

The Impact of Host-galaxy Properties on Supernova Classification with Hierarchical Labels

V. Ashley Villar^{1,2} , Sebastian Gomez³ , Edo Berger^{1,2} , and Alex Gagliano^{1,2} 

¹Center for Astrophysics | Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138-1516, USA

²The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, Boston, MA, USA

³Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA

Received 2024 July 17; revised 2024 August 28; accepted 2024 September 10; published 2024 December 11

Abstract

With the advent of the Vera C. Rubin Observatory, the discovery rate of supernovae (SNe) will surpass the rate of SNe with real time spectroscopic follow-up by 3 orders of magnitude. Accurate photometric classifiers are essential to both select interesting events for follow-up in real time and for archival population-level studies. In this work, we investigate the impact of observable host-galaxy information on the classification of SNe, both with and without additional light-curve and redshift information. We find that host-galaxy information alone can successfully isolate relatively pure (>90%) samples of Type Ia SNe with or without redshift information. With redshift information, we can additionally produce somewhat pure (>70%) samples of Type II SNe and superluminous SNe. Additionally with redshift information, host-galaxy properties do not significantly improve the accuracy of SN classification when paired with complete light curves. In the absence of redshift information, however, galaxy properties significantly increase the accuracy of photometric classification. As a part of this analysis, we present the first formal application of a new objective function, the weighted hierarchical cross entropy, to the problem of SN classification. This objective function more naturally accounts for the hierarchical nature of SN classes and, more broadly, transients. Finally, we present a new set of SN classifications for the Pan-STARRS Medium Deep Survey of SNe that lack spectroscopic redshift, increasing the full photometric sample to >4400 events.

Unified Astronomy Thesaurus concepts: Supernovae (1668); Light curve classification (1954); Neural networks (1933)

Materials only available in the [online version of record](#): machine-readable table

1. Introduction

Early classification of extragalactic transients, in particular supernovae (SNe), is paramount to enabling multiwavelength and spectroscopic analysis in real time. Currently, $\simeq 10\%$ of SNe receive a spectroscopic classification, the traditional means for understanding the underlying nature of an SN. The fraction of SNe which remain spectroscopically unclassified will significantly grow with the upcoming Legacy Survey of Space and Time (LSST) conducted by Vera C. Rubin Observatory and expected to commence in 2025. LSST will discover over one million SNe annually (R. Kessler et al. 2019); without additional spectroscopic resources, $\lesssim 0.1\%$ of all SNe will have spectroscopic follow-up. Large-scale spectroscopic follow-up campaigns, such as 4MOST (E. Swann et al. 2019), will increase this to $\sim 1\%$. As a result of this mismatch in discovery versus follow-up rates, the time-domain community has placed significant emphasis on photometric classifiers in the years leading up to LSST.

Photometric classifiers aim to classify SNe into their historically spectroscopic classes based on photometric data alone. The taxonomy of SNe has actively evolved as increasingly large samples unveil new diversity in SNe observables, leading to a branching hierarchical structure. Broadly speaking, Type II SNe are those which show spectroscopic signatures of hydrogen near

the peak of their optical emission. Within this class, the spectra of Type IIP/L SNe have broad H emission (typically with P-Cygni profiles), while Type IIn SNe are dominated by narrow H emission. In contrast, Type I SNe lack said features, with Type Ib SNe showing signs of helium and Type Ic SNe lacking signs of either. Types Ibc and II SNe arise from the core collapse of massive stars. Type Ia SNe, arising from thermonuclear explosions of white dwarfs, lack H and He in their optical spectra but show strong signatures of Si II near the peak. A broad overview of the hierarchical SN taxonomy is presented in A. Gal-Yam (2017).

Some photometric classification methods can be conducted in real time (e.g., D. Muthukrishna et al. 2019; A. Möller & T. de Boissière 2020; R. Carrasco-Davis et al. 2021; H. Qu & M. Sako 2022; A. Gagliano et al. 2023) while others rely on complete light curves (e.g., G. Hosseinzadeh et al. 2020; K. Boone 2021) for feature extraction. It can be especially challenging to identify and measure useful SN features in the early photospheric phases of the light curves. However, prompt identification of SNe is key to capturing early observational phenomena (e.g., flash spectroscopy for young core-collapse SNe (CCSNe); D. Khazov et al. 2016; R. J. Bruch et al. 2021; W. Jacobson-Galán et al. 2024), to guide multiwavelength follow-up and to optimize spectroscopic resources. In order to perform this early classification, all available information must be utilized, including the contextual information provided by the host galaxy of the transient.

Specifically, host-galaxy information is known to correlate with transient properties. For example, while thermonuclear



Original content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](#). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

(Type Ia) SNe are broadly observed across all galaxy types, CCSNe occur only in galaxies with recent or ongoing star formation (J. Leaman et al. 2011; A. Hakobyan et al. 2012; M. Childress et al. 2013). Type Ib and Ic SNe, specific subtypes of stripped-envelope SNe (SESNe), prefer slightly more-massive (and typically higher-metallicity) host galaxies compared to their hydrogen-rich counterparts of Type II(b) SNe (S. Schulze et al. 2021). Similarly, SESNe are over-represented in disturbed galaxies compared to their H-rich counterparts (S. M. Habergham et al. 2012). Rare CCSN classes seem to show stronger and more exotic preferences. Both Type Ic-BL (high-energy, “broad-lined” events) and H-poor superluminous SNe (SLSNe) prefer low-mass (metal-poor) galaxies with high specific star formation rates (P. L. Kelly & R. P. Kirshner 2012; S. Schulze et al. 2021). On local scales, SESNe strongly trace H α (a tracer of ongoing star formation), while Type II SNe tend to show larger spreads in local environment properties (J. P. Anderson et al. 2012). In contrast the Type IIn SNe, which show signs of interaction between the SN blast wave and dense, preexisting circumstellar material, show highly heterogeneous local environments (C. L. Ransome et al. 2022). The rarer Ca-rich SNe of yet unknown progenitor origin show strong preference for high offsets in their host galaxies, although they are not preferentially found in star-forming hosts (M. M. Kasliwal et al. 2012; Y. Dong et al. 2022). These correlations are clues toward the underlying progenitor populations for each of these SN classes. Here, however, we utilize the correlations between global host properties and transients as a tool to help classify the underlying physics of the SNe.

Classifiers that utilize galaxy information for SN identification have been explored in the literature, although they have been primarily limited to binary classification tasks. R. J. Foley & K. Mandel (2013) introduced the first galaxy-based SN classifier based on a Naive Bayes architecture with the LOSS SN sample. In a similar vein, A. Baldeschi et al. (2020) showed that galaxy morphology and star formation could be used to increase the purity of thermonuclear and CCSN samples compared to randomly guessing. A. Gagliano et al. (2021) introduced a random forest classifier based on host-galaxy properties from the GHOST sample of \simeq 16,000 SNe associated with host galaxies, finding that such a method could perform thermonuclear versus CCSNe classification with \simeq 68% accuracy. S. Gomez et al. (2020) presented one of the first host-aware classifiers that extend beyond Ia/CCSNe classification, focusing instead on a Type I SLSNe versus non-SLSNe classification from a combination of host and light-curve properties; a classifier based on host-galaxy properties alone is not explored in that work. M. Kisley et al. (2023) presented a hierarchical SN classifier based solely on optical and (near)-infrared host-galaxy photometry from the THEX catalog (Y.-J. Qin et al. 2022). They use a likelihood-based approach reminiscent of a Naive Bayes classifier, using a series of binary classifications to distinguish 11 SN classes (in addition to tidal disruption events). They were able to classify Type Ia subtypes and Type II SNe with a purity statistically above random guessing. Most recently, A. Gagliano et al. (2023) presented a neural-network-based classifier for Zwicky Transient Facility (ZTF) Bright Transient Survey data which utilizes the light-curve and host-galaxy information for Type Ia, II, and Ibc SN classification, achieving an accuracy of \simeq 82% within 3 days (observer frame) of discovery.

Furthermore, the hierarchical nature of SN taxonomy is rarely used in photometric classification. P. Sánchez-Sáez et al. (2021) presented a broad hierarchical classifier for ZTF data. There, multiple “flat” (i.e., those lacking a hierarchical structure) random forest algorithms are trained at each “level” of the hierarchical classifier. For example, all SNe are classified as transient phenomena, then a second flat classifier is trained to classify them as one of four SN types. M. Kisley et al. (2023) consider the hierarchical nature of SNe in the sense of a prior distribution, which is then used in the classifier’s objective function. For example, the rate of Type Ia-91bg events are included when calculating the rate of Type Ia SNe, which is then used to define a Type Ia prior probability within the objective function. However, this prior information is not used in their primary analysis.

Here, we present a broad, hierarchical classifier based on host-galaxy properties alone and explore how galaxy properties can improve our transient classifications. We specifically test how simple, measurable properties of the host impact our classification accuracy both with and without redshift information. The paper is organized as follows. In Section 2, we review the data used in this work and the features selected for our SN classifiers. In Section 3 we present the weighted hierarchical cross-entropy score to more naturally account for the hierarchical nature of SN classification. We additionally discuss the architecture of our neural-network-based classifier used here for classification. In Section 4, we discuss the results of our classifiers for different subsets of SN classes, feature sets, and the inclusion/exclusion of redshift information. In Section 5, we present new classifications of SNe within the Pan-STARRS Medium Deep Survey (PS1-MDS) sample. We conclude in Section 6. Throughout this work, we assume a flat Λ CDM cosmology with $H_0 = 67.8 \text{ km s}^{-1} \text{ Mpc}^{-1}$ and $\Omega_m = 0.308$ (Planck Collaboration et al. 2020).

2. Data and Feature Selection

We aim to understand if and when host-galaxy features can aid in SN photometric classification. Therefore, we require a SN sample for which both a light-curve-only classifier and a host-galaxy classifier are both available. While there now exist two large catalogs of host galaxies with associated SNe, GHOST (A. Gagliano et al. 2021) and THEX (Y.-J. Qin et al. 2022), we instead opt to use a uniform survey in order to directly compare the light-curve versus host features. For this reason, we utilize the PS1-MDS catalog of SN-like light curves published by V. A. Villar et al. (2020) and G. Hosseinzadeh et al. (2020). In total, our sample includes 557 spectroscopically identified SNe, in five classes: Type Ia SNe (404 objects), Type II SNe (93 objects), Type IIn SNe (24 objects), Type Ib/c SNe (19 objects), and Type I SLSNe (17 objects). Our Pan-STARRS sample has the additional benefit of being a close analog to Vera C. Rubin data in terms of cadence, filter selection, and depth, offering a realistic view of how our algorithm will perform on the LSST data stream.

