A Classifier-Based Approach to Multi-Class Anomaly Detection Applied to Astronomical Time-Series

Rithwik Gupta [1 2] **Daniel Muthukrishna** [1] **Michelle Lochner** [3 4]

Abstract

Automating anomaly detection is an open problem in many scientific fields, particularly in time-domain astronomy, where modern telescopes generate millions of alerts per night. Currently, most anomaly detection algorithms for astronomical time-series rely either on hand-crafted features or on features generated through unsupervised representation learning, coupled with standard anomaly detection algorithms. In this work, we introduce a novel approach that leverages the latent space of a neural network classifier for anomaly detection. We then propose a new method called Multi-Class Isolation Forests (MCIF), which trains separate isolation forests for each class to derive an anomaly score for an object based on its latent space representation. This approach significantly outperforms a standard isolation forest when distinct clusters exist in the latent space. Using a simulated dataset emulating the Zwicky Transient Facility (54 anomalies and 12,040 common), our anomaly detection pipeline discovered 46 ± 3 anomalies ($\sim 85\%$ recall) after following up the top 2,000 ($\sim 15\%$) ranked objects. Furthermore, our classifier-based approach outperforms or approaches the performance of other state-of-the-art anomaly detection pipelines when applied to the dataset used in Perez-Carrasco et al. (2023). Our novel method demonstrates that existing and new classifiers can be effectively repurposed for real-time anomaly detection. The code used in this work, including a Python package, is publicly available.

1. Introduction

Astronomical surveys measure light (or flux) in specific regions of the night sky. In time-domain astronomy, observations are made periodically, forming a light curve that represents the object's brightness variations over time. Most light curves exhibit minimal or gradual changes and are relatively unremarkable. However, when a significant deviation in brightness is detected with a high signal-to-noise ratio (S/N), it indicates the presence of a transient event in the observed galaxy. Transient events encompass a wide range of astrophysical phenomena, including various types of supernovae, which are explosive endings of stellar life cycles, and rare occurrences such as microlensing, where the light from a distant source is gravitationally amplified by an intervening massive object. Examples of light curves exhibiting transient events are presented in Appendix B.

With the advancement of these survey telescopes and the advent of large-scale transient surveys, we are entering a new paradigm for astronomical study. The Vera Rubin Observatory's Legacy Survey of Space and Time (LSST) is expected to observe ten million transient alerts per night (Ivezić et al., 2019). The traditional approach of manual examination of astronomical data, which has led to some of the biggest discoveries in astronomy, is no longer feasible. As a result, there is a growing need to develop methods that can automate the serendipity that has so far played a pivotal role in scientific discovery.

The literature on anomaly detection for astronomical transients presents two distinct problem definitions. Some approaches, categorized as unsupervised methods, focus on extracting anomalies from large datasets without relying on prior information (e.g. Villar et al., 2021; Webb et al., 2020; Giles & Walkowicz, 2019). Numerous differing approaches exist for unsupervised anomaly detection. Villar et al. (2021) used an unsupervised recurrent variational autoencoder to learn a representative latent space mapping of the light curves to then derive anomaly scores using an isolation forest. Webb et al. (2020) used user-defined feature extraction and then active learning to identify anomalies.

In contrast, our work, among others (e.g. Perez-Carrasco et al., 2023; Muthukrishna et al., 2022), uses previous, ei-

[1] Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA [2] Irvington High School, 41800 Blacow Rd, Fremont, CA 94538, USA [3] Department of Physics and Astronomy, University of the Western Cape, Bellville, Cape Town, 7535, South Africa [4] South African Radio Astronomy Observatory, 2 Fir Street, Black River Park, Observatory, 7925, South Africa. Correspondence to: Daniel Muthukrishna <danmuth@mit.edu>.

AI4Science Workshop of the 41st International Conference on Machine Learning. Copyright 2024 by the authors.

ther simulated or real, transients to determine whether a new light curve is anomalous. This approach is often referred to as novelty detection or supervised anomaly detection. Previous novelty detection approaches (e.g. Muthukrishna et al., 2022; Soraisam et al., 2020) are often variations of one-class classification (Schölkopf et al., 1999). One-class classifiers attempt to model a set of *normal* samples and then classify new transients as either part of that sample or as outliers. One-class methods have been shown to be effective at anomaly detection (Ruff et al., 2018a), but they do not capture the complexity of the population of known astronomical transients, that are grouped into numerous classes with intrinsically different qualities. Perez-Carrasco et al. (2023) extended the one-class classifier to multiple classes after training on features extracted from full light curve data. Their method adapts the single-class loss function to multiple classes by encouraging light curves of the same class to cluster together.

In this work, we leverage a light-curve classifier to address the one-class challenge and distinguish between the various classes of transients. Our approach demonstrates promising clustering in the feature space, the penultimate layer of the classifier, and shows a substantial level of discrimination in anomaly scores. Notably, similar feature extraction methods have shown potential in the field of astronomical image analysis (e.g. Etsebeth et al., 2023; Walmsley et al., 2022).

Once a feature space has been identified using one of the previously mentioned methods, several prior works have employed an isolation forest (Liu et al., 2008) to generate anomaly scores. While this approach has demonstrated success in previous research (e.g Villar et al., 2021; Ishida et al., 2021; Pruzhinskaya et al., 2019), it faces challenges when dealing with a complex latent space that contains multiple clusters of intrinsically different transient classes. Consequently, the application of a single isolation forest may have limitations in accurately identifying certain anomalies, as it may struggle to adequately capture the distinct properties of each cluster. Singh et al. (2022) also recognized the problem of using a single anomaly detector in a multi-class setting and introduced a method training an autoencoder for each class to then derive an anomaly score.

In response to this limitation, we propose the use of Multi-Class Isolation Forests (`MCIF`): a method that involves training a separate isolation forest for each known class and extracting the minimum score among them as the final anomaly score for a given sample. Our experimental results suggest that `MCIF` holds promise in improving anomaly detection performance for astronomical transients when there are defined clusters in the latent space.

