree.

In order to optimize the performance of our RF classifiers, we tested several possible combinations of the hyperparameters

that typically affect the most the building process of the ensemble of decision trees and the way predictions are made. Specifically, we resorted to a grid search, which essentially requires as an input a grid of values for the various hyperparameters one aims at tuning, and then performs an exhaustive search over this grid to find their best combination via cross-validation and based on specific scoring metrics. Of course the possible combinations to test are infinite and the process is computationally expensive. Hence, one has to limit the number of possible values for each hyperparameter. Here we chose to tune the following hyperparameters:

- *n_estimators*: This defines the number of decision trees to be used to build a forest and, ideally, it should optimize the accuracy of the classification over a reasonable computational time; we tested the values 100, 300, 500;
- *min_samples_split*: This sets a minimum threshold for the number of objects required to split an internal node; we tested the values 2, 5, 10;
- *max_depth*: This defines the depth of each tree, aiming at avoiding underfitting/overfitting; we tested the values 10, 20, 30, None, where the last one means that the tree ramification goes on until all leaves are pure (i.e., they only contain sources from one class) or until the stopping criterion defined by *min_samples_split* is fulfilled;
- *min_samples_leaf*: This sets the minimum number of objects required for a node not to be merged with its parent node, hence preventing an excessive growth of the tree; we tested the values 1, 2, 4;
- *max_features*: This defines the number of features to consider for the splitting in each node; we tested the options *sqrt* and *log2*, respectively setting this number to the square root or the $\log_2$ of the total number of features.

Based on the chosen values, we tested 216 combinations for each classifier. We chose to evaluate the performance of our models via the balanced accuracy, which is essentially an average of recall for the positive and negative classes and thus means that we are taking into account that our LS is unbalanced. We report details about the obtained results in Appendix A.

Our classifiers were built making use of different sets of features for each test, but always including the morphology indicator and all the colors from Table 2.

- *D*21 test: This aimed at repeating what was done in De Cicco et al. (2021). Hence, the RF classifier made use of the same set of features there used; but, for the sake of consistency with the rest of the present work, we used the same LS – made of 2543 instead of 2414 sources – which we introduced in Sect. 2.2; this is the only case where we used the total number of points of the *r*-band light curves, the maximum being 54, and did not limit this number to 33 as in the rest of this work. Again, the choice of using the full set of points for each light curve was in order to be consistent with De Cicco et al. (2021). With this test we aimed at assessing how important the number of visits – and hence of points in a light curve – is in the AGN selection process.
- *r*-band test: Here the RF classifier made use of the same set of features used in De Cicco et al. (2021), the LS consisting of the above-mentioned 2543 sources. The difference with the previous test in this list is that the light curves of the sources here used consist of up to 33 points, as we aimed at properly comparing the results from this classifier to the ones that we would obtain from *g*-band data; the light curves can include a maximum of four synthetic points, based on what we explained in Sect. 3.1 and showed in Table 1.
- *g*-band test: The RF classifier made use of the same morphology indicator and colors used in the previous two tests, but the *r*-band features were replaced by the corresponding *g*-band features. We stress that, as in the previous test, the light curves of the sources used for this test consist of up to 33 points but, in this case, a maximum of 16 points can be synthetic. This means that, with this test, we were also trying to assess whether the nature (real or synthetic) of the light curve points affects the final classification.
- *rg* test: The RF classifier made use of all the features used in the previous two tests, i.e.: *r*-band features, *g*-band features, plus the morphology indicator and colors introduced in Table 2. With this test we explored the option of using two bands (by using the synthetic points that we added with the imputation) for the selection process, instead of using only one.
- *rg* + bivariate feature test: The RF classifier made use of all the features selected for this work, that is, the ones used for the previous classifier plus the bivariate features reported in Table 3. This test was meant to assess the relevance of features combining simultaneous observations in the two bands used, in addition to features computed from single-band observations.
- $(g-r)_{\rm feat}$ test: The RF classifier made use of features defined as the difference between each *r*-band feature and the homologous *g*-band feature plus, as usual, the morphology indicator and the colors from Table 2. This test investigated the variability of the "colors" of the various features initially defined for each band, thus identifying possible features that vary significantly from one band to another.
- $(g-r)_{\rm mag}$ test: The RF classifier made use of variability features computed from the light curves obtained as the magnitude difference between the *g* and the *r* band for each source plus, as usual, additional features, these being the morphology indicator and the colors from Table 2. This test investigated the variability of the features obtained from "color light curves".

Table 4 reports some metrics that typically characterize the performance of a binary classifier, and that were derived from the confusion matrices obtained from the various classifiers tested in this work. We recall that, for a binary classifier, the confusion matrix consists of four frames, reporting the number of:

- true positives (TPs), that is to say, known AGN correctly classified as AGN;
- true negatives (TNs), that is to say, known non-AGN correctly classified as non-AGN;
- false positives (FPs), that is to say, known non-AGN erroneously classified as AGN;
- false negatives (FNs), that is to say, known AGN erroneously classified as non-AGN.