We use the full set of light-curve features available from SuperRAENN (44 in total; V. A. Villar et al. 2020), which includes:

1–8: Nonlinear features extracted from an autoencoder, trained on the PS1-MDS data set. The correlations between these data-driven features and “typical” observational features are explored in more detail in V. A. Villar et al. (2020).

9–21: The *griz* rise times, calculated using 1, 2, and 3 mag below peak. Note that these are calculated using light curves which have been interpolated via a 2D Gaussian process. These are, when redshift is utilized, calculated in the rest frame of the SN (using only the cosmological k -corrections).

22–32: The *griz* decline times, calculated for 1, 2, and 3 mag below peak.

33–36: The *griz* peak magnitude. When redshift is utilized, these are absolute magnitudes.

37–40: The median slope measured in *griz* between 10 and 30 days postpeak in the observer frame. Note that these features in particular help distinguish between Type II and Type Ibc SNe.

41–44: The integrals of the interpolated *griz* light curves.

To associate SNe with host galaxies and extract galaxy features, we use the “Finding Luminous and Exotic Extragalactic Transients” (FLEET) pipeline (S. Gomez et al. 2020, 2023). FLEET is a classification methodology designed to classify rare, extragalactic transients (SLSNe and tidal disruption events) using a combination of host-galaxy and light-curve information. FLEET queries a 1' region of the Pan-STARRS 3 π survey around a given transient to identify the most likely host galaxy using the follow algorithm. First, a probability of being a galaxy (as opposed to a star) is assigned to every object in the field, where 0 means most like a star, and 1 means most likely a galaxy. This probability is estimated using a custom k-nearest neighbors algorithm trained on data from the Canada–France–Hawaii Telescope Legacy Survey, which has complete star/galaxy labels down to ≈ 26 mag (P. Hudelot et al. 2012). Then, the probability of chance coincidence P_{cc} of every galaxy in the search region is calculated using the J. S. Bloom et al. (2002) method described in E. Berger (2010). More precisely, we follow Equation (2) of S. Gomez et al. (2020):

$$P_{cc} = 1 - e^{-\pi(d^2+4R^2)\Sigma(\leq m)} \quad (1)$$

$$\Sigma(\leq m) = \frac{10^{0.33(m-24)-2.44}}{0.33 \ln(10)}, \quad (2)$$

where d is the angular separation between the transient and its host in arcseconds; R is the half-light radius of the host also in arcseconds; and m is the i -band magnitude of the host (or the r band if the i band is not measured). We select the galaxy with the lowest P_{cc} as the host galaxy of the transient, or (for SNe in which $P_{cc} > 0.1$) consider the transient to be “hostless.” Hostless transients are retained in the data set, as these are often associated with rare CCSN types (e.g., SLSNe and Type IIn). The selection of $P_{cc} > 0.1$ is somewhat arbitrary, but found to be a reasonable threshold in S. Gomez et al. (2020).

Many host-galaxy features are available via FLEET, including multisurvey observables (e.g., magnitudes), derived properties (e.g., photo- z), and inferred properties more directly related to the transient (e.g., offset). We restrict ourselves to data from the Pan-STARRS 3 π survey, in order to minimize missing features in our data set. We use the following features from the pipeline:

1–5: g -, r -, i -, z - and y -band Kron magnitudes, estimated using SExtractor.

6–8: g -, r - and i -band Kron radii, estimated using SExtractor.

9: Host-galaxy separation from the transient in arcseconds. If the transient is deemed “hostless,” the offset value is set to zero.

10: Point-source score, a measurement of probability of the object being an extended object (score of 0) or a point source (score of 1). This is a property within the PS1 3 π Survey.

11: Number of detections in all bands from the independent images used to generate the stacked image.

12: Host-galaxy half-light radius in i band or in r band if i band is not available.

13: Host-galaxy nature, a measurement of probability of the object being a galaxy (score of 1) or a star (score of 0). This is a custom function within FLEET.

14: Probability of chance coincidence, as calculated by FLEET.

All magnitudes are corrected for Milky Way extinction. In some of our pipelines, we additionally include the redshift (spectroscopically measured from the host or transient) as a feature.

Although we attempt to minimize missing data, a small fraction of events are missing some observational features. Most often (in four cases of the 557 objects in the spectroscopic data set), the point-source score and uncertainties on the measured magnitudes are missing due to a transient being “hostless”. We use a K-means imputation method (KNNImputer in scikit-learn) to fill in the missing data. This method utilizes the information from the $K=3$ neighbors in the 13-dimensional galaxy feature space, and fills in missing values with an average from these neighbors. This imputation method works as expected with no ad-hoc corrections: for hosts that are not detected due to missing data (e.g., they were never observed in the z band), the nearest neighbor provides a reasonable estimate of the missing properties. For hosts near the survey limit, the method naturally fills in the missing data with dim apparent magnitudes and small observed radii. In testing, we find that changing the value of k has minimal impact on results. Finally, we normalize our data such that each feature values between 0 and 1 (MinMaxScaler in scikit-learn).

We note that there are clear correlations between SN type and host-galaxy properties, visible without the aid of a specialized classifier. We show a number of representative feature spaces in Figure 1. Brighter objects (SLSNe, Type IIn SNe) tend to occur in dimmer galaxies. This is likely an observational bias of PS1-MDS, as our SLSNe and Type IIn SNe sample skew to higher redshifts (see Figure 1 of V. A. Villar et al. 2019). The color of these galaxies also clearly strongly correlates with type. For any given magnitude, CCSNe (especially Type II SNe) are more likely to occur in bluer galaxies (i.e., those with ongoing star formation).

3. Classification Methods

3.1. A Novel Hierarchical Loss Function

We explicitly include the hierarchical nature of SN taxonomy within our analysis via a weighted hierarchical cross-entropy (WHXE) objective function (also presented in V. A. Villar et al. 2023 for SNe and variable stars). L. Bertinetto et al. (2020) originally introduced the HXE to classify images with a similar hierarchical taxonomy. They compare the performance of HXE to standard cross entropy, finding that the two perform similarly in overall accuracy but that HXE “makes better mistakes,” i.e., the HXE enforces a graph structure that encourages the classifier to

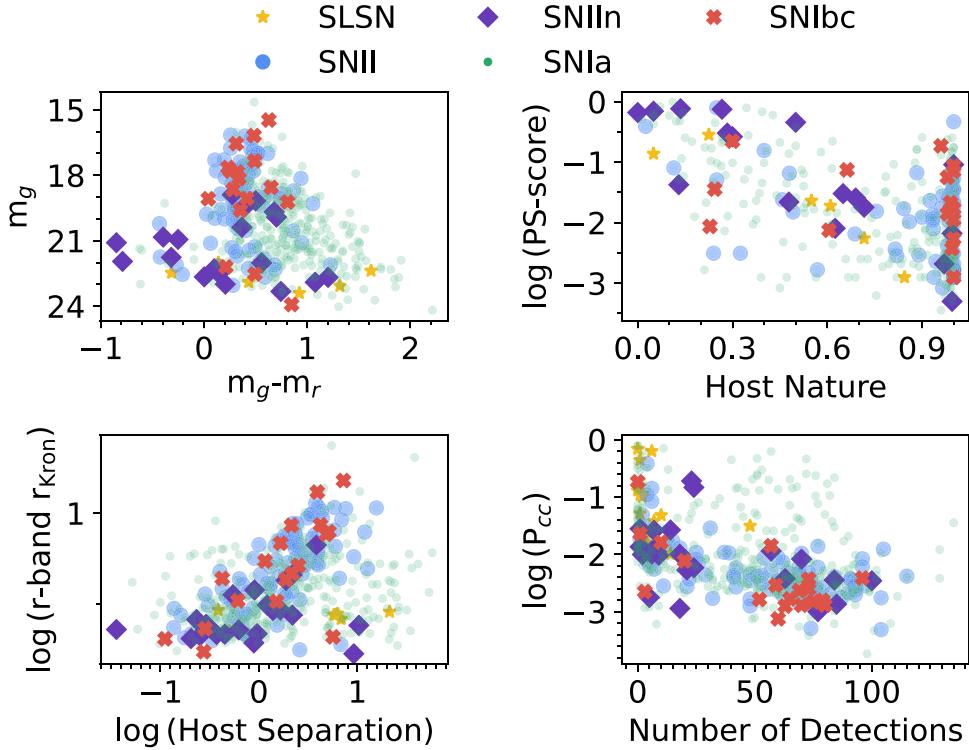


Figure 1. Scatter plot of various host-galaxy features to classify SNe in this work (see Section 2 for feature definitions). SNe are spectroscopically labeled as part of PS1-MDS, and for features where the host-galaxy features were not observed, they are inferred via K-means imputation. Even in these simple feature spaces, clear clustering is seen.

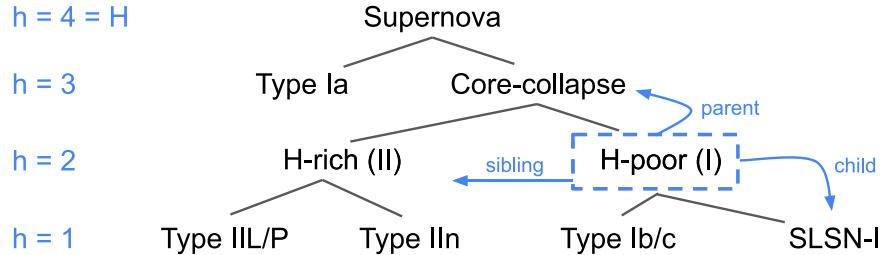


Figure 2. Hierarchical graph structure used in this work. Blue text is meant as a guide to the various components of a generic tree.

place objects in the correct broader category. In our SN context, the WHXE allows us to (1) train a multilayered classifier that can easily perform binary and general classification and (2) encourage transients to be misclassified within their broader class (e.g., CCSNe are less likely to be classified as Type Ia SNe). In this section, we define the graph structure used for our SN classification and WHXE in detail.