2. Dataset

In this work, we use a collection of simulated light curves that match the observing properties of the Zwicky Transient Facility (ZTF, Bellm et al., 2018). This dataset is described in § 2 of Muthukrishna et al. (2022) and is based on the simulations developed for PLAsTiCC (Kessler et al., 2019). Each transient in the dataset has flux and flux error measurements in the g and r passbands (two different light filters) with a median cadence of roughly 3 days in each passband.

The 17 transient classes we consider in this work are SNIa, SNIa-91bg, SNIax, SNIb, NIc, SNIc-BL, SNII, SNIIb, SNIIn, SLSN-I, PISN, KNe, AGN, TDE, ILOT, CaRT, and uLens-BSR. Due to their low occurrence in nature, **KNe, ILOT, CaRT, PISN, and uLens-BSR** are considered the **anomalous classes** in this work, and all remaining classes are considered the "common" classes. Example light curves from each of these classes are illustrated in Appendix B.

To emulate the real world, where scientists do not necessarily know what anomalies they are looking for, we ensure all transients from the anomalous classes are unseen by our model until final evaluation. Further, the goal of this work is to detect anomalies in general, not specifically transients of the aforementioned anomalous classes. Hence, we do not use physical priors of any transient type to aid in detection. Finally, because anomalies are inherently rare, but our simulated dataset is relatively class balanced, we perform evaluation by down-sampling the objects in the anomalous classes to create a more realistic evaluation dataset.

3. Methods

3.1. Overview

Figure 1 summarizes our methodology. First, we train a Recurrent Neural Network (RNN) to classify the common classes of transients. Then, we remove the final layer of the trained model and use the remaining architecture as an encoder. To effectively extract anomalies from a well-represented space, it is essential to ensure that transients from similar classes cluster together. In our encoder, the latent space is directly used for light curve classifications, which should naturally lead to clustering of similar transients.

Once we have established this representation space, we must extract anomalies from it. However, when dealing with multiple clusters, a single isolation forest may struggle to capture each cluster equally (for further details, refer to Section 4.5). This challenge motivated our approach, `MCIF`, where we train an isolation forest for each class, representing a distinct cluster, and select the minimum anomaly score as the final score. This minimum score should come from the cluster to which the latent observation is closest, providing

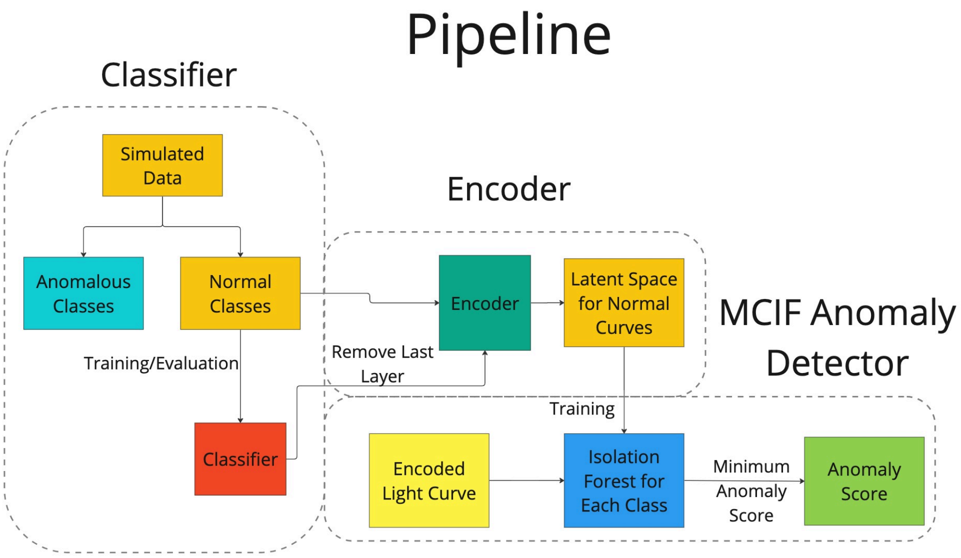


Figure 1. A visual summary of the architecture described in this work. Our approach first trains a classifier, then repurposes it as an encoder, and finally applies Multi-Class Isolation Forests (`MCIF`), proposed in this work, for anomaly detection.

the desired functionality.

3.2. Classifier

We train a DNN (Deep Neural Network) classifier that maps a matrix of multi-passband light-curve data $\boldsymbol{X}_s$ for a transient s to a $1 \times N_c$ vector of probabilities, reflecting the likelihood of the given light curve being from each of the aforementioned non-anomalous transient classes, where N_c is the number of classes.

The transient classifier utilizes a Recurrent Neural Network (RNN) with Gated Recurrent Units (GRU, Cho et al., 2014) to handle the sequential time series data. GRUs have been shown to perform better than typical Recurrent Neural Networks (RNNs), have quicker training times than LSTMs[1] (Chung et al., 2014), and have shown promise in the domain of astronomical time-series (Muthukrishna et al., 2022). The input for each transient, $\boldsymbol{X}_s$, is a $4 \times N_T$ matrix where N_T is the maximum number of timesteps for any input sample. N_T is 656 in this work, but most transients have much fewer observations. Each row of the input matrix is composed of the following vector,

$$\boldsymbol{X}_{sj} = [f_{sj}, \epsilon_{sj}, t_{sj}, \lambda_p], \tag{1}$$

where f_{sj} is the scaled flux for the jth observation of transient s, ϵ_{sj} is the corresponding scaled uncertainty, t_{sj} is the scaled time of when the measurement was taken, and λ_p is the central wavelength of the passband from which the measurement comes from.

After the recurrent layers of the DNN, we pass some contextual information into the classifier, which has been shown to be helpful for light curve classification (Foley & Mandel, 2013). In this work, we use the Milky Way extinction and the host galaxy's spectroscopic redshift as additional inputs to the network. We train our neural network for 40 epochs using the `adam` optimizer and counteract class imbalance in our dataset by using class weights inversely proportional to the frequency of the class while training. Our model takes roughly 10 minutes to train on a 16GB Tesla V100 GPU core.