Consistent with De Cicco et al. (2021), the metrics we used in this work are accuracy (*A*), precision (*P*, also known as purity), recall (*R*, also known as completeness), and *F*1, defined as follows:

$$A = \frac{\mathrm{TPs} + \mathrm{TNs}}{\mathrm{Tot.Sample}},$$

which tells how often the classification is correct for either class and is hence computed with respect to the whole sample of sources in the LS;

$$P = \frac{\mathrm{TPs}}{\mathrm{TPs} + \mathrm{FPs}},$$

which tells how often the classification as AGN is correct and therefore only refers to the sources classified as AGN;

$$R = \frac{\mathrm{TPs}}{\mathrm{TPs} + \mathrm{FNs}},$$

which tells how often known AGN are classified correctly and is hence computed with respect to all the known AGN in the LS;

$$F1 = 2 \times \frac{P \times R}{P + R},$$

which is the harmonic mean of P and R and therefore provides a different estimate of the accuracy, which also takes into account the sources for which the classification is wrong.

Specifically, the upper section of Table 4 refers to the various classifiers introduced above. Since De Cicco et al. (2021) also tested the use of various LSs, including different (sub)samples of AGN, we specify that the percentages reported in this work refer to the case where the full LS of 2543 sources is used. This is indeed the most interesting case if we consider that, in general, we do not know a priori what class of AGN we are dealing with, which depends on the specific properties of the survey and on the sample selection criteria.

The analysis of the results obtained from this first series of tests led to a series of remarks, which we discuss in what follows:

- The recall (quantified by the true positive ratio, TPR), that is, the largest fraction of correctly classified AGN, is generally consistent from test to test, except for the two classifiers combining r and g features; this also holds for the recall of obscured AGN, while we obtained a slightly higher value for the recall of unobscured AGN when we used 54 instead of 33 visits. The mild decrease in the TPR for the two classifiers *rg* and *rg + bivar* suggests that, if we aim at a higher recall, using only one band at a time is preferable, and also that the bivariate features we chose are not adding any relevant information to the tested classifiers.
- When combining the r- and the g-band data (*rg* and *rg+bivar* classifiers), on the other hand, the precision we obtain is higher than in all the tests where only one band is used. This is due to a slightly higher TNR, and this result is consistent with several works from the literature showing that combining more bands usually returns less contaminated samples (e.g., Sánchez-Sáez et al. 2021; Savić et al. 2023), as well as with other yet-to-be-published results based on these same data where, combining three bands (g, r, and i), we obtain a purer sample of AGN candidates than the ones obtained from individual bands.
- Comparing the g- and r-band tests, we find that the recall of either class of AGN is slightly lower for the g-band classifier, while all the other metrics are generally slightly higher for the g-band classifier. If we focus on the TPRs, there are 0.4% more AGN correctly classified when using g-band features, which may suggest that, having a fixed number of 33 visits, the g-band dataset, where up to 16 points can be synthetic, is doing better than the r-band dataset, where only up to four points can be synthetic. In principle, this could be explained by the fact that we typically observe larger AGN variability in bluer bands (e.g., Petrecca et al. 2024). Nonetheless, we caution that we are comparing results in two different bands, one with mostly real visits and the other where about half the visits are synthetic. A more appropriate comparison would require making use of data from the same band to test the effect of the inclusion of synthetic visits, which we explore in the next section.
- The results from the test where color light curves are used are generally consistent with the others, except for the precision, which is the lowest in the whole upper section of the table.
- We stress that AGN recall, in general, is highly affected by the depth of the sample of sources under investigation. As a test, we computed the recall values for either class of AGN corresponding to different depths, obtained from the *gr* classifier, and we found that, for unobscured AGN, we go from a 100% recall value[4] when the average r-band magnitude is <20 mag to a 97.0% value for <22 mag, to the 98.2% value reported in Table 4 for <23.5 mag; for obscured AGN we obtained 16.0%, (39.4 ± 0.8)%, and (43.1 ± 1.0)% for <20 mag, <22 mag, and <23.5 mag samples, respectively. In this last case, the large uncertainties are due to the small sizes of the available samples of obscured AGN (6, 46, and 104, respectively), which get larger with depth. This is a crucial point to keep in mind if we attempt a proper comparison of our results with other works from the literature, where the samples of AGN used are typically brighter than ours and thus mainly consist of unobscured AGN and therefore generally return recall values higher than ours (e.g., Sánchez-Sáez et al. 2021).

4.1. Testing the impact of synthetic visits

As we mentioned in the previous section, when comparing the r- and g-band classifiers, we are comparing results in two different bands where the datasets have different properties: one contains mostly real visits, while in the other about half of the visits are synthetic. Here we propose a more appropriate comparison making use of data from the same band, which allows us to test the impact of adding synthetic visits to a "real" dataset. With this in mind, we based our analysis on 33 visits, that is, the number of visits that we had been using for most of our tests so far. We therefore extracted a set of 33 visits from the original r-band dataset, which includes 54 visits. We selected them so as to cover the full baseline of 3.3 yr, and we required 16 of them to be replaceable with synthetic visits following the criterion described in Sect. 3.1, that is, if we exclude one of them, the time difference between the two adjacent visits must be ≤15 days, so that we can replace that visit via linear interpolation. In this way, we could replicate with our r-band data a similar visit configuration to the one that we have for the g band, with 17 real visits and 16 synthetic visits. Once identified this set of 33 real visits, we proceeded with the following tests, where we progressively replaced four, eight, 12, and 16 real visits with as many synthetic visits, building each time an RF classifier that made use of r-band features only:

- 33 real visits and no synthetic visits;
- 29 real visits and four synthetic visits (25% of the maximum number of synthetic visits used in this work; this means that 12% of the total number of visits are synthetic);
- 25 real visits and eight synthetic visits (50% of the maximum number of synthetic visits used in this work; 24% of the total number of visits are synthetic);
- 21 real visits and 12 synthetic visits (75% of the maximum number of synthetic visits used in this work; 36% of the total number of visits are synthetic);
- 17 real visits and 16 synthetic visits (maximum number of synthetic visits used in this work; 48% of the total number of visits are synthetic).