The graphical nature of our classification hierarchy is shown in Figure 2. Given the small data set size, we have not subdivided Type Ia, Type IIP/L, and Type Ib/c SNe into additional subclasses. We emphasize, however, that this is not the only choice for a classification graph. For example, some bright Type IIn SNe share similar hosts to SLSNe. One could include galaxy information in the graph structure (e.g., preferring to classify bright transients in low-luminosity galaxies together), although the classification bias would be less physically motivated. Alternatively, one could follow the “classic” classification schema, first distinguishing Type I versus Type II SNe (rather than thermonuclear versus core collapse). This may aid in isolating hydrogen-rich transients, but is not explored in this work.

We next describe the WHXE objective function, largely following the notation of V. A. Villar et al. (2023). The classic categorical cross entropy, most often used in classification tasks, is defined as

$$\mathcal{L} = - \sum_i^C t_i \log(p(c_i)), \quad (3)$$

where C is the total number of classes, c_i is a specific SN class, and t_i is an indicator variable for the true SN class. This indicator function means that the cross-entropy score only rewards the classifier based on assigning a high likelihood to the true class. There is no reward structure for how probabilities are distributed over the incorrect classes.

We contrast this to the WHXE, which utilizes the hierarchical graph structure of the SN taxonomy by factorizing each class by its leaf nodes. We represent the SN class “height” on the tree as $c_i^{(h)}$. The root (the “SN” parent class) is represented as $c_i^{(H)}$ (where $H=4$ is the height of the tree), and a lowest leaf (e.g., the Type Ib/c class) is represented as $c_i^{(1)}$. For any given class c_i with height h' , the classification probability

can be written as

$$p(c_i) = \prod_{h=h'}^{H-1} p(c_j^{(h)}|c_k^{(h+1)}), \quad (4)$$

where c_j and c_k are ancestor nodes of class c_i (i.e., nodes on the path from the leaf to the root). Here, we are assuming that the probability of being an SN (the root node) is equal to 1. The conditional probabilities can be written more explicitly as

$$p(c_j^{(h)}|c_k^{(h+1)}) = \frac{\sum_{c_a \in \text{Siblings}(c_j^{(h)})} p(c_a)}{\sum_{c_b \in \text{Siblings}(c_k^{(h+1)})} p(c_b)}, \quad (5)$$

where $\text{Siblings}(c)$ represent the sibling set of node c (i.e., those sharing the same parent). As an example, the probability that an SN is Type Ib/c given it is H-poor is equal to the probability that it is a Type Ib/c (as opposed to SLSN-I) divided by the probability that it is H-poor (as opposed to H-rich).

Analogous to Equation (3), we can define the WHXE as

$$\mathcal{L}_{\text{WHXE}}(c_i^{(h)}) = - \sum_{h=h'}^{H-1} W(c^{(h)}) \lambda(c^{(h)}) \log p(c_j^{(h)}|c_k^{(h+1)}), \quad (6)$$

where $W(c^{(h)})$ weights each SN class, and $\lambda(c^{(h)})$ is a weighting which emphasizes each level of the taxonomy tree (as a function of class height, h). For $W(c^{(h)})$, we choose to weight each SN by $N_{\text{All}}/(N_{\text{Labels}} \times N_{c_i})$, where N_{All} is the total number of SNe in our training set, N_{Labels} is the number of SN classes (all nodes), and N_{c_i} is the number of SNe of type c . L. Bertinetto et al. (2020) suggest the following form for $\lambda(c^{(h)})$:

$$\lambda(c^{(h)}) = \exp(-\alpha h), \quad (7)$$

where α is a free parameter. Larger values of α weight the top of the hierarchy more strongly (i.e., Ia versus CC classification); lower values of α weight each level of the hierarchy equally, emphasizing fine-grained classifications; α is therefore a hyperparameter of our model which we optimize.

Based on our graph in Figure 2, our objective function requires nine outputs from our classifier, reflecting a probability at each node of the graph. The top node (the “SN” designation) is always equal to 1. A softmax function is then applied to each branch of the tree such that the neural-network output can be interpreted as a conditional probability. For example, one output represents the probability⁴ of being H-rich and another represents the probability of being H-poor. As these are a pair of siblings, these outputs will be renormalized such that their sum is equal to 1. The height of each output is also tracked, with an appropriate weighting (a function of α) applied. A working PyTorch implementation of the WHXE is presented in V. A. Villar et al. (2023) and available via Zenodo (V. A. Villar 2024) and GitHub.⁵

3.2. Multilayer Perceptron Classifier

Throughout this work, we use a fully connected multilayer perceptron (MLP, a simple neural network) to classify the SN subclasses. The MLP transforms an input feature vector into an output probability vector which optimizes the WHXE objective

⁴ Note that these are not “true” probabilities as they are not properly calibrated. Instead, they share qualitative similarities to a probability vector: they contain nonnegative values which sum to one.

⁵ <https://github.com/VTDA-Group/hxe-for-tda/>

function. Between these are a series of “hidden” layers with optimizable weights and nonlinear activation functions. Here, we use the standard rectified linear unit. We optimize the MLP using the standard Adam optimizer (D. P. Kingma & J. Ba 2014), a momentum-based gradient descent algorithm. Our model and training procedure is built in Pytorch.

In total, our hyperparameters of the model are: (1) the hierarchy weighting, $\alpha \in \{0.0, 0.1, 1.0, 3.0\}$; (2) the learning rate, $\beta \in \{0.001, 0.005, 0.01, 0.05, 0.1\}$; (3) the batch size $\in \{16, 32, 64, 128, 256\}$; and (4) the number of neurons per layer, a number from the set $\in \{3, 5, 10\}$. Using each combination of hyperparameters, we train an MLP for 300 epochs with early stopping and select the optimal set for each feature set. We train six unique classifiers using different light-curve/host features (see Section 4 for details). However, the best hyperparameter values (as determined by tracking the F1-score) do not vary greatly. Typically, the best hierarchy weighting is $\alpha \simeq 1$, the learning rate on the higher end $\beta \simeq 0.05$, the batch size is on the higher end (128–256), and the number of neurons is on the higher end (5–10). Our model takes minutes on a single CPU to complete training. We note that we also attempted to use synthetic minority oversampling techniques to reweight our training set (as opposed to including the class weights in our objective function), finding somewhat worse results.

4. Classification Results and Discussion

Our goal is to understand how contextual host-galaxy information improves classification performance of SNe both with and without redshift information. Therefore, we train and compare six classifiers in total: (1) one which uses solely the observer-frame host-galaxy information; (2) one which uses observer-frame host-galaxy information and redshift; (3) one which uses solely the observer-frame light-curve information; (4) one which uses the rest-frame light-curve and redshift information; (5) one which uses the galaxy information and the observer-frame light-curve information; and (6) one which uses the galaxy, rest-frame light-curve, and redshift information.

We track the purity, completeness, and F1-score of each classifier, defined as

$$\text{Purity} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (8)$$

$$\text{Completeness} = \frac{\text{TP}}{\text{TP} + \text{TN}} \quad (9)$$

$$\text{F1} = \frac{2 \times (\text{Purity} \times \text{Completeness})}{\text{Purity} + \text{Completeness}} \quad (10)$$

TP is the true positive rate or the fraction of SNe within a given class correctly identified as belonging to said class. TN is the true negative rate or the fraction of SNe not in a given class and correctly identified as not being a member of said class. Finally, FP is the false positive rate or the fraction of SNe which are identified as belonging to a given class but in fact are not of that class. The F1-score is the harmonic mean of the purity and completeness and a commonly used metric for evaluating classifiers.

We report the purity, completeness, and F1-score for each SN class, as well as the class-averaged and the “weighted” versions of these metrics. Here, the “weighted” averages reweight each class to represent the total number of objects in

each class. In this case, the statistic is dominated by the majority class, Type Ia SNe. Note that we do not use accuracy to evaluate each classifier, which can be a particularly poor metric of success for highly imbalanced training sets; however, when appropriate (i.e., when comparing to other works), we report accuracy of specific models.

We test our classifier on three classification tasks: (1) five-way classification (SLSN, Type II, Type IIn, Type Ia, and Type Ibc); (2) three-way classification (Type II, Type Ia, and stripped envelope); and (3) binary classification (core collapse versus Type Ia). As described above, the classifier does not need to be retrained for each of these tasks; instead, we calculate the conditional probabilities of each, assigning the final label of each object as the category with the highest conditional probability, e.g., for Type IIn SNe:

$$p(\text{IIn}) = p(\text{CC})P(\text{H-rich}|\text{CC})P(\text{IIn}|\text{H-rich}). \quad (11)$$

In total, we compare 18 combinations of feature sets and output classes. Our results are fully summarized in Table 1 and visually summarized in Figure 3. Throughout this section, we will exclude uncertainties (which have been calculated using 10 random seeds for each MLP) for purity, completeness, and F1-scores. However, these values are listed within Table 1 and are typically $\lesssim 0.05$.

4.1. Classification Performance with Host-galaxy Information Only

We first explore the performance of SN classification from host-galaxy information, with and without redshift information (the first two columns of Table 1). Our results are primarily visualized in the confusion matrices of Figure 4.

In the five-way task without redshift information, we achieve a poor average F1-score of 0.36 across all classes. We find class completeness ranging from 0.28 (for Type Ibc SNe) to 0.59 (for Type Ia SNe). We see a much wider spread in purity, from just 0.11 (for Type Ibc SNe) to 0.86 (for Type Ia SNe). Unsurprisingly, the majority of incorrectly labeled SNe are dominated by Type Ia SNe, making up $\gtrsim 50\%$ of each sample. Including redshift information, our average F1-score greatly improves from 0.36 to 0.45. Classification purity and completeness also increase to 0.41 and 0.51, respectively. The class completeness ranges from 0.27 (for Type IIn SNe) to 0.83 (for SLSNe). The improvement in the latter is not surprising, as SLSNe are intrinsically luminous and can therefore be uniquely found at higher redshift for a magnitude-limited survey like PS1-MDS (and this will similarly be true in LSST). The spread in purity is still quite high, ranging from 0.09 (for Type Ibc SNe) to 0.91 (for Type Ia SNe). If we impose a confidence cutoff ($p > 0.8$), we find a substantial improvement in overall performance, with the (unweighted) average purity increasing from 0.32 to 0.41, and the completeness from 0.40 to 0.50, suggesting that purer samples of SNe can be collected with these cuts. Type Ia purity, in particular, increases to 0.94, suggesting that a highly pure sample of cosmological Type Ia SNe can be selected from host-galaxy properties alone. Similarly, SLSNe and Type II SNe can achieve relatively high purity (> 0.7) with high thresholds ($p > 0.8$) when redshift information is included. Unfortunately, the purity of Type IIn and Type Ibc SNe remain low ($\lesssim 0.2$) even at this high cutoff.