One of the advantages of using a neural network-based architecture over hand-selected features is that it is a data-driven model, which should make it more sensitive to identifying out-of-distribution data. This inherent quality of neural networks makes them especially good for anomaly detection. However, the lack of interpretability of DNN models is a drawback and means that we can't discern why a certain object is marked anomalous.

3.3. Anomaly Detection

Once the classifier is trained, we remove the last layer and use the remaining architecture to map any light curve to the latent space. We define this encoder as a function $E(\boldsymbol{X}_s)$, that takes the aforementioned preprocessed light curve data, $\boldsymbol{X}_s$, and maps it to a 100-dimensional latent space $\boldsymbol{z}_s$

[1] We empirically find that there is little difference between an LSTM and GRU model, in both classification accuracy and anomaly detection.

$$\boldsymbol{z}_s = E(\boldsymbol{X}_s) \tag{2}$$

For anomaly detection, we now want to compute the anomaly score, $a_s = A(\boldsymbol{z}_s)$, where $A(\boldsymbol{z}_s)$ is a function that evaluates the anomaly score a_s for a latent observation $\boldsymbol{z}_s$. The goal of this work is to generate relatively large anomaly scores for anomalous transients and smaller anomaly scores for non-anomalous transients.

We propose a new framework where an isolation forest is trained separately on data from every class, using the minimum anomaly score from any isolation forest as the final anomaly score[2]. We call this approach Multi-Class Isolation Forests (MCIF).

We define 12 isolation forests, $I_c(\boldsymbol{z}_s)$, trained on latent space observations from the common transient class c. The final anomaly score is defined as

$$A(\boldsymbol{z}_s) = \min_{\forall c}\Big(-I_c(\boldsymbol{z}_s)\Big) \tag{3}$$

The function $I_c(\boldsymbol{z}_s)$ is positive for less anomalous transients and negative for anomalous ones, to be consistent with the sklearn implementation of Isolation Forests. We negate the scores as we prefer defining transients with higher anomaly scores to be more anomalous, but this makes no difference to the results. All isolation forests used in this work are trained with 200 estimators. The results of using a single isolation forest and the benefits of using Multi-Class Isolation Forests are explored further in Section 4.5.

4. Evaluation

4.1. Latent Space

After repurposing the classifier as an encoder, we obtain a 100-dimensional latent space. We can visualize this latent space with UMAP (McInnes et al., 2020), a manifold embedding technique, to determine if there is visible clustering[3]. In Figure 2 [left], we plot the UMAP representations of the test data. While it is difficult to examine some of the overlapping classes in this embedded space, there is clear clustering of many of the classes. In Figure 2 [right], we color all of the common classes grey and include a sample of transients from the anomalous classes. We see that the anomalous classes cluster together in the embedded space and separate from the common transients despite the model not being trained on these objects. This level of clustering suggests that our encoder may be discovering generalizable patterns within light curves, and this property may have potential use cases beyond anomaly detection in few-shot classification. It is important to note that we only use UMAP for visualisation purposes and that the latent space used for anomaly detection is obtained directly from the penultimate layer of the classifier.

4.2. Anomaly Detection

In Figure 3 [left], we plot the distribution of anomaly scores predicted by MCIF from the latent space for each class. The plot demonstrates the distinction in anomaly scores of common and anomalous transients as there is a significant skew towards larger anomaly scores for the anomalous classes. However, Calcium Rich Transients (CaRTs), despite being one of our anomalous classes, tend to have lower anomaly scores. CaRTs are notoriously difficult to photometrically classify as anomalous due to their resemblance to other common supernova classes (see Fig. 8 of Muthukrishna et al. 2019 for example).

4.3. Detection Rates in a Representative Population

The previous results do not acknowledge a key difficulty of anomaly detection: anomalies are inherently infrequent. While the frequency of anomalous transients in nature is not known, a good estimate for the expected population frequency was presented in Kessler et al. (2019) for the PLAsTiCC dataset (The PLAsTiCC team et al., 2018). Using PLAsTiCC frequencies for each class, the rate of common transients is roughly 220 times that of anomalous transients. We used this rate to randomly select a more realistic test dataset that contained 12,040 normal transients and 54 anomalies. Randomly selecting a representative sample of only 54 anomalies is subject to significant variance. Therefore, we created 50 sample datasets to perform 50-fold cross-validation. Information on the exact composition of these test sets is listed in Table 2.

For each validation set, we ranked the transients by the anomaly scores predicted by MCIF. We then selected the top 2,000 highest-scoring transients (roughly 15% of the dataset) as the candidate pool. Across 50 repeated trials, we identified 46 ± 3 out of the 54 true anomalies in our dataset (recalling $\sim 85\%$ of the anomalies). In Figure 4, we plot the fraction of anomalies recalled and the total number of anomalies recovered for thresholds up to the top 2,000 transients. MCIF recalls the majority of true anomalies among candidates having the highest anomaly scores, followed by a tapering as fewer anomalies remain.

[2] We also tested using an SVM and the distance from the cluster's center, but an Isolation Forest empirically worked the best, as is seen in similar literature.

[3] We use the umap-learn implementation in python using the hyperparameters "minimum distance" set to 0.5 and "number of neighbors" set to 500.