We report the results obtained from each of these tests in the middle section of Table 4, as well as in Fig. 2, for a more immediate visualization. Essentially, we observe no noteworthy trend, which suggests that our imputation strategy is suitable to our purposes. There is a mild decrease in the recall of obscured AGN in the two bottom tests, as the fraction of synthetic visits increases, yet these values are consistent with the other ones in this section

[4] We do not report uncertainties here when they are <0.1%.

Table 4. Confusion matrix values for the various classifiers tested.

	TPR (recall)	TNR	Accuracy	Precision	Unobscured AGN recall	Obscured AGN recall	$F1$
r, 54 visits	79.5 ± 0.2	98.76 ± 0.06	95.88 ± 0.05	91.8 ± 0.3	99.54 ± 0.00	48.4 ± 0.6	85.2 ± 0.2
r, 33 visits	79.6 ± 0.9	98.23 ± 0.15	95.45 ± 0.22	88.8 ± 0.9	98.6 ± 0.4	50.1 ± 1.2	83.9 ± 0.8
g, 33 visits	80.0 ± 0.5	98.42 ± 0.15	95.67 ± 0.13	89.9 ± 0.9	98.2 ± 0.3	47.8 ± 1.6	84.7 ± 0.4
rg, 33 visits	76.8 ± 0.5	98.88 ± 0.07	95.59 ± 0.08	92.3 ± 0.5	98.2 ± 0.0	43.1 ± 1.0	83.9 ± 0.3
$rg + bivar.$, 33 visits	73.3 ± 0.7	99.06 ± 0.09	95.21 ± 0.14	93.2 ± 0.6	97.9 ± 0.2	36.9 ± 1.4	82.0 ± 0.6
$(g-r)_{feat}$, 33 visits	80.3 ± 0.7	98.34 ± 0.09	95.64 ± 0.12	89.4 ± 0.5	99.1 ± 0.2	48.7 ± 1.6	84.6 ± 0.5
$(g-r)_{mag}$, 33 visits	80.0 ± 0.4	97.84 ± 0.08	95.18 ± 0.09	86.7 ± 0.4	98.1 ± 0.3	50.0 ± 1.5	83.2 ± 0.3
r, 33 real, 0 synthetic	79.5 ± 0.5	98.30 ± 0.06	95.50 ± 0.09	89.2 ± 0.4	99.49 ± 0.15	48.4 ± 1.4	84.1 ± 0.4
r, 29 real, 4 synthetic	79.7 ± 0.5	98.16 ± 0.08	95.40 ± 0.09	88.4 ± 0.5	99.4 ± 0.2	48.5 ± 1.0	83.8 ± 0.3
r, 25 real, 8 synthetic	79.5 ± 0.4	98.07 ± 0.12	95.30 ± 0.09	87.9 ± 0.6	99.2 ± 0.3	48.2 ± 1.5	83.5 ± 0.3
r, 21 real, 12 synthetic	78.9 ± 1.0	98.29 ± 0.15	95.40 ± 0.17	89.0 ± 0.9	99.59 ± 0.15	45 ± 2	83.7 ± 0.6
r, 17 real, 16 synthetic	79.3 ± 0.4	98.23 ± 0.11	95.40 ± 0.09	88.7 ± 0.6	98.9 ± 0.4	46.3 ± 1.2	83.7 ± 0.3
ks, 25feat, 33 visits	82.2 ± 0.6	98.22 ± 0.08	95.82 ± 0.11	89.0 ± 0.4	98.7 ± 0.2	55 ± 2	85.5 ± 0.4
ks, 9feat, 33 visits	85.2 ± 0.3	97.75 ± 0.10	95.87 ± 0.07	86.9 ± 0.5	99.0 ± 0.2	66.7 ± 1.0	86.0 ± 0.2
ks, 8feat, 33 visits	85.8 ± 0.4	97.55 ± 0.11	95.80 ± 0.10	86.0 ± 0.5	99.1 ± 0.3	68.1 ± 1.2	85.9 ± 0.3
ks, 7feat, 33 visits	85.4 ± 0.6	97.54 ± 0.14	95.73 ± 0.15	85.9 ± 0.7	98.9 ± 0.2	67 ± 2	85.7 ± 0.5

Notes. The table compares true positive ratio (TPR) and true negative ratio (TNR); accuracy, overall precision values, recall for unobscured and obscured AGN, and $F1$ values obtained in this work are also included, the overall value of the recall being the same as the TPR. All values are to be read as percent values. The percentage errors represent the standard deviation from the mean value derived from a set of ten simulations per classifier. In each simulation, the classifier builds a number of trees as detailed in Appendix A to determine the final classification for each source. What is new in this work is the introduction of additional features (see Sect. 3.2). The last four lines report the corresponding values obtained for four RF classifiers discussed in Sect. 4.2: they are built with the aim of optimizing the identification of obscured AGN in this work.

of the table within their uncertainties. For each feature we compared the distributions obtained when all the 33 visits are real to the corresponding distributions obtained when 16 out of 33 visits are synthetic. We resorted to the Kolmogorov-Smirnov (K-S) test to identify the pairs of distributions that changed the most; we found that, for seven of them, the distance D returned from the test is >0.78 and the probability to obtain by chance a larger value is $<10^{-31}$. These features are, in descending order of D: `MaxSlope`, η^e, $\mathtt{IAR}_\phi$, `GP_DRW_tau`, `LinearTrend`, R_{cs}, and `Period_fit`. In Table 5 we report for each of these seven features the position in the importance ranking obtained for the two cases that we are comparing (33 real visits versus 17 real and 16 synthetic visits), in order to show where these features place themselves and to assess whether the changes in their distributions affect the ranking. It is apparent that the features that are higher in the ranking when all 33 visits are real (η^e, R_{cs}, and $\mathtt{IAR}_\phi$, which are in the top quartile) slide down to lower rankings when synthetic visits are introduced. This might be – at least in part – responsible for the mild drop that we observe for the recall of obscured AGN. Indeed, we anticipate that these features belong to the subset of features that we will select as the most suitable to identify obscured AGN; this is discussed in the next section, where we further investigate the issue of the identification of this class of AGN.