In the three-way classification task, the galaxy-only classifier achieves a notably higher F1-score of 0.49 even when

excluding redshift information. The average purity (0.46) and completeness (0.52) also increase. Here, again, the minority classes (stripped-envelope and H-rich SNe) suffer from generally low purity scores (0.14 and 0.36, respectively), while Type Ia SNe can be classified with a high purity of 0.88. Including redshift information, the purity and completeness again show positive trends for each class, particularly for SESNe, whose purity increases from 0.14 to 0.22. The F1-score, again, significantly improves from 0.49 to 0.54.

In the binary classification task (Type Ia versus CCSN), we find average class purity of 0.64 and completeness of 0.67. This final classifier is most readily compared to literature studies. The binary average accuracy of our classifier (0.66) is very similar to that of GHOST (A. Gagliano et al. 2021), which reported a class-average accuracy of 0.68. Including redshift information, we increase this accuracy to $\simeq 0.74$, showing state-of-the-art performance in the binary classification task.

Similarly, in the three-way classification task, we find that our classifier outperforms state-of-the-art in the literature. In particular, we compare the performance of our host-only classifications to those of A. Gagliano et al. (2023). A. Gagliano et al. (2023) uses early (within three days of detection) light-curve information, redshift, and host information to classify SNe. They find, using the same three-way classification schema, a class-balanced F1-score of 0.48. Without redshift information, we achieve an F1-score of 0.49 and with redshift information this substantially increases to 0.54.

We additionally compare the performance of our classifier on the specialized FLEET classifier for SLSNe. S. Gomez et al. (2020) present three versions of the FLEET classifier: one which only uses host-galaxy information and early light-curve information (e.g., rise); one which uses host-galaxy information and redshift; and one which uses host properties, the full light curve, and redshift information. Here, we compare to both the first and second versions. Without redshift information, we achieve low purity in SLSNe (with a maximum purity of $\simeq 25\%$ with a threshold $p > 0.5$). In contrast, FLEET boasts a purity of $\simeq 85\%$ (with a corresponding completeness of $\simeq 20\%$) without redshift information. However, we find that—given redshift information—our classifier notably outperforms that of FLEET. We achieve a purity $\simeq 85\%$ at a corresponding completeness of $\sim 50\%$. FLEET achieves a similar purity at a completeness of $\simeq 25\%$.

The key takeaways from the galaxy-based classifier are threefold. First, we are able to successfully isolate a pure ($\simeq 0.9$) sample of Type Ia SNe with host-galaxy information alone at a reasonably high completeness ($\simeq 0.6$) even without any redshift information. Second, using redshift information, we can produce reasonably pure samples of Type II SNe and SLSNe, but not Type IIn and Type Ibc SNe. Finally, in the absence of any light-curve information, redshift does improve the performance of our classifier for all classes, suggesting that accurate photo- z information will be valuable for rapid classification.

4.2. Classification Performance Using Light-curve and Galaxy Features

We next turn our attention to a combined feature set that includes both light-curve and host-galaxy features. First, we focus on the case where redshift information is not known. Although photometric redshift estimates are often available in current surveys and will be available in the era of LSST, the

Table 1
Classification Performance

		Gal. without z	Gal. with z	LC without z	LC with z	Gal. + LC without z	Gal. + LC with z
Five-way Classification							
SLSN	F1	0.19 (0.03)	0.63 (0.11)	0.23 (0.02)	0.77 (0.05)	0.29 (0.05)	0.73 (0.05)
	Purity	0.12 (0.02)	0.50 (0.14)	0.20 (0.02)	0.73 (0.07)	0.25 (0.06)	0.73 (0.06)
	Completeness	0.47 (0.06)	0.83 (0.09)	0.27 (0.04)	0.80 (0.05)	0.33 (0.09)	0.73 (0.08)
II	F1	0.38 (0.02)	0.46 (0.06)	0.52 (0.03)	0.71 (0.03)	0.56 (0.04)	0.73 (0.02)
	Purity	0.39 (0.03)	0.41 (0.04)	0.55 (0.05)	0.79 (0.04)	0.62 (0.04)	0.79 (0.03)
	Completeness	0.36 (0.03)	0.53 (0.14)	0.51 (0.05)	0.65 (0.03)	0.51 (0.05)	0.67 (0.02)
IIn	F1	0.16 (0.02)	0.16 (0.03)	0.34 (0.04)	0.38 (0.06)	0.37 (0.04)	0.41 (0.04)
	Purity	0.10 (0.02)	0.11 (0.02)	0.29 (0.05)	0.32 (0.08)	0.29 (0.04)	0.41 (0.06)
	Completeness	0.33 (0.05)	0.27 (0.10)	0.42 (0.09)	0.46 (0.06)	0.50 (0.07)	0.42 (0.05)
Ia	F1	0.70 (0.02)	0.73 (0.04)	0.88 (0.01)	0.94 (0.02)	0.91 (0.01)	0.94 (0.01)
	Purity	0.86 (0.01)	0.91 (0.01)	0.93 (0.01)	0.95 (0.01)	0.94 (0.01)	0.93 (0.01)
	Completeness	0.60 (0.03)	0.61 (0.06)	0.84 (0.02)	0.93 (0.04)	0.89 (0.01)	0.95 (0.01)
Ibc	F1	0.16 (0.03)	0.13 (0.05)	0.27 (0.04)	0.31 (0.04)	0.29 (0.03)	0.36 (0.04)
	Purity	0.12 (0.03)	0.09 (0.04)	0.19 (0.04)	0.24 (0.04)	0.22 (0.03)	0.32 (0.04)
	Completeness	0.26 (0.07)	0.29 (0.19)	0.47 (0.06)	0.47 (0.08)	0.42 (0.08)	0.42 (0.06)
Average	F1	0.36 (0.01)	0.45 (0.03)	0.46 (0.01)	0.63 (0.02)	0.49 (0.02)	0.64 (0.02)
	Purity	0.32 (0.01)	0.41 (0.03)	0.43 (0.02)	0.61 (0.02)	0.46 (0.02)	0.64 (0.02)
	Completeness	0.40 (0.02)	0.51 (0.06)	0.50 (0.03)	0.66 (0.02)	0.53 (0.03)	0.64 (0.02)
Weighted Avg.	F1	0.60 (0.02)	0.65 (0.03)	0.76 (0.01)	0.85 (0.02)	0.79 (0.01)	0.85 (0.01)
	Purity	0.69 (0.01)	0.75 (0.01)	0.79 (0.01)	0.86 (0.01)	0.81 (0.01)	0.85 (0.01)
	Completeness	0.53 (0.03)	0.58 (0.05)	0.73 (0.02)	0.84 (0.03)	0.77 (0.01)	0.85 (0.01)
Three-way Classification							
II	F1	0.44 (0.02)	0.48 (0.04)	0.66 (0.02)	0.75 (0.04)	0.71 (0.02)	0.75 (0.02)
	Purity	0.36 (0.02)	0.37 (0.04)	0.65 (0.03)	0.79 (0.07)	0.72 (0.02)	0.79 (0.03)
	Completeness	0.58 (0.03)	0.70 (0.10)	0.68 (0.04)	0.72 (0.03)	0.70 (0.03)	0.72 (0.02)
Ia	F1	0.68 (0.03)	0.70 (0.05)	0.88 (0.01)	0.94 (0.02)	0.91 (0.01)	0.94 (0.01)
	Purity	0.88 (0.02)	0.92 (0.01)	0.93 (0.01)	0.95 (0.01)	0.94 (0.01)	0.93 (0.01)
	Completeness	0.56 (0.04)	0.57 (0.06)	0.83 (0.02)	0.92 (0.04)	0.88 (0.01)	0.95 (0.01)
SE	F1	0.20 (0.02)	0.31 (0.06)	0.30 (0.03)	0.48 (0.04)	0.38 (0.03)	0.52 (0.03)
	Purity	0.14 (0.02)	0.22 (0.06)	0.22 (0.03)	0.38 (0.04)	0.29 (0.03)	0.47 (0.04)
	Completeness	0.41 (0.06)	0.51 (0.10)	0.47 (0.05)	0.65 (0.05)	0.53 (0.06)	0.59 (0.05)
Average	F1	0.49 (0.01)	0.54 (0.03)	0.63 (0.01)	0.73 (0.02)	0.68 (0.01)	0.74 (0.01)
	Purity	0.46 (0.01)	0.50 (0.02)	0.60 (0.01)	0.71 (0.03)	0.65 (0.01)	0.73 (0.02)
	Completeness	0.52 (0.03)	0.59 (0.05)	0.66 (0.02)	0.76 (0.03)	0.70 (0.02)	0.75 (0.02)
Weighted Avg.	F1	0.62 (0.02)	0.66 (0.03)	0.80 (0.01)	0.87 (0.02)	0.83 (0.01)	0.87 (0.01)
	Purity	0.72 (0.01)	0.75 (0.01)	0.82 (0.01)	0.88 (0.02)	0.85 (0.01)	0.87 (0.01)
	Completeness	0.55 (0.03)	0.59 (0.05)	0.78 (0.02)	0.86 (0.03)	0.82 (0.01)	0.88 (0.01)
Binary Classification							
Ia	F1	0.54 (0.02)	0.62 (0.03)	0.77 (0.02)	0.85 (0.03)	0.81 (0.01)	0.85 (0.02)
	Purity	0.44 (0.02)	0.49 (0.03)	0.73 (0.03)	0.85 (0.06)	0.78 (0.01)	0.88 (0.02)
	Completeness	0.68 (0.04)	0.84 (0.04)	0.82 (0.02)	0.86 (0.02)	0.83 (0.01)	0.82 (0.02)
CC	F1	0.74 (0.02)	0.76 (0.04)	0.90 (0.01)	0.94 (0.02)	0.92 (0.01)	0.94 (0.01)
	Purity	0.84 (0.01)	0.91 (0.01)	0.93 (0.01)	0.94 (0.01)	0.93 (0.01)	0.93 (0.01)
	Completeness	0.65 (0.04)	0.65 (0.05)	0.87 (0.02)	0.94 (0.04)	0.91 (0.01)	0.96 (0.01)
Average	F1	0.66 (0.01)	0.72 (0.02)	0.84 (0.01)	0.90 (0.02)	0.86 (0.01)	0.90 (0.01)
	Purity	0.64 (0.01)	0.70 (0.02)	0.83 (0.01)	0.90 (0.03)	0.86 (0.01)	0.91 (0.01)
	Completeness	0.67 (0.03)	0.74 (0.03)	0.85 (0.01)	0.90 (0.02)	0.87 (0.01)	0.89 (0.01)
Weighted Avg.	F1	0.69 (0.02)	0.74 (0.02)	0.86 (0.01)	0.92 (0.02)	0.89 (0.01)	0.92 (0.01)
	Purity	0.73 (0.01)	0.79 (0.01)	0.87 (0.01)	0.92 (0.02)	0.89 (0.01)	0.92 (0.01)
	Completeness	0.66 (0.03)	0.70 (0.04)	0.86 (0.01)	0.92 (0.03)	0.88 (0.01)	0.92 (0.01)

Note. Classification performance (quantified by accuracy, purity, and completeness) using a five-way, three-way, and two-way split. Optimal feature sets are bolded for each category.

fraction of catastrophic outlier redshift estimates will be $\simeq 0.1$ in the first half of LSST (M. L. Graham et al. 2018). Transients which are primarily found in intrinsically low-luminosity galaxies (e.g., SLSNe) will have particularly unreliable redshift information. As a baseline comparison, we will contrast our

results with both the galaxy-only classifier and a classifier that relies solely on light-curve features. Our results can be primarily visualized by the confusion matrices in Figure 5.