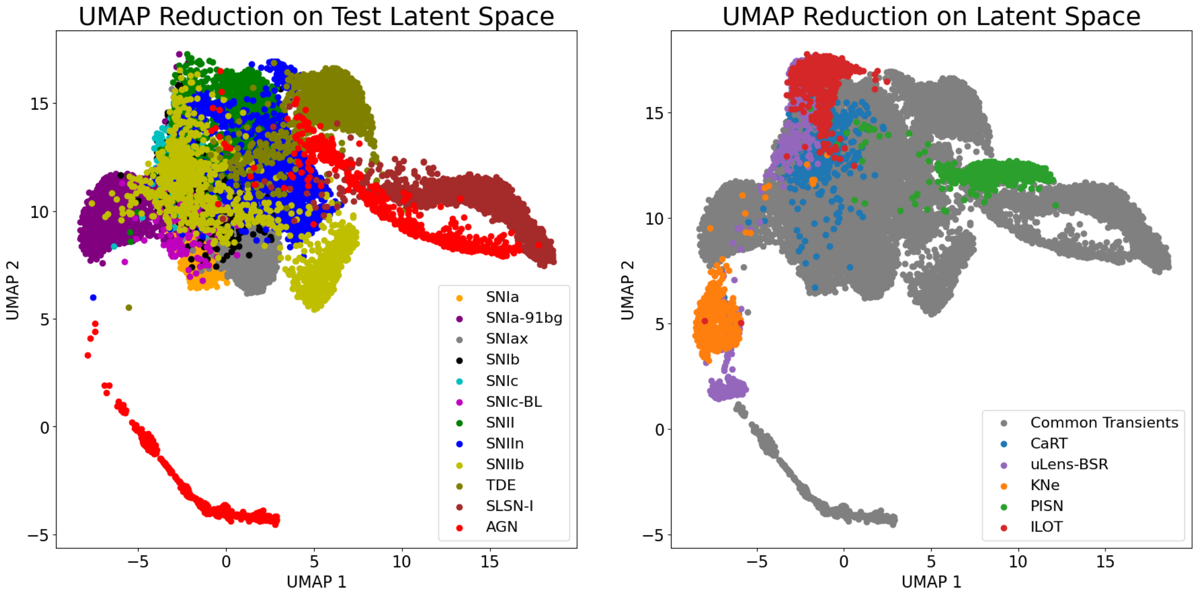


Figure 2. The UMAP reduction of the latent space derived from the test set, which includes 10% of the common transients reserved for testing the classifier [left] and randomly sampled anomalous transients from the unseen anomaly dataset [right]. Despite not being trained on this data, the learned features still exhibit clear visual structure and anomalous transients form distinct clusters separate from the common classes. It is important to note that the UMAP reduction is used only for visualization purposes, and the actual anomaly detection is performed on the nine-dimensional latent space.

	Transient				**Stochastic**					**Periodic**				
Method	SLSN	SNII	SNIa	SNIbc	AGN	Blazar	CV/Nova	QSO	YSO	CEP	DSCT	E	RRL	LPV
IForest (Liu et al., 2008)	0.640 ±0.014	0.721 ±0.021	0.428 ±0.032	0.490 ±0.038	0.573 ±0.017	0.710 ±0.009	**0.975** **±0.001**	0.468 ±0.016	**0.913** **±0.003**	0.359 ±0.007	0.295 ±0.012	0.469 ±0.021	0.549 ±0.033	**0.971** **±0.007**
OCSVM (Schölkopf et al., 1999)	0.577 ±0.014	0.587 ±0.014	0.434 ±0.021	0.492 ±0.011	0.532 ±0.008	0.443 ±0.002	0.909 ±0.001	**0.517** **±0.005**	0.792 ±0.005	0.432 ±0.004	**0.557** **±0.005**	0.555 ±0.003	0.539 ±0.004	0.943 ±0.001
AE (Rumelhart & McClelland, 1987)	**0.736** **±0.022**	**0.807** **±0.021**	0.438 ±0.015	0.537 ±0.019	**0.701** **±0.010**	**0.762** **±0.006**	**0.980** **±0.016**	0.443 ±0.004	**0.990** **±0.001**	0.564 ±0.024	0.367 ±0.015	**0.864** **±0.009**	**0.907** **±0.015**	**0.996** **±0.000**
VAE (Kingma & Welling, 2014)	0.669 ±0.015	0.690 ±0.023	0.404 ±0.018	0.522 ±0.025	0.596 ±0.007	0.597 ±0.010	0.849 ±0.028	**0.500** **±0.009**	0.795 ±0.009	0.442 ±0.010	0.417 ±0.007	0.561 ±0.007	0.451 ±0.006	0.936 ±0.007
Deep SVDD (Ruff et al., 2018b)	0.644 ±0.043	0.731 ±0.043	0.475 ±0.040	0.507 ±0.040	0.496 ±0.025	0.607 ±0.044	0.932 ±0.015	0.411 ±0.008	0.901 ±0.022	0.707 ±0.027	0.482 ±0.054	0.636 ±0.055	0.774 ±0.068	0.785 ±0.025
MCDSVDD (Perez-Carrasco et al., 2023)	**0.686** **±0.051**	**0.828** **±0.024**	**0.624** **±0.039**	**0.584** **±0.032**	**0.706** **±0.069**	0.512 ±0.113	0.770 ±0.127	0.483 ±0.080	0.854 ±0.041	**0.858** **±0.025**	**0.819** **±0.015**	**0.945** **±0.006**	**0.953** **±0.003**	0.953 ±0.008
Classifier + IForest (This work)	**0.757** **±0.047**	**0.811** **±0.017**	**0.619** **±0.073**	0.556 ±0.039	**0.715** **±0.028**	**0.720** **±0.032**	0.945 ±0.015	0.456 ±0.041	**0.977** **±0.003**	**0.766** **±0.066**	0.504 ±0.111	**0.811** **±0.038**	**0.907** **±0.026**	**0.969** **±0.016**
Classifier + MCIF (This work)	0.567 ±0.091	0.699 ±0.046	**0.536** **±0.061**	**0.560** **±0.034**	0.615 ±0.048	0.701 ±0.045	0.882 ±0.050	**0.605** **±0.051**	0.893 ±0.025	**0.875** **±0.036**	**0.742** **±0.044**	0.773 ±0.031	0.808 ±0.046	0.779 ±0.107
MCIF (This work)	0.503 ±0.018	0.668 ±0.008	0.532 ±0.007	**0.643** **±0.005**	0.614 ±0.02	**0.745** **±0.008**	**0.966** **±0.003**	0.446 ±0.007	0.907 ±0.007	0.514 ±0.013	0.433 ±0.009	0.476 ±0.021	0.447 ±0.011	0.959 ±0.004

Table 1. Performance of each model when applied to the dataset used in Perez-Carrasco et al. (2023). Each row represents a different anomaly detection algorithm and each column represents a different class being chosen as the anomalous class. The performance is evaluated using the AUROC score of detected anomalies. The top 3 metrics per class are marked in bold. The AUROC scores for the first 5 methods are taken directly from and are reported in Perez-Carrasco et al. (2023). A visual representation of this table is shown in Figure 7.