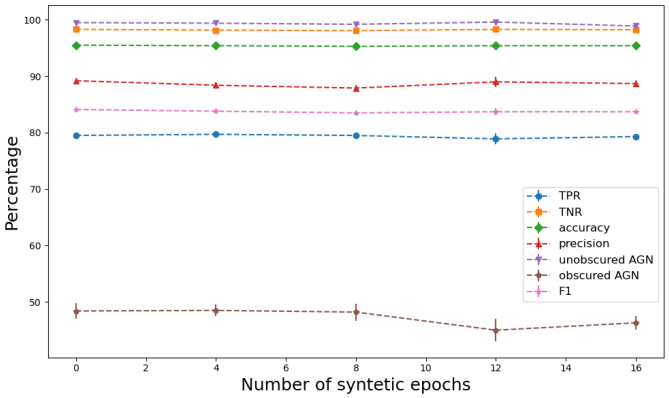

Fig. 2. Comparison of the results obtained from the five tests where real visits were progressively replaced by synthetic visits for the various metrics used in this work. The total number of visits is always 33, and we replaced part of them, four by four, with synthetic visits, up to a maximum of 16. The only error bars large enough to be visible correspond to the recall for obscured AGN.

4.2. Selection of obscured AGN

An interesting comparison concerns the recall of the samples of AGN retrieved using different classifiers: indeed, while it is well known that AGN selection based on optical variability is highly efficient in identifying unobscured AGN, it is also well known that it is generally not very effective in unearthing obscured AGN, as our series of VST-COSMOS works has widely proven (De Cicco et al. 2015, 2019, 2021, and references therein); as a consequence, even a small improvement in the recall of the obscured AGN that we are able to retrieve via optical variability is relevant. The largest value here obtained so far for the recall of obscured AGN is (50.1 ± 1.2)%: though this is still much lower than the corresponding value obtained for unobscured AGN, it undoubtedly shows a significant improvement if compared to the initial 6% value obtained in De Cicco et al. (2015) with a five month baseline, or the 18% value from De Cicco et al. (2019), where the baseline was the same as in this work, but the selection was based on a traditional approach, where we considered the r.m.s. deviation distribution and selected as variable AGN candidates all the sources with an r.m.s. deviation in excess of

Table 5. Position in the importance ranking for the seven features that were mostly affected by the replacement of 16 real visits with as many synthetic visits in order to test the impact of our data imputation strategy.

Feature	Ranking 33 real, 0 synthetic visits	Ranking 17 real, 16 synthetic visits
MaxSlope	26	24
η^e	4	9
IAR_ϕ	9	19
GP_DRW_tau	8	10
LinearTrend	18	15
R_{cs}	6	17
Period_fit	21	31

the 95% percentile.

We know that, when trying to unearth obscured AGN, the main difficulty our classifiers have to face is separating them from inactive galaxies. With this in mind, we inspected the distributions of all the features used in this work, comparing via the K-S test the ones obtained for obscured AGN and the corresponding distributions for inactive galaxies, aiming at selecting those features that seem to better disentangle the two classes of sources. Based on the results of the K-S test, we selected the features where the distance between the two distributions in a pair is large and the corresponding probability to get by chance a larger distance is small. We identified a possible threshold for distances $D > 0.25$, which equals keeping 25 of the initial set of 162 features, and we built a classifier using only the selected features, which are reported in Table 6. We note that no bivariate features are part of this selection; 12 out of 25 are *r*-band features, three are *g*-band features, four are obtained as differences between the two bands, five are colors, and one is the morphology indicator. Together with the classification, we obtained the feature importance ranking, shown in Fig. 3. The importance for each feature was computed as the mean of the importance values that that feature has in each of the trees built by the classifier where that feature was used, where the total importance of a feature in a single tree is defined as the sum of the impurity reductions across all nodes where that feature is used. The importance of a feature is indeed strictly connected to the reduction of the impurity that results from using that feature to split the sample, thus generating a node. Error bars were obtained as the standard error associated with the mean value for each feature, and only the trees where the feature were used (i.e., were part of the bootstrapped sample of features) were included in the calculation.

Based on the obtained feature importance ranking, we proceeded as follows: we excluded from our new set of features the least important one, than tested a new classifier using the remaining 24 features. We repeated this procedure eliminating each time the last feature in the ranking and building a new classifier with all the other features. We stopped when we were left with seven features, as we noticed a trend in the results obtained test after test and, based on that, at some point we did not expect a further reduction in the number of features to lead to any improvements. We report the metrics obtained from the classifiers using 25 features plus the ones using nine to seven features, in descending order, in the bottom section of Table 4. We omit the ones in between as the results are not particularly interesting. What is apparent from this section of the table is that

Table 6. Features selected as the ones that better disentangle obscured AGN and inactive galaxies based on the results of the K-S test comparing, for each feature, the corresponding distributions for these two classes of sources.