In the five-way classification task, our “baseline” classifier with only light-curve information achieves an average F1-score

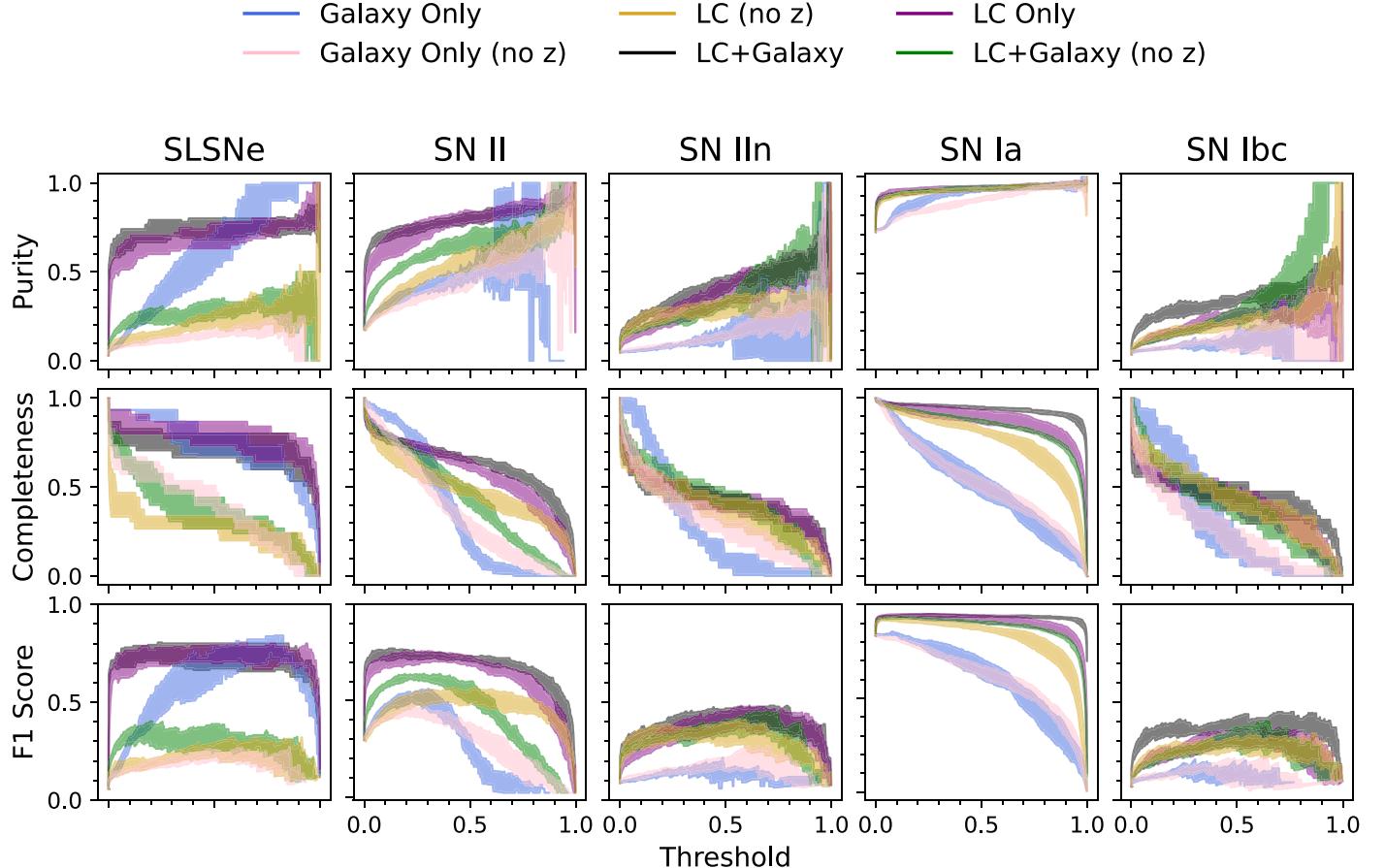


Figure 3. Purity (top row), completeness (middle row), and F1-score (bottom row) achieved with the various feature sets and confidence thresholds as a function of SN type (columns). Shaded regions represent 1σ uncertainties, computed using the same classifier with different random model initializations during training.

of 0.46. This is only marginally improved with the inclusion of galaxy information, increasing to an average F1-score of 0.49. Although the purity and completeness of all classes improve with the addition of galaxy information, although most are not statistically significant ($>1\sigma$). Notably, the purity (0.20–0.25), completeness (0.27–0.33), and F1-score (0.23–0.29) of SLSNe substantially improve, with the F1-score having a statistically significant increase. This is not surprising, as SLSNe are known to prefer unusually low-mass galaxies. The contextual information of their hosts is therefore likely a useful feature. Imposing a higher confidence threshold ($p > 0.8$) notably improves the purity of Type II SNe (from 0.62 to 0.70), Type IIn SNe (from 0.29 to 0.52), and Type Ibc SNe (from 0.22 to 0.50) when including galaxy information. These improvements are not seen for Type IIn and Type Ibc SNe when excluding galaxy information.

In both the three-way and binary classification tasks, we see even greater performance improvement. In the three-way task, including galaxy information improves our average F1-score to 0.65, compared to the light-curve-only baseline of 0.60; this is a $>3\sigma$ performance improvement. We again see improvement in purity, completeness, and F1-scores for all classes. This improvement is substantial for both Type II and SLSNe, with the F1-score improving from 0.66 to 0.71 and from 0.30 to 0.38, respectively. For the binary classification, the average F1-score improves from 0.84 to 0.86 ($>1\sigma$), with both Type Ia purity and CC completeness most notably improving (both by $>1\sigma$).

Finally, we explore how galaxy information impacts classification performance when redshift is known. Across all three tasks (five-way, three-way, and binary classifications), there is no statistical difference of average F1-scores between classifier with and without galaxy information (see Figure 6). The only statistically significant difference ($>1\sigma$) is an improvement on the purity of Type Ibc SNe (from 0.24 to 0.32) when including galaxy information, although this comes with a decrease in completeness (from 0.47 to 0.42, within 1σ uncertainties). We note that including galaxy information leads to higher SN Ia completeness (0.95 versus 0.98; a $>2\sigma$ result). As expected, a high-confidence cutoff ($p > 0.8$), improves accuracy, completeness, and purity across all classification tasks with or without galaxy information. To understand if the relative unimportance of host-galaxy features is due to the imbalance of host versus light-curve features, we train a simple random forest to perform a five-way classification, which allows us to calculate a relative “importance” score for all features. We find that no galaxy features appear in the top 10 important features, and five appear in the upper half of all features. This is in contrast to the importance of features calculated without redshift information. There, five of the top 10 important features come from host galaxies.

In short, contextual host-galaxy information improves classification performance when redshift is not known. However, when redshift is known, galaxy information does not significantly improve classification performance. As a caveat, host-galaxy information does improve classification performance for Type Ia SNe (with or without redshift