4.4. Comparison Against Other Approaches

In the field of anomaly detection in time-domain astronomy, there is no comprehensive baseline on which to evaluate different detection methods. This is largely because of the vastly differing definitions of what *anomaly detection* is, for example, the difference between unsupervised and novelty detection methods as described in Section 1. Baselining all existing anomaly detection methods is a much needed line of future work, especially as there is no consensus on which method will work best on the deluge of data that will available when LSST is running.

Despite these challenges, Perez-Carrasco et al. (2023) evaluated 5 different approaches to anomaly detection (see Table 1 for all benchmarked approaches), and we use their dataset (which was inspired by Sánchez-Sáez et al. 2021) to evaluate our classifier-based approach. In contrast to our dataset

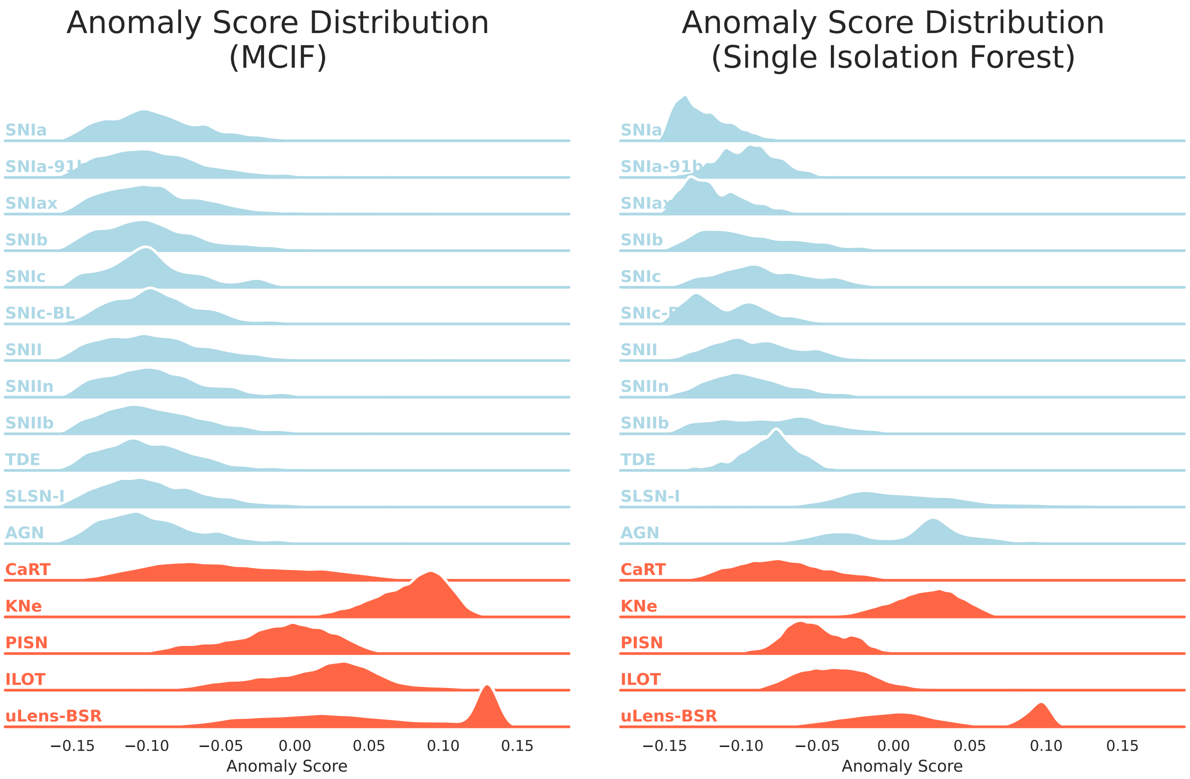


Figure 3. The distribution of anomaly scores for each class, computed using MCIF [left] or a single isolation forest [right] on the latent representations derived from full light curves. The scores are plotted using 100% of the anomalous dataset (unseen during training) and the test dataset of common classes. The anomalous classes (bottom five in red) generally show higher anomaly scores with positively skewed distributions when using MCIF, however this is less true when using a single isolation forest. The common classes and CaRTs all have low anomaly scores when using MCIF.

of raw light curve data, this dataset consists of *features* extracted from light curves. We evaluate three new techniques for anomaly detection on this dataset: using a classifier with MCIF, a classifier with just a single Isolation Forest, and MCIF on its own[4]. The dataset is split into 3 hierarchical categories with 4-5 transient classes each. Evaluation is performed separately for each class, each time counting that transient class as anomalous and the rest of its hierarchical category as common. Full evaluation is performed across 5 folds of testing data for cross-validation.

As seen in Table 1 (and visually in Figure 7), our classifier-based approach with an isolation forest is one of the top approaches for most transient classes, showing the power of using a classifier's latent space for anomaly detection. Using a classifier with MCIF also preforms promisingly, however is sometimes worse than using a classifier with a single isolation forest. This is not the case on our dataset and is discussed further in the next section.

4.5. Advantages of MCIF

To evaluate MCIF, we compare it to the performance of using a normal isolation forest to detect anomalies from the latent representation z_s of a light curve[5]. We train an isolation forest on the latent represenation of our training data using 2400 estimators (the same number used by all of the isolation forests in MCIF combined). To account for the class imbalance in our training data, we weight samples from underrepresented classes more heavily during the training of the isolation forest, using the same weighting scheme used in the classifier. The anomaly score function $A(z_s)$ is now simply the negated anomaly score output from a single isolation forest trained on all the latent representations of the training data.