Feature	D	prob
Autocor_length_r	0.611	10^{-31}
SF_ML_amplitude_r	0.577	10^{-28}
SF_ML_gamma_r	0.547	10^{-25}
Autocor_length_g	0.448	10^{-17}
i-z	0.376	10^{-12}
r-i	0.366	10^{-11}
MedianBRP_diff	0.353	10^{-10}
R_{cs}_r	0.334	10^{-9}
u-B	0.317	10^{-8}
z-y	0.315	10^{-8}
Autocor_length_diff	0.296	10^{-7}
ch21	0.295	10^{-7}
Eta_e_r	0.291	10^{-7}
class_star_hst	0.287	10^{-7}
P_{var}_r	0.285	10^{-7}
ExcessVar_r	0.283	10^{-7}
MHAOV_Period_r	0.283	10^{-7}
GP_DRW_tau_r	0.281	10^{-7}
GP_DRW_sigma_r	0.275	10^{-6}
SF_ML_amplitude_diff	0.269	10^{-6}
Q31_g	0.267	10^{-6}
MedianAbsDev_g	0.262	10^{-6}
MaxSlope_diff	0.258	10^{-6}
Period_fit_r	0.256	10^{-5}
IAR_ϕ_r	0.251	10^{-5}

Notes. The features are listed in descending order of the distance D returned by the test, which represents the maximum distance between the two cumulative distributions compared; the table also reports the corresponding probability to obtain by chance a larger value for this distance.

there is no classifier whose performance is consistently better than the others. Hence, the choice of the features to use for a model depends on the results we aim at. If our goal is obtaining a sample of obscured AGN that be as complete as possible, then we should base our selection on the eight features used in the classifier named *ks8*. Indeed, in this case we managed to retrieve (68.1 ± 1.2)% of known obscured AGN, a result that almost doubles the one obtained in De Cicco et al. (2021). Of course, as it usually happens, this higher recall comes at the expense of a higher contamination, reflected by the lower TNR and the corresponding lower precision. Specifically, we note that the TNR obtained from this classifier is 1.51% lower than the highest value in the whole table, which corresponds to the *rg + bivar* classifier. We note that the recall for unobscured AGN is consistent through the various *ks* classifiers here discussed.

The feature importance ranking for the *ks8* classifier is shown in Fig. 4: we can see that the most important feature is, consistent with all the tests performed in this work as well as in De Cicco et al. (2021), the MIR color ch21; of the remaining features, three more are colors, one is the morphology indicator, and three are variability features. This final selection confirms once again the importance of combining light curves and color information to get improved results; see also Sect. 5.4 of De Cicco et al. (2021), which discusses the effects of using

colors alone, without variability features, to select AGN via an RF classifier. We also note that the last g-band feature surviving the iterative reduction in the number of features used is Q31_g, and it disappeared from the list when we were left with 12 features. From this point on, the only variability features used came from r-band light curves.

While in the bottom section of Table 4 the classifier with the highest TPR is *ks8*, an overall look at the values obtained for TNR, precision, accuracy, and $F1$ shows that the *ks9* classifier performs generally better than *ks8*. Therefore, in principle, we could choose to keep one more feature – namely, IAR_phi_r – at the expense of selecting 1.4% less obscured AGN. If we now look again at Table 4 as a whole, we can see that the reduction in the number of features used to build the various *ks* classifiers does not imply a significant drop in the accuracy (−0.08% if we compare the *ks8* value to the highest accuracy value in the table, obtained from the *r54* classifier), and it also returns a higher $F1$ while, as we mention above, the precision drops by 7.2% if we compare the *ks8* to the *rg+bivar* classifier, or by 5.8% if we compare the *ks8* classifier to one-band only classifiers and select the one with the highest precision, namely the *r*54 classifier. Hence, once again, while this specific part of our work is focused on optimizing the selection of obscured AGN, in general, depending on the purpose and on how one plans to use the sample of sources classified as AGN, one might want to favor recall over contamination or vice versa.

Figure 5 shows redshift as a function of the average r-band magnitude for all the AGN in the LS, separating the ones correctly classified (TPs) from the misclassified ones (FNs), based on the results from the *ks8* classifier. It is apparent that, while the two subsamples span quite uniformly the whole magnitude range covered by our dataset, the misclassified AGN are mainly lower-redshift sources and, based on what we showed in Fig. 1 and also on the recall values reported in Table 4, we know that these misclassified AGN are mostly obscured sources.