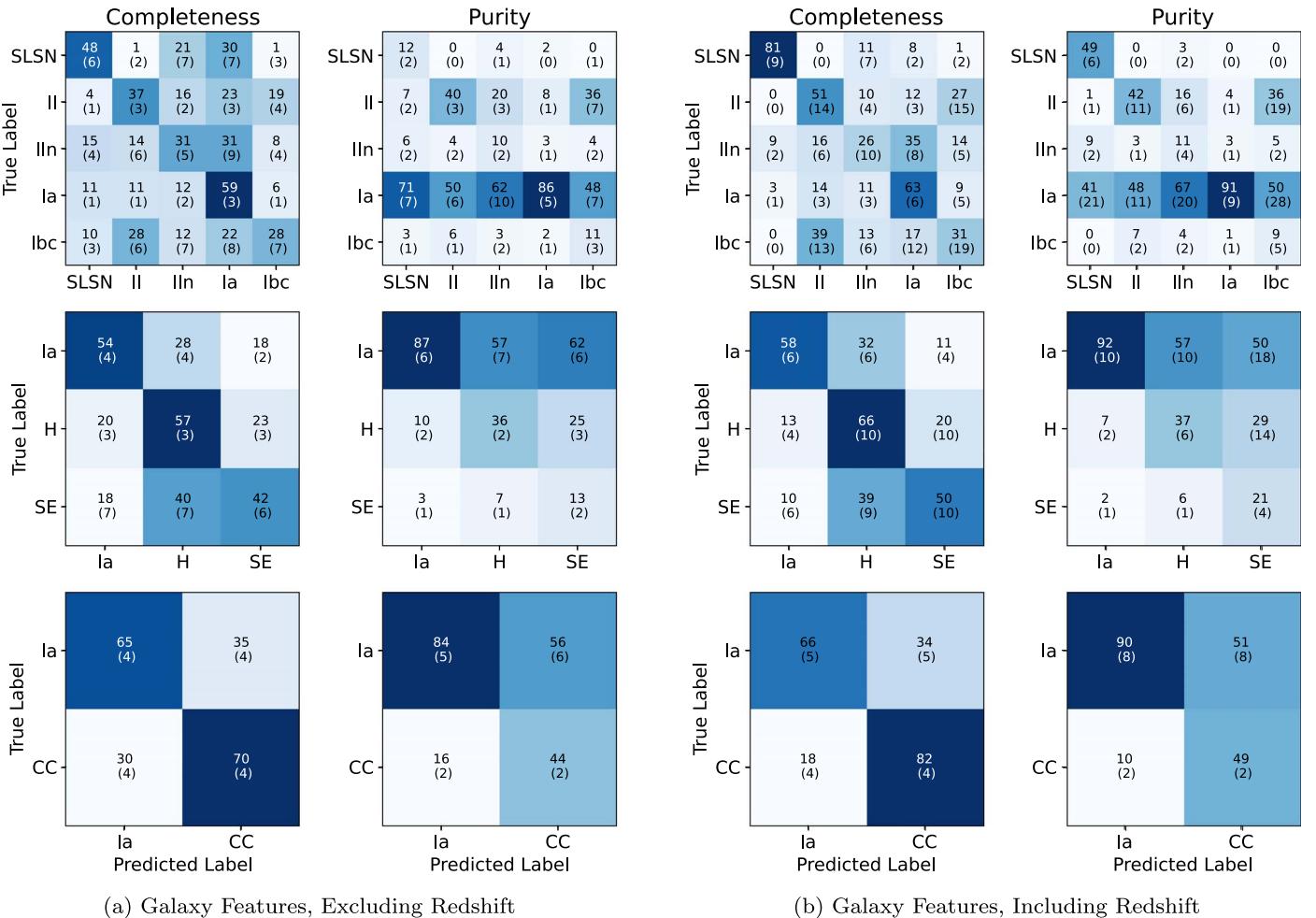


Figure 4. Confusion matrices for the classifier using only host galaxy with (left) and without (right) redshift information. Numbers are overall percent, while uncertainties are given in parentheses (e.g., the completeness of SLSNe in the five-way classifier is 0.48 ± 0.06). The classifier reaches state-of-the-art results for the Type Ia vs. CC classification, but fails to achieve high accuracy in the five-way class split.

information) when a high-confidence threshold is used. These findings hold true for complete light curves, but may change in the case where partial light-curve information is known. For example, A. Gagliano et al. (2023) found that host information led to a $\simeq 2\sigma$ increase in binary classification accuracy when utilizing redshift information.

5. Classifications of the PS1-MDS Sample

We use our redshift-independent classifier, which uses both host and light-curve information, to classify the full set of 4407 SN-like transients from the PS1-MDS originally presented in V. A. Villar et al. (2020) and G. Hosseinzadeh et al. (2020). This sample includes all objects that are not spectroscopically classified and not identified as variables or otherwise “bad” objects. The full classifications are provided in Table 2.

The class fractions, as derived by our new classifier, are shown in Figure 7. Compared to the spectroscopic data set, we find an underrepresentation of the two majority classes (Type Ia and Type II SNe) in our photometric classifications, and an overrepresentation of the minority classes (SLSNe and Type Ibc). G. Hosseinzadeh et al. (2020) showed that the expected number of misclassifications can be used to correct the expected class breakdown (by dotting the purity matrix with the final classifications). When we apply this correction to our

final class breakdown, we do recover a five-way class breakdown similar to the spectroscopic sample, lending credence to the idea that we correctly capture the biases of our imperfect classifier, even when applied to a new test set.

We more quantitatively compare our results to the SNe photometrically classified in V. A. Villar et al. (2020) and G. Hosseinzadeh et al. (2020) by analyzing the agreement (A) between these three classifiers. For a given SN class, the agreement between two classifiers is the number of events that have been identified as the same class by both classifiers divided by the size of the class in the older classifier. G. Hosseinzadeh et al. (2020) showed that the agreement between two classifiers, assuming independent biases, can be calculated as

$$A = P^T C \quad (12)$$

where P is the purity matrix of the new classifier and C is the completeness matrix of the old classifiers. Using this, the expected agreement for each class and either classifier is shown in Table 3. We will note that the definition of agreement listed above assumes that the two classifiers are independent, which is not true in our case, as both SuperRAENN and our new classifier use the same feature set. This may lead to more agreement than expected between the classifiers. We also note

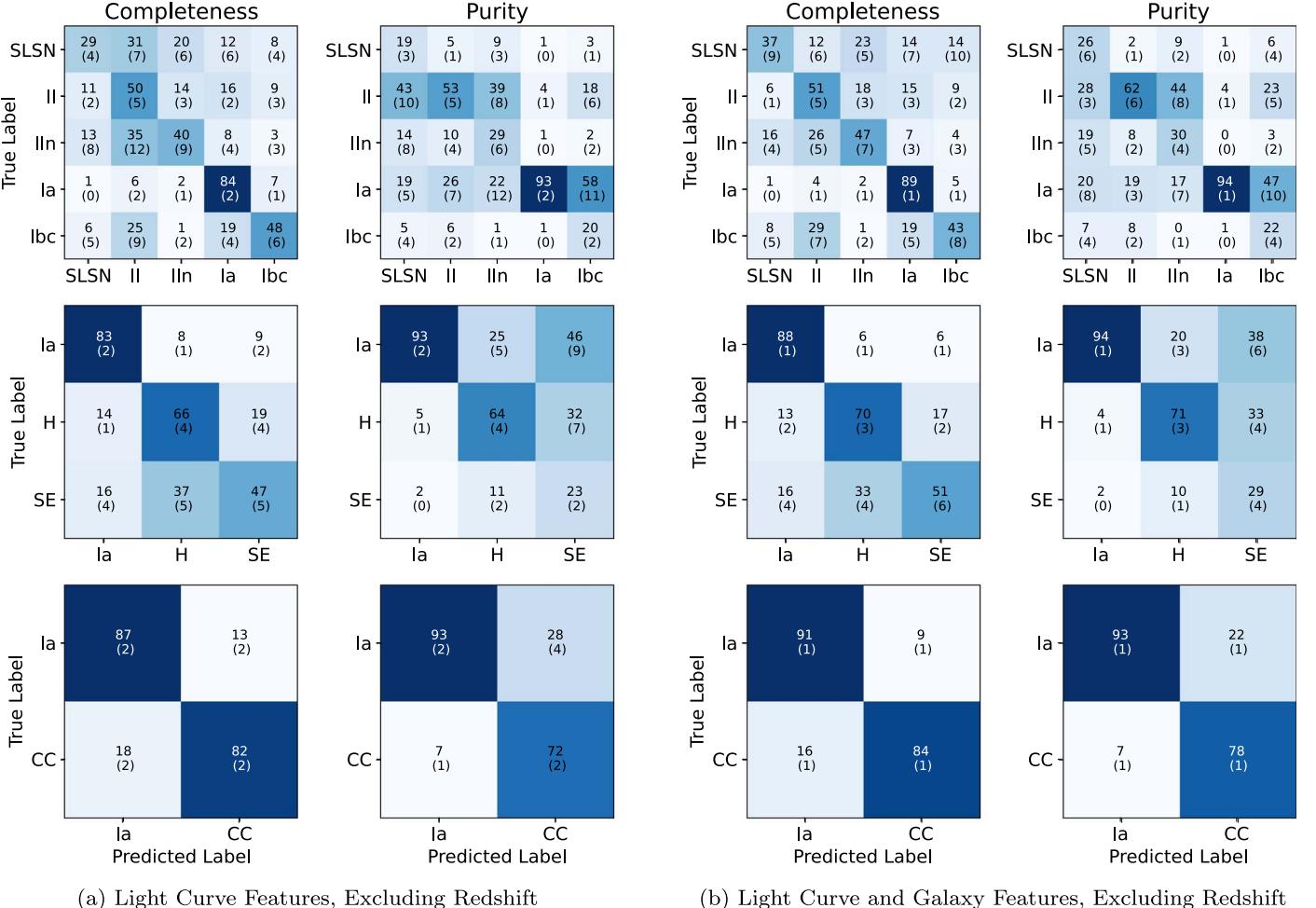


Figure 5. Confusion matrices for the classifier using solely light-curve features (left) and a combination of galaxy and light-curve features (right), both excluding redshift information. Inclusion of galaxy features does not improve classification for any one class to a statistically significant degree, and the overall performance only moderately increases.

that both SuperRAENN and SuperPhot utilize redshift information, which is not used in our new classifier.

Averaged across the five SN classes, we find 67% agreement between the new classifications and those from SuperRAENN, and 59% agreement between our new classifier and SuperPhot. For each SN class, the agreement between classifiers ranges from $\simeq 5\%$ to 90%. Our classifier shows roughly expected agreement for both Type Ia and Type II classes. Type Ia SNe, in particular, have strong agreement between the three classifiers (80%–85%), giving high confidence in the purity of our Type Ia sample. Type Ibc SNe have notably higher agreement ($\simeq 20\%$) than expected. Interestingly, this is similar to what was found in G. Hosseinzadeh et al. (2020) when comparing SuperPhot to SuperRAENN. Type IIn and SLSNe, on the other hand, have much smaller agreement than expected by a factor of 2–4; however, these two classes are commonly identified as the other.

We draw further attention to the rare class of SLSNe. Most objects classified by SuperRAENN or SuperPhot as SLSNe were classified as Type Ia SNe by the new classifier. Interestingly, of the $\simeq 150$ SLSNe identified by the three classifiers, only one object is identified as a SLSN by all three —PSc010186. Examination of the light curve and a crossmatch with known active galactic nuclei suggest that object is a $z \simeq 1$ active galactic nucleus undergoing regular, long-term variation

(B. Hsu et al. 2022). No objects are classified as SNe by our new classifier and one of the two original classifiers. We also note two other events presented in B. Hsu et al. (2022) that are not classified as SLSNe by our new classifier: PSc000036 (classified as a Type Ia) and PSc000553a (classified as a Type IIn). The former is the highest-redshift ($z = 2.026$) and brightest ($M_g = -24$) object in the sample of B. Hsu et al. (2022), which likely led to the initial classification. We take these results as an important warning: without redshift information, our photometric classifiers seemingly fail to produce consistent and pure samples of SLSNe. Furthermore, accurate AGN classifiers may be particularly important in isolating SLSNe photometrically.