As shown in Figure 3 [right], there is little distinction in the anomaly scores of most anomalous and common classes when using a single isolation forest. Surprisingly, the common classes SLSN-I and AGN are classified as relatively

[4] We can use MCIF on its own as this is a dataset of features extracted from time-series, not the raw time-series.

[5] Note that this evaluation is done on the dataset described in Section 2, not the one used for comparitive analysis in Section 4.4.

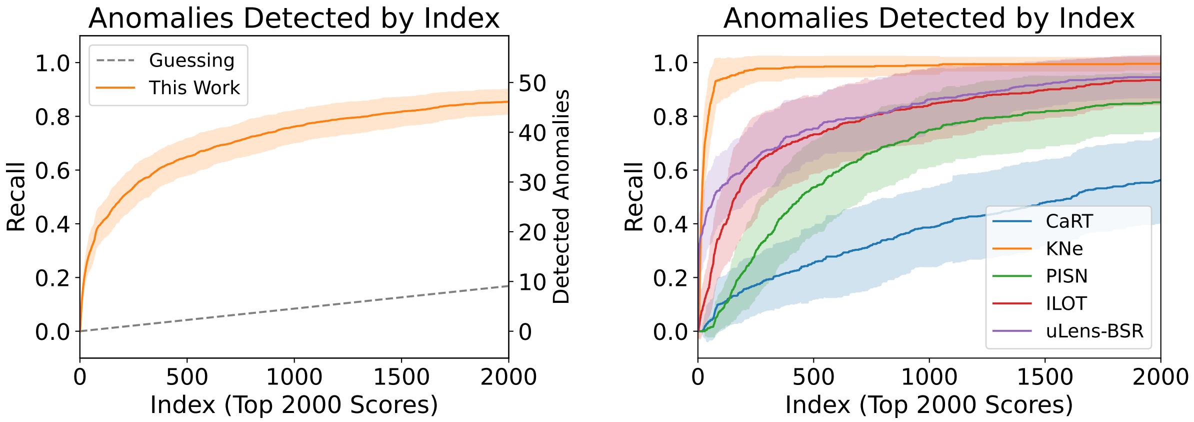


Figure 4. Anomalies detected in the 2,000 top-ranked transients by `MCIF` anomaly score index, using a test sample reflecting the estimated frequency of anomalies in nature. In the sample of 12,040 common transients and 54 anomalous transients, the model recalls 46 ± 3 ($\sim 85\%$) of the anomalies after following up the top 2,000 ranked transients. The left plot aggregates all anomalies and the right plot delineates per class. To control for the variance imposed by the small anomaly sample size, we repeat the sampling 50 times. The mean and standard deviation of detected anomalies are plotted as the solid lines and shaded regions, respectively.

more anomalous than all the other classes.

The UMAP reduction of the latent space of our classifier, as depicted in Figure 2, provides insight into this behaviour. The SLSN-I and AGN classes are located far from the main cluster formed by other classes and are nearly perfectly classified by our classifier (shown in the confusion matrix and ROC curves in Figure 9 in Appendix C). In fact, the near-perfect classification hinted at their potential to be misidentified as anomalies, suggesting that their distinct characteristics make them easily separable from other classes and, consequently, more likely to be flagged as anomalous by a single isolation forest. On the other hand, while SNIa also deviate from the central cluster in the UMAP visualization, they are among the most challenging classes to classify accurately and are the most frequently observed transient class in real surveys. Thus, they are likely a part of the central cluster in the full 100-dimensional latent space. Hence, while an isolation forest is good at detecting anomalies, it struggles to capture the structure of a latent space with numerous well-defined clusters. This drawback of using a single isolation forest could explain why other works report high anomaly scores for SLSN-I and AGN (e.g. Villar et al., 2021). Using a class-by-class (or cluster-by-cluster) anomaly detector, such as `MCIF`, can mitigate this. A direct comparison of the anomaly score distributions in Figure 3 empirically demonstrates the advantages of `MCIF` on our dataset.

Further analysis of MCIF's performance on the comparative evaluation dataset (Section 4.4) reveals that, contrary to the results shown in Figure 3, a single isolation forest generally outperforms MCIF (Table 1). Investigating the UMAP representations of the latent space for classes exhibiting this discrepancy offers insights. When SNII is considered anomalous, the latent space (Figure 5 [left]) lacks clear separation between SNIbc and SNIa, likely due to poor generalization caused by the limited number of SNIbc transients in the training set, explaining the single isolation forest's superior performance. However, for the DSCT class (Figure 5 [right]), distinct visual clusters are present, and MCIF achieves better results. These findings suggest that MCIF enhances performance when majority classes are well-separated, a characteristic seemingly inherent to the dataset rather than the classifier-based latent space identification approach, as a single isolation forest surpasses MCIF on the raw data for most classes where it also outperforms MCIF on the classifier's latent space. Future research should explore the factors influencing MCIF's effectiveness based on the separability of raw data, with the SNII case indicating a partial dependence on data quantity, as increased data improves the DNN's generalization ability.

4.6. Scaling the Latent Space

Anomaly detection presents a unique challenge in terms of evaluation, as the true anomalies are only revealed during the final testing phase. Consequently, we refrain from tuning hyperparameters for model selection and instead retrospectively analyze the effects of different hyperparameter choices, particularly the size of the latent space.