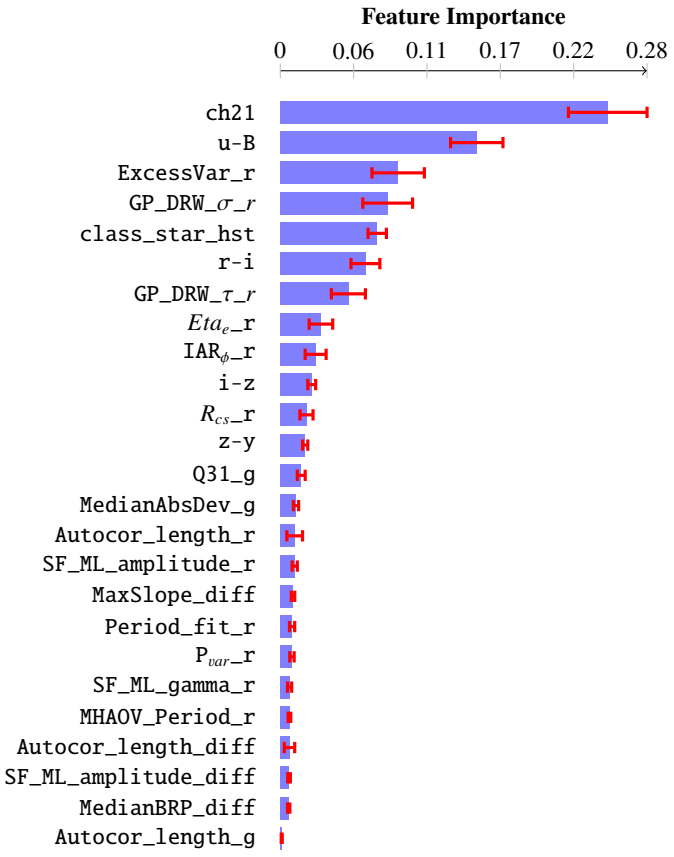

Fig. 3. Importance ranking for the top 25 features that, based on the results of the K-S test, allow a better separation between obscured AGN and inactive galaxies.

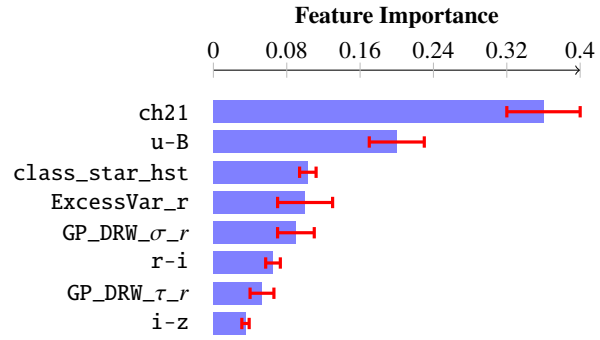

Fig. 4. Importance ranking for the eight features that were used to build the RF classifier identifying the highest fraction (i.e., returning the highest recall) of obscured AGN.

5. Summary and conclusions

This study has evaluated the effectiveness of an RF classifier trained on various feature sets to identify AGN. The chosen features mostly characterize the optical variability of a source and were derived from light curves in two different bands, used individually, jointly, combined as bivariate features, and subtracted as "color" indicators for each feature or light curve. In particular, we focused on how to optimize the selection of obscured AGN, which are typically more challenging to detect through optical variability. Of course we do not expect any AGN photometric selection techniques that rely on optical variability to be able to return a completeness for obscured AGN that be anywhere near 100%, as we know that their optical emission should be at least in part hidden because of the presence of a dust torus or whatever structure might be responsible for the obscuration of the accretion disk. Nonetheless, in this work we show the way to extract sizable samples of obscured AGN, and we aim at testing this method on other datasets, keeping in mind that its efficiency will always depend on the amount of obscuration, intrinsic luminosity, and so on.

In order to draw broader conclusions from our analysis, and also considering the various findings from our series of studies using VST-COSMOS data (De Cicco et al. 2015, 2019, 2021), we confirm the well-known fact that the observing cadence – and consequently, the total number of visits used for selection – is relevant, but we obtain comparable results from other classifiers. When defining the feature set obtained by single-band data alone, the best-performing algorithm among the ones tested is the only one that utilizes a larger (54) number of visits. In the various tests where the maximum number of visits is fixed to 33, the nature of the data points (real or synthetic) does not significantly impact the results, with slightly improved performance – except for the recall – when more synthetic points are added to the light curves (g-band test). However, tests examining the effect of synthetic points on real light curves suggest a possible decrease in the recall of obscured AGN as the fraction of synthetic to total number of points increases, while one important aim of this work is tailoring the selection method to increase the recall for obscured AGN.

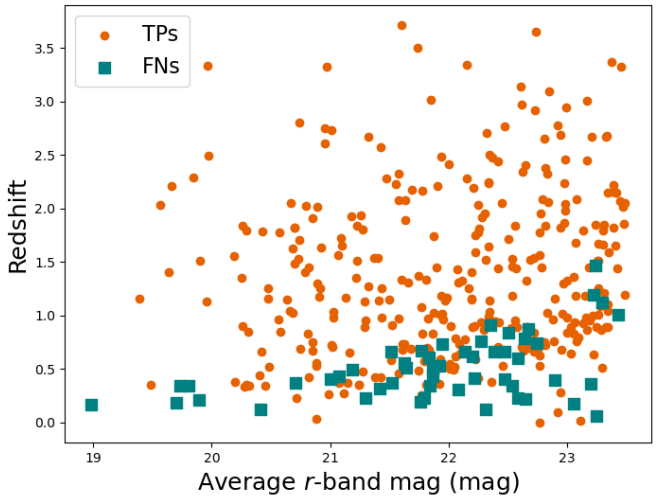

Fig. 5. Redshift as a function of the average *r*-band magnitude for the AGN correctly classified (TPs, dark green dots) and the misclassified AGN (FNs, magenta squares) from the *ks8* classifier.

Examining the three sections of Table 4 together reveals that the *ks8* classifier returns the highest fraction of obscured AGN, this being $(68.1 \pm 1.2)\%$, at the expense of a precision that is of several percents lower than the values obtained from the classifiers using more features (upper section of Table 4). This decrease in the precision originates from a larger contamination by false positives. This test also confirms the crucial role of the only color based on MIR data, as well as of optical/NIR colors (three out of five are among the top eight features in the ranking), and also shows how *r*-band features dominate over *g*-band ones; a possible explanation for this could be, as mentioned, the larger presence of synthetic visits in the *g*-band dataset, but it can be also correlated with the fact that for obscured AGN bluer wavelengths, such as the *g* band, are more absorbed with respect to the *r* band.