Finally, we visualize our agreement in Figure 8. In general, Type Ia SNe tend to dominate the sample of objects for which both our new classifier and V. A. Villar et al. (2020) confidently agree. It is reassuring that all disagreements occur with the new confidence level < 0.5 , and are evenly distributed across the classes. However, a small sample of events have a highly confident classification from V. A. Villar et al. (2020) despite disagreement with the new classifier. In particular, 14 events have a confidence score of < 0.05 in this work and > 0.95 for the same class. Of these, six were originally classified as SLSNe and are now labeled as Type Ia SNe; one of these is the aforementioned PSc310006. An additional six

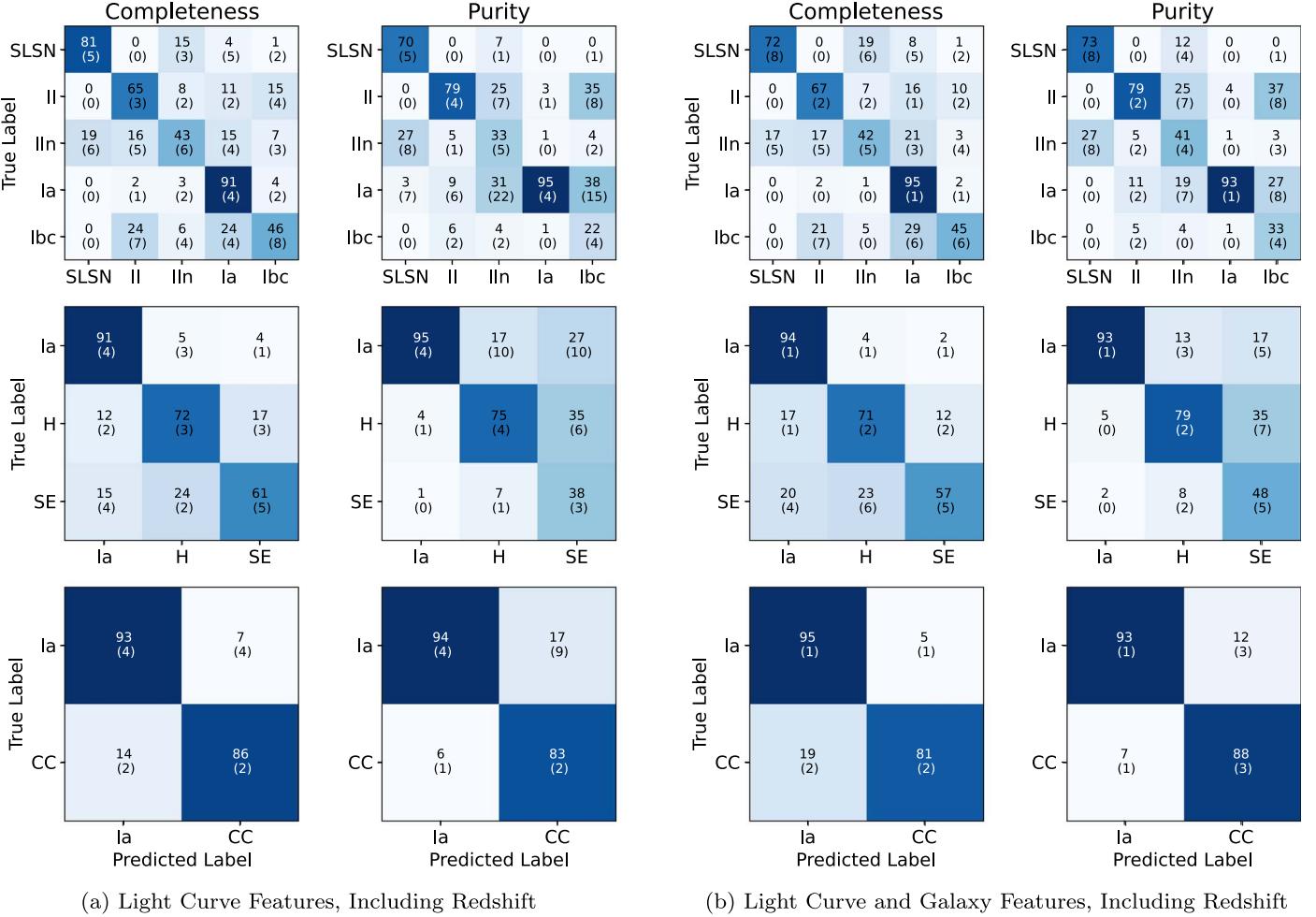


Figure 6. Confusion matrices for the classifier using redshift information, trained using only light-curve features (left) and with both light-curve and galaxy features (right). Including host information provides similar overall classification accuracy.

Table 2
Classification Probabilities of All SN-like Objects in PS1-MDS

Name	$p(\text{Ia})$	$p(\text{CC})$	$p(\text{H-rich})$	$p(\text{H-poor})$	$p(\text{SLSN})$	$p(\text{II})$	$p(\text{IIn})$	$p(\text{Ibc})$
PS000012	$1.00 \pm < 0.01$	$0.00 \pm < 0.01$	$0.00 \pm < 0.01$	$0.00 \pm < 0.01$	$0.00 \pm < 0.01$	$0.00 \pm < 0.01$	$0.00 \pm < 0.01$	$0.00 \pm < 0.01$
PS000013	0.98 ± 0.02	0.02 ± 0.02	0.01 ± 0.01	0.01 ± 0.01	0.00 ± 0.00	0.01 ± 0.01	0.00 ± 0.01	$0.00 \pm < 0.01$
PS000015	$1.00 \pm < 0.01$	$0.00 \pm < 0.01$	$0.00 \pm < 0.01$	$0.00 \pm < 0.01$	$0.00 \pm < 0.01$	$0.00 \pm < 0.01$	$0.00 \pm < 0.01$	$0.00 \pm < 0.01$
PS00246	0.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.67 ± 0.18	0.33 ± 0.18
PS00248	0.99 ± 0.01	0.01 ± 0.01	0.01 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.01	0.00 ± 0.00	0.00 ± 0.00
PS00260	0.85 ± 0.08	0.15 ± 0.08	0.14 ± 0.09	0.00 ± 0.01	0.00 ± 0.00	0.13 ± 0.07	0.01 ± 0.02	0.00 ± 0.01
PS00263	0.00 ± 0.00	1.00 ± 0.00	0.51 ± 0.42	0.49 ± 0.42	0.44 ± 0.42	0.08 ± 0.09	0.43 ± 0.38	0.04 ± 0.05

(This table is available in its entirety in machine-readable form in the [online article](#).)

were originally Type Ia SNe, and are now labeled as CCSNe (Type II and Ibc).

A key takeaway of this analysis is that expected classification accuracies and “agreement” between photometric classifiers are highly variable, and any population-level studies of photometrically classified SNe should take care to understand underlying biases and misclassifications in the observed SNe population.

6. Conclusions

We have presented an analysis on the impact of galaxy properties on the photometric classification of SNe, focusing on

CCSN subtypes. We present our key findings and products below:

1. In corroboration with previous results, we find that the observable properties of host galaxies visibly correlate with SN type. We are able to successfully distinguish CCSNe from Type Ia SNe with $\simeq 70\%$ accuracy, matching similar studies in the literature.
2. We are able to produce relatively pure ($> 90\%$) samples of Type Ia SNe using host-galaxy information alone, with or without redshift information. We are able to produce reasonably pure ($> 70\%$) samples of SLSNe and Type II SNe when using host-galaxy and redshift information.

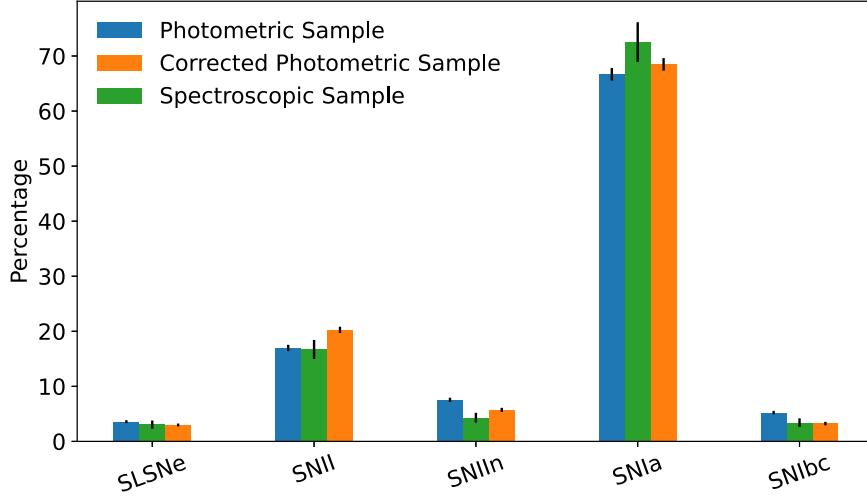


Figure 7. Breakdown of SN subclasses in our new photometrically classified set of SNe without redshift information. Compared to the spectroscopic sample (green), our classifier (blue) predicts a smaller fraction of Type Ia and Type II SNe and a larger fraction of SLSNe and Type Ibc SNe. However, using our purity matrix (see text), we can correct our classification breakdowns accounting for expected misidentifications (orange). Note that error bars are shown in black and calculated using Poisson uncertainties. We see that, following this correction, our class breakdowns are largely consistent with the spectroscopic sample. This indicates that although our classifier is biased, it is biased in a way which can be understood via our training set and can likely be carefully corrected for population-level studies.

Table 3
Expected and Actual Agreements between Classifiers

Class	SuperRAENN (Expected)	SuperRAENN (Actual)	SuperPhot (Expected)	SuperPhot (Actual)
SLSNe	0.14	0.06	0.16	0.04
SN II	0.48	0.43	0.52	0.41
SN IIn	0.15	0.08	0.13	0.09
SN Ia	0.85	0.88	0.78	0.85
SN Ibc	0.07	0.24	0.07	0.19

3. We present the first application of the weighted hierarchical cross-entropy score (see V. A. Villar et al. 2023), which better accounts for the hierarchical taxonomy of SN classification.
4. Given redshift, we find that contextual information—i.e., the host-galaxy observable properties—does not necessarily aid in SN classification when a full light curve is available. Host-galaxy information is most helpful in improving the classification accuracy of Type Ia and Type II SNe, but does not necessarily increase the accuracy of classification for other subtypes.
5. Given no redshift information, host-galaxy features consistently improve the classification accuracy of SNe across all subtypes. Galaxy information can greatly increase the completeness of Type Ia samples.
6. Finally, we present an updated classification set for the PS1-MDS set of SN-like transients without redshift information. We explore the agreement between these classifications and those originally presented in G. Hosseinzadeh et al. (2020) and V. A. Villar et al. (2020), finding strong agreement in some classes (Type Ia, Type Ibc) and weaker than expected agreement in others (most notably, SLSNe).