To assess the impact of latent space size on anomaly detection performance, we train multiple models with varying latent dimensions and evaluate them using the AUROC. As shown in Figure 6, increasing the latent size beyond 50 leads to significant improvements in anomaly detection performance, with diminishing returns observed after 70

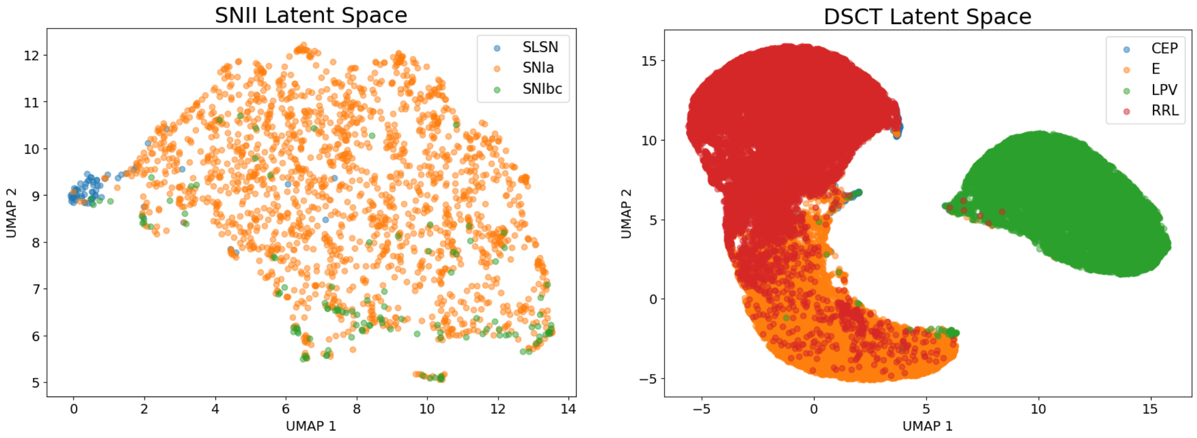


Figure 5. The UMAP reduction of the training data in the latent space for a classifier trained for detecting the class SNII [left] and DSCT [right] as anomalous using the data introduced in (Perez-Carrasco et al., 2023) and used in Section 4.4. As the UMAP only plots the training data, it includes all the classes in the respective hierarchical category (seen in Table 1) but the one set aside as anomalous.

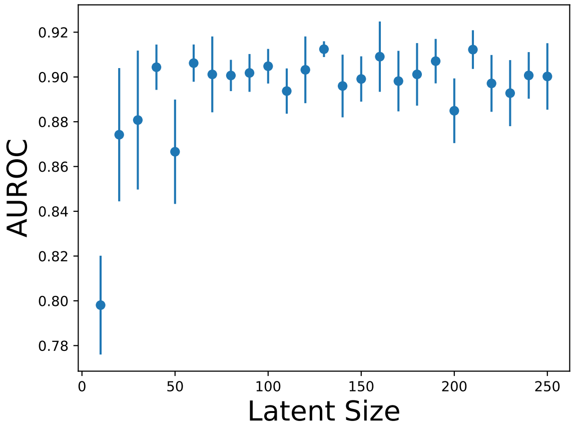


Figure 6. Anomaly detection performance (AUROC) of models trained with different latent space sizes. A significant improvement is observed when increasing the latent size up to 50 dimensions, with performance plateauing thereafter.

dimensions. Smaller models generally exhibit lower average performance and higher variance. Interestingly, we do not observe a performance drop in high-dimensional latent spaces, despite the presence of numerous correlated features. This robustness can be attributed to the ability of isolation forests and ensemble methods to effectively handle high-dimensional data. Our classifier's 100-neuron penultimate layer is one of the best-performing hyperparameter settings, with any reasonably large latent space yielding comparable results.

It is worth noting that while classifiers demonstrate effectiveness in anomaly detection, we find little correlation between classification accuracy and anomaly detection performance. This highlights a key drawback in terms of interpretability in both DNN frameworks and our approach, warranting further investigation.

5. Conclusion

In this work, we have introduced a novel approach that leverages the latent space of a neural network classifier for identifying anomalous transients. Our pipeline, which combines a deep recurrent neural network classifier with our novel Multi-Class Isolation Forest (`MCIF`) anomaly detection method, demonstrates promising performance on simulated data matched to the characteristics of the Zwicky Transient Facility and when compared to other state-of-the-art anomaly detection methods.

The key advantages of our approach are:

1. The recurrent neural network (RNN) classifier maps light curves into a low-dimensional latent space that naturally clusters similar transient classes together, providing an effective representation for anomaly detection. We repurposed the penultimate layer of this classifier as the feature space for anomaly detection.

2. Our novel `MCIF` method addresses the limitations of using a single isolation forest on the complex latent space by training separate isolation forests for each known transient class and taking the minimum score as the final anomaly score.

A significant contribution of this work is the demonstration that a well-trained classifier can be effectively repurposed for anomaly detection by leveraging the clustering properties of its latent space. The flexibility of our approach allows for the adaptation of any classifier to an anomaly detector. For example, using existing classifiers as feature extractors for astronomical spectra, images, or time series from other domains, we can build effective anomaly detectors.

References

Bellm, E. C., Kulkarni, S. R., Graham, M. J., Dekany, R., Smith, R. M., Riddle, R., Masci, F. J., Helou, G., Prince, T. A., Adams, S. M., Barbarino, C., Barlow, T., Bauer, J., Beck, R., Belicki, J., Biswas, R., Blagorodnova, N., Bodewits, D., Bolin, B., Brinnel, V., Brooke, T., Bue, B., Bulla, M., Burruss, R., Cenko, S. B., Chang, C.-K., Connolly, A., Coughlin, M., Cromer, J., Cunningham, V., De, K., Delacroix, A., Desai, V., Duev, D. A., Eadie, G., Farnham, T. L., Feeney, M., Feindt, U., Flynn, D., Franckowiak, A., Frederick, S., Fremling, C., Gal-Yam, A., Gezari, S., Giomi, M., Goldstein, D. A., Golkhou, V. Z., Goobar, A., Groom, S., Hacopians, E., Hale, D., Henning, J., Ho, A. Y. Q., Hover, D., Howell, J., Hung, T., Huppenkothen, D., Imel, D., Ip, W.-H., Ivezić, Ž ., Jackson, E., Jones, L., Juric, M., Kasliwal, M. M., Kaspi, S., Kaye, S., Kelley, M. S. P., Kowalski, M., Kramer, E., Kupfer, T., Landry, W., Laher, R. R., Lee, C.-D., Lin, H. W., Lin, Z.-Y., Lunnan, R., Giomi, M., Mahabal, A., Mao, P., Miller, A. A., Monkewitz, S., Murphy, P., Ngeow, C.-C., Nordin, J., Nugent, P., Ofek, E., Patterson, M. T., Penprase, B., Porter, M., Rauch, L., Rebbapragada, U., Reiley, D., Rigault, M., Rodriguez, H., van Roestel, J., Rusholme, B., van Santen, J., Schulze, S., Shupe, D. L., Singer, L. P., Soumagnac, M. T., Stein, R., Surace, J., Sollerman, J., Szkody, P., Taddia, F., Terek, S., Sistine, A. V., van Velzen, S., Vestrand, W. T., Walters, R., Ward, C., Ye, Q.-Z., Yu, P.-C., Yan, L., and Zolkower, J. The zwicky transient facility: System overview, performance, and first results. *Publications of the Astronomical Society of the Pacific*, 131(995):018002, dec 2018. doi: 10.1088/1538-3873/aaecbe. URL https://doi.org/10.1088%2F1538-3873%2Faaecbe.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734. Association for Computational Linguistics, 2014. doi: 10.3115/v1/D14-1179.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