Given the challenges we have faced in previous studies in the effort of retrieving larger fractions of obscured AGN, the achieved recall of 68.1% indicates that building an accurate training sample and testing the optimal feature set is crucial to identify the obscured AGN population, and the feature set here identified could be valuable for further testing on different datasets, allowing us to assess their effectiveness in varying contexts, especially in view of wide-field surveys – such as the already mentioned LSST – which will provide us with much larger and richer source samples to investigate: indeed, with the LSST *ugrizy* filters we will have densely sampled light curves in more bands than the ones we used in this work. In particular, the addition of the *izy* filters will open up to a possible extension of our method to redder bands; if we focus once again on obscured AGN, typically enshrouded by dust and less affected by extinction compared to bluer bands, we expect their selection to benefit from the use of redder bands. This can therefore help detect variability that might be suppressed at the bluer wavelengths. In addition, analyzing their variability in the redder filters will allow us to better separate them from "inactive" galaxies, especially when combined with infrared data, which would allow detection of variability originating from dust reprocessing over longer scales. An important contribution in this wavelength regime is expected from the Euclid Mission (Euclid Collaboration: Mellier et al. 2025), with a plan for creating multiband catalogs of AGN and their host galaxies as a part of a set of Rubin-Euclid Derived Data Products (DDP; Guy et al. 2022). Another point is that, while our sample of obscured AGN does not extend to $z \gtrsim 1.5$, multiband variability can probe higher-redshift AGN as their variability signatures will be shifted into redder bands.

Our next goals while we wait for LSST data include the testing of our selection method optimized for the identification of obscured AGN on other datasets and, of course, once a sample of obscured AGN candidates is identified, we will need to validate it via other diagnostics and, when possible, via spectroscopic follow-up. In particular, we aim at testing our method over datasets of comparable or larger depth since, as also discussed in Sect. 4.2, a larger depth is necessary for the sample of obscured AGN to increase in size.

Acknowledgements. DD acknowledges PON R&I 2021, CUP E65F21002880003, and Fondi di Ricerca di Ateneo (FRA), linea C, progetto TORNADO. DD, MP, and VP acknowledge the financial contribution from PRIN-MIUR 2022 and from the Timedomes grant within the "INAF 2023 Finanziamento della Ricerca Fondamentale". SC acknowelges the ASI-INAF TI agreement, 2018-23-HH.0 "Attività scientifica per la missione Euclid - fase D", and PRIN MUR 2022 (20224MNC5A), "Life, death and after-death of massive stars", funded by European Union – Next Generation EU.

References

Allevato, V., Paolillo, M., Papadakis, I., & Pinto, C. 2013, ApJ, 771, 9
Amin, A., Anwar, S., Adnan, A., et al. 2016, IEEE Access, 4, 7940
Antonucci, R. 1993, ARA&A, 31, 473
Botticella, M. T., Cappellaro, E., Greggio, L., et al. 2017, A&A, 598, A50
Breiman, L. 2001, Mach. Learn., 45, 5
Brusa, M., Civano, F., Comastri, A., et al. 2010, ApJ, 716, 348
Bruzual, G., & Charlot, S. 2003, MNRAS, 344, 1000
Burbidge, G. R., Burbidge, E. M., & Sandage, A. R. 1963, Rev. Mod. Phys., 35, 947
Capaccioli, M., & Schipani, P. 2011, The Messenger, 146, 2
Cappellaro, E., Botticella, M. T., Pignata, G., et al. 2015, A&A, 584, A62
Cavuoti, S., De Cicco, D., Doorenbos, L., et al. 2024, A&A, 687, A246
Cha, S. H. 2007, Int. J. Math. Model. Meth. Appl. Sci., 1
De Cicco, D., Paolillo, M., Covone, G., et al. 2015, A&A, 574, A112
De Cicco, D., Paolillo, M., Falocco, S., et al. 2019, A&A, 627, A33
De Cicco, D., Bauer, F. E., Paolillo, M., et al. 2021, A&A, 645, A103
De Cicco, D., Bauer, F. E., Paolillo, M., et al. 2022, A&A, 664, A117
Donley, J. L., Koekemoer, A. M., Brusa, M., et al. 2012, ApJ, 748, 142
Edelson, R. A., Alexander, T., Crenshaw, D. M., et al. 1996, ApJ, 470, 364
Euclid Collaboration (Mellier, Y., et al.) 2025, A&A, 697, A1
Eyheramendy, S., Elorrieta, F., & Palma, W. 2018, MNRAS, 481, 4311
Falocco, S., Paolillo, M., Covone, G., et al. 2015, A&A, 579, A115
Fu, L., Liu, D., Radovich, M., et al. 2018, MNRAS, 479, 3858
Graham, M. J., Djorgovski, S. G., Drake, A. J., et al. 2017, MNRAS, 470, 4112
Green, P. J., Pulgarin-Duque, L., Anderson, S. F., et al. 2022, ApJ, 933, 180
Guy, L. P., Cuillandre, J.-C., Bachelet, E., et al. 2022, https://doi.org/10.5281/zenodo.5836022
Huijse, P., Estévez, P. A., Förster, F., et al. 2018, ApJS, 236, 12
Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111
Jeni, L., Cohn, J., & De la Torre, F. 2013, in Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013, 2013
Kim, D.-W., Protopapas, P., Byun, Y.-I., et al. 2011, ApJ, 735, 68
Kim, D.-W., Protopapas, P., Bailer-Jones, C. A. L., et al. 2014, A&A, 566, A43
Klesman, A., & Sarajedini, V. 2007, ApJ, 665, 225
Koekemoer, A. M., Aussel, H., Calzetti, D., et al. 2007, ApJS, 172, 196
Kormendy, J., & Richstone, D. 1995, ARA&A, 33, 581
Laigle, C., McCracken, H. J., Ilbert, O., et al. 2016, ApJS, 224, 24
LaMassa, S. M., Cales, S., Moran, E. C., et al. 2015, ApJ, 800, 144
Liu, D., Fu, L., Liu, X., et al. 2018, MNRAS, 478, 2388
Liu, D., Deng, W., Fan, Z., et al. 2020, MNRAS, 493, 3825
LSST Science Collaboration (Abell, P. A., et al.) 2009, arXiv e-prints [arXiv:0912.0201]
MacLeod, C. L., Ross, N. P., Lawrence, A., et al. 2016, MNRAS, 457, 389
Marchesi, S., Civano, F., Elvis, M., et al. 2016, ApJ, 817, 34
McLaughlin, M. A., Mattox, J. R., Cordes, J. M., & Thompson, D. J. 1996, ApJ, 473, 763