In the era of LSST, thousands of new SNe are expected to be discovered nightly. Our results show that host galaxy information alone can distinguish between core-collapse and thermonuclear SNe, even in the absence of redshift

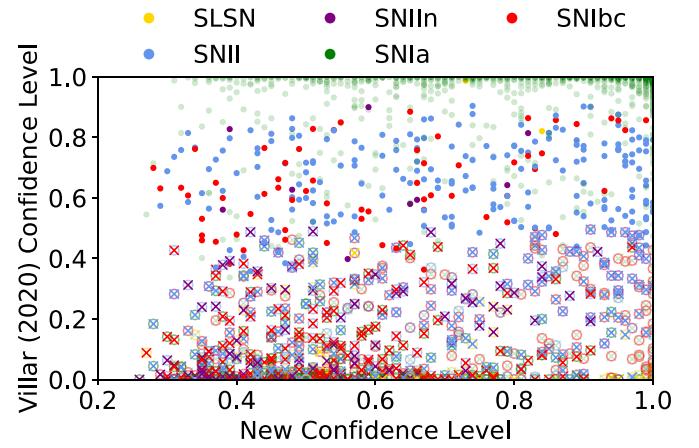


Figure 8. Comparison of SN classifications from V. A. Villar et al. (2020) and this work. On the abscissa, we plot the confidence level of the V. A. Villar et al. (2018) classifier, while the ordinate shows the confidence of the new classifier for the same class. Objects plotted as a point have the same classification with both methods; objects plotted as an “x” have different classifications with either method. The color of the central point (“x” or “o”) identifies the label from this paper. In the case of mismatches (the “x” objects), the colored circle represents the label from V. A. Villar et al. (2020).

information. However, accurate photometric redshift estimates from host galaxies will likely greatly increase our ability to rapidly classify further subtypes (particularly Type Ia and SLSNe). Given the expected on-sky density of galaxies ($\sim 10^4$ per square degree), reliable host-galaxy association techniques, such as GHOST, are essential in utilizing this contextual information.

Finally, the variation on agreement between classifiers explored here emphasizes the need for methodologies to calibrate classification confidence and build reliable populations of photometrically classified SNe. SLSNe stand out as particularly challenging. Here, we find just a single object labeled as a SLSN by all three classifiers out of dozens of events. Future work should investigate the use of probabilistic labels into population-level inference, e.g., similar to M. Kunz et al. (2007).

Acknowledgments

V.A.V. acknowledges support by the NSF through grant AST-2108676. This work is also supported by the NSF through PHY-2019786 (the NSF AI Institute for Artificial Intelligence and Fundamental Interactions). Some of the computations in this paper were run on the Cannon cluster supported by the FAS Division of Science, Research Computing Group at Harvard University. The Pan-STARRS1 Surveys (PS1) and the PS1 public science archive have been made possible through contributions by the Institute for Astronomy, the University of Hawaii, the Pan-STARRS Project Office, the Max-Planck Society and its participating institutes, the Max Planck Institute for Astronomy, Heidelberg, and the Max Planck Institute for Extraterrestrial Physics, Garching, Johns Hopkins University, Durham University, the University of Edinburgh, the Queen’s University Belfast, the Harvard-Smithsonian Center for Astrophysics, the Las Cumbres Observatory Global Telescope Network Incorporated, the National Central University of Taiwan, the Space Telescope Science Institute, the National Aeronautics and Space Administration under grant No. NNX08AR22G issued through the Planetary Science Division of the NASA Science Mission Directorate, the National Science Foundation grant No. AST-1238877, the University of Maryland, Eotvos Lorand University (ELTE), the Los Alamos National Laboratory, and the Gordon and Betty Moore Foundation.

Software: Astropy (T. P. Robitaille et al. 2013; Astropy Collaboration et al. 2018, 2022), FLEET (S. Gomez et al. 2020), Matplotlib (J. D. Hunter 2007), NumPy (C. R. Harris et al. 2020), Superphot (G. Hosseinzadeh et al. 2020), SuperRAENN (V. A. Villar et al. 2020), Pytorch (A. Paszke et al. 2019).

ORCID iDs

V. Ashley Villar  <https://orcid.org/0000-0002-5814-4061>
 Sebastian Gomez  <https://orcid.org/0000-0001-6395-6702>
 Edo Berger  <https://orcid.org/0000-0002-9392-9681>
 Alex Gagliano  <https://orcid.org/0000-0003-4906-8447>

References

Anderson, J. P., Habergham, S. M., James, P. A., & Hamuy, M. 2012, *MNRAS*, 424, 1372
 Astropy Collaboration, Price-Whelan, A. M., Lim, P. L., et al. 2022, *ApJ*, 935, 167
 Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, *AJ*, 156, 123
 Baldeschi, A., Miller, A., Stroh, M., Margutti, R., & Coppejans, D. L. 2020, *ApJ*, 902, 60
 Berger, E. 2010, *ApJ*, 722, 1946

Bertinetto, L., Mueller, R., Tertikas, K., Samangooei, S., & Lord, N. A. 2020, in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (Piscataway, NJ: IEEE) 12506
 Bloom, J. S., Kulkarni, S. R., & Djorgovski, S. G. 2002, *AJ*, 123, 1111
 Boone, K. 2021, *AJ*, 162, 275
 Bruch, R. J., Gal-Yam, A., Schulze, S., et al. 2021, *ApJ*, 912, 46
 Carrasco-Davis, R., Reyes, E., Valenzuela, C., et al. 2021, *AJ*, 162, 231
 Childress, M., Aldering, G., Antilogus, P., et al. 2013, *ApJ*, 770, 107
 Dong, Y., Milisavljevic, D., Leja, J., et al. 2022, *ApJ*, 927, 199
 Foley, R. J., & Mandel, K. 2013, *ApJ*, 778, 167
 Gagliano, A., Contardo, G., Mackey, D. F., Malz, A. I., & Aleo, P. D. 2023, *ApJ*, 954, 6
 Gagliano, A., Narayan, G., Engel, A., et al. 2021, *ApJ*, 908, 170
 Gal-Yam, A. 2017, in Handbook of Supernovae, ed. A. Alsabti & P. Murdin (Cham: Springer), 195
 Gomez, S., Berger, E., Blanchard, P. K., et al. 2020, *ApJ*, 904, 74
 Gomez, S., Berger, E., Blanchard, P. K., et al. 2023, *ApJ*, 949, 114
 Graham, M. L., Connolly, A. J., Ivezić, Z., et al. 2018, *AJ*, 155, 1
 Habergham, S. M., James, P. A., & Anderson, J. P. 2012, *MNRAS*, 424, 2841
 Hakobyan, A., Adibekyan, V. Z., Aramyan, L., et al. 2012, *A&A*, 544, A81
 Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, *Natur*, 585, 357
 Hosseinzadeh, G., Dauphin, F., Villar, V. A., et al. 2020, *ApJ*, 905, 93
 Hsu, B., Hosseinzadeh, G., Villar, V. A., & Berger, E. 2022, *ApJ*, 937, 13
 Hudelot, P., Cuillandre, J.-C., Withington, K., et al. 2012, *yCat*, II/317
 Hunter, J. D. 2007, *CSE*, 9, 90
 Jacobson-Galán, W., Dessart, L., Davis, K., et al. 2024, *ApJ*, 970, 189
 Kasliwal, M. M., Kulkarni, S. R., Gal-Yam, A., et al. 2012, *ApJ*, 755, 161
 Kelly, P. L., & Kirshner, R. P. 2012, *ApJ*, 759, 107
 Kessler, R., Narayan, G., Avelino, A., et al. 2019, *PASP*, 131, 094501
 Khazov, D., Yaron, O., Gal-Yam, A., et al. 2016, *ApJ*, 818, 3
 Kingma, D. P., & Ba, J. 2014, arXiv:1412.6980
 Kisley, M., Qin, Y.-J., Zabludoff, A., Barnard, K., & Ko, C.-L. 2023, *ApJ*, 942, 29
 Kunz, M., Bassett, B. A., & Hlozek, R. A. 2007, *PhRvD*, 75, 103508
 Leaman, J., Li, W., Chornock, R., & Filippenko, A. V. 2011, *MNRAS*, 412, 1419
 Möller, A., & de Boissière, T. 2020, *MNRAS*, 491, 4277
 Muthukrishna, D., Narayan, G., Mandel, K. S., Biswas, R., & Hložek, R. 2019, *PASP*, 131, 118002
 Paszke, A., Gross, S., Massa, F., et al. 2019, arXiv:1912.01703
 Planck Collaboration, Aghanim, N., Akrami, Y., et al. 2020, *A&A*, 641, A6
 Qin, Y.-J., Zabludoff, A., Kisley, M., et al. 2022, *ApJS*, 259, 13
 Qu, H., & Sako, M. 2022, *AJ*, 163, 57
 Ransome, C. L., Habergham-Mawson, S. M., Darnley, M. J., James, P. A., & Percival, S. M. 2022, *MNRAS*, 513, 3564
 Robitaille, T. P., Tollerud, E. J., Greenfield, P., et al. 2013, *A&A*, 558, A33
 Sánchez-Sáez, P., Reyes, I., Valenzuela, C., et al. 2021, *AJ*, 161, 141
 Schulze, S., Yaron, O., Sollerman, J., et al. 2021, *ApJS*, 255, 29
 Swann, E., Sullivan, M., Carrick, J., et al. 2019, *Msngr*, 175, 58
 Villar, V. A. 2024, VTDA-Group/hxe-for-tda: AAS Release, v1.0.0, Zenodo, doi:10.5281/zenodo.13386740
 Villar, V. A., Berger, E., Miller, G., et al. 2019, *ApJ*, 884, 83
 Villar, V. A., de Soto, K., & Gagliano, A. 2023, arXiv:2312.02266
 Villar, V. A., Hosseinzadeh, G., Berger, E., et al. 2020, *ApJ*, 905, 94
 Villar, V. A., Nicholl, M., & Berger, E. 2018, *ApJ*, 869, 166