Etsebeth, V., Lochner, M., Walmsley, M., and Grespan, M. Astronomaly at scale: Searching for anomalies amongst 4 million galaxies, 2023.

Foley, R. J. and Mandel, K. CLASSIFYING SUPERNOVAE USING ONLY GALAXY DATA. *The Astrophysical Journal*, 778(2):167, nov 2013. doi: 10.1088/0004-637x/778/2/167. URL https://doi.org/10.1088%2F0004-637x%2F778%2F2%2F167.

Giles, D. and Walkowicz, L. Systematic serendipity: a test of unsupervised machine learning as a method for anomaly detection. *Monthly Notices of the Royal Astronomical Society*, 484(1):834–849, Mar 2019. doi: 10.1093/mnras/sty3461.

Ishida, E. E. O., Kornilov, M. V., Malanchev, K. L., Pruzhinskaya, M. V., Volnova, A. A., Korolev, V. S., Mondon, F., Sreejith, S., Malancheva, A. A., and Das, S. Active anomaly detection for time-domain discoveries. *Research in Astronomy and Astrophysics*, 650:A195, June 2021. doi: 10.1051/0004-6361/202037709.

Ivezić, Ž., Kahn, S. M., Tyson, J. A., Abel, B., Acosta, E., Allsman, R., Alonso, D., AlSayyad, Y., Anderson, S. F., Andrew, J., and et al. LSST: From Science Drivers to Reference Design and Anticipated Data Products. *ApJ*, 873:111, March 2019. doi: 10.3847/1538-4357/ab042c.

Kasen, D. Seeing the Collision of a Supernova with Its Companion Star. *ApJ*, 708(2):1025–1031, January 2010. doi: 10.1088/0004-637X/708/2/1025.

Kessler, R., Narayan, G., Avelino, A., Bachelet, E., Biswas, R., Brown, P. J., Chernoff, D. F., Connolly, A. J., Dai, M., Daniel, S., Stefano, R. D., Drout, M. R., Galbany, L., Gonzá lez-Gaitán, S., Graham, M. L., Hložek, R., Ishida, E. E. O., Guillochon, J., Jha, S. W., Jones, D. O., Mandel, K. S., Muthukrishna, D., O'Grady, A., Peters, C. M., Pierel, J. R., Ponder, K. A., Prša, A., Rodney, S., and and, V. A. V. Models and simulations for the photometric LSST astronomical time series classification challenge (PLAsTiCC). *Publications of the Astronomical Society of the Pacific*, 131(1003):094501, jul 2019. doi: 10.1088/1538-3873/ab26f1. URL https://doi.org/10.1088%2F1538-3873%2Fab26f1.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. Isolation forest. In *2008 eighth ieee international conference on data mining*, pp. 413–422. IEEE, 2008.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

Muthukrishna, D., Narayan, G., Mandel, K. S., Biswas, R., and Hložek, R. RAPID: Early classification of explosive transients using deep learning. *Publications of the Astronomical Society of the Pacific*, 131(1005):118002, sep 2019. doi: 10.1088/1538-3873/ab1609. URL https://doi.org/10.1088%2F1538-3873%2Fab1609.

Muthukrishna, D., Mandel, K. S., Lochner, M., Webb, S., and Narayan, G. Real-time detection of anomalies in large-scale transient surveys. *Monthly Notices of the Royal Astronomical Society*, 517(1):393–419, sep 2022. doi: 10.1093/mnras/stac2582. URL https://doi.org/10.1093%2Fmnras%2Fstac2582.

Perez-Carrasco, M., Cabrera-Vives, G., Hernandez-García, L., Förster, F., Sanchez-Saez, P., Arancibia, A. M. M., Arredondo, J., Astorga, N., Bauer, F. E., Bayo, A., Catelan, M., Dastidar, R., Estévez, P. A., Lira, P., and Pignata, G. Alert classification for the alerce broker system: The anomaly detector. *The Astronomical Journal*, 166(4):151, sep 2023. doi: 10.3847/1538-3881/ace0c1. URL https://dx.doi.org/10.3847/1538-3881/ace0c1.

Pruzhinskaya, M. V., Malanchev, K. L., Kornilov, M. V., Ishida, E. E. O., Mondon, F., Volnova, A. A., and Korolev, V. S. Anomaly detection in the open supernova catalog. *Monthly Notices of the Royal Astronomical Society*, aug 2019. doi: 10.1093/mnras/stz2362. URL https://doi.org/10.1093%2Fmnras%2Fstz2362.

Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., and Kloft, M. Deep one-class classification. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4393–4402. PMLR, 10–15 Jul 2018a. URL https://proceedings.mlr.press/v80/ruff18a.html.

Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., and Kloft, M. Deep one-class classification. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4393–4402, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018b. PMLR. URL http://proceedings.mlr