Nakos, T., Willis, J. P., Andreon, S., et al. 2009, A&A, 494, 579
Nun, I., Protopapas, P., Sim, B., et al. 2015, arXiv e-prints [arXiv:1506.00010]
Petrecca, V., Papadakis, I. E., Paolillo, M., De Cicco, D., & Bauer, F. E. 2024, A&A, 686, A286
Poulain, M., Paolillo, M., De Cicco, D., et al. 2020, A&A, 634, A50
Rees, M. J. 1984, ARA&A, 22, 471
Richards, J. W., Starr, D. L., Butler, N. R., et al. 2011, ApJ, 733, 10
Salpeter, E. E. 1964, ApJ, 140, 796
Sammut, C., & Webb, G. I. 2010, Leave-One-Out Cross-Validation (Boston, MA: Springer, US), 600
Sánchez-Sáez, P., Lira, P., Cartier, R., et al. 2019, ApJS, 242, 10
Sánchez-Sáez, P., Reyes, I., Valenzuela, C., et al. 2021, AJ, 161, 141
Sánchez-Sáez, P., Arredondo, J., Bayo, A., et al. 2023, A&A, 675, A195
Sarajedini, V. L., Koo, D. C., Klesman, A. J., et al. 2011, ApJ, 731, 97
Savić, Đ. V., Jankov, I., Yu, W., et al. 2023, ApJ, 953, 138
Schmidt, K. B., Marshall, P. J., Rix, H.-W., et al. 2010, ApJ, 714, 1194
Scoville, N., Abraham, R. G., Aussel, H., et al. 2007a, ApJS, 172, 38
Scoville, N., Aussel, H., Brusa, M., et al. 2007b, ApJS, 172, 1
Trevese, D., Pittella, G., Kron, R. G., Koo, D. C., & Bershady, M. 1989, AJ, 98, 108
Trevese, D., Kron, R. G., Majewski, S. R., Bershady, M. A., & Koo, D. C. 1994, ApJ, 433, 494
Trevese, D., Boutsia, K., Vagnetti, F., Cappellaro, E., & Puccetti, S. 2008, A&A, 488, 73
Tschopp, M., & Hernandez-Rivera, E. 2017, Quantifying Similarity and Distance Measures for Vector-Based Datasets: Histograms, Signals, and Probability Distribution Functions
Ulrich, M. H., Courvoisier, T. J. L., & Wamsteker, W. 1993, ApJ, 411, 125
Urry, C. M., & Padovani, P. 1995, PASP, 107, 803
Vanden Berk, D. E., Wilhite, B. C., Kron, R. G., et al. 2004, ApJ, 601, 692

Appendix A: Best hyperparameters obtained per classifier

Table A.1. Set of best values obtained from a grid search-based optimization of the five hyperparameters that typically have the most influence in the performance of an RF classifier.

RF classifier	n_estimators	max_depth	min_samples_split
r, 54 visits	500	10	10
r, 33 visits	100	10	10
g, 33 visits	300	20	2
rg, 33 visits	500	10	10
rg + *bivar.*, 33 visits	500	10	10
$(g-r)_{feat}$, 33 visits	100	10	10
$(g-r)_{mag}$, 33 visits	500	10	10
r, 33 real, 0 synthetic	300	10	2
r, 29 real, 4 synthetic	300	10	10
r, 25 real, 8 synthetic	100	10	10
r, 21 real, 12 synthetic	100	10	2
r, 17 real, 16 synthetic	300	10	10
rg, 25feat, 33 visits	300	10	10
rg, 9feat, 33 visits	100	20	2
rg, 8feat, 33 visits	100	20	10
rg, 7feat, 33 visits	100	10	2

Notes. The optimized hyperparameters are *n_estimators*, *max_depth*, *min_samples_split*, *min_samples_leaf*, and *max_features*; these were introduced in Sect. 4). The first column in the table lists the various classifiers tested in this work, whose performance metrics are reported in Table 4. We do not include a column for the last two hyperparameters since our tests always returned the same best values $min\_samples\_leaf = 4$ and $max\_features = sqrt$, the only exception being the *ks8* classifier, for which $max\_features = log2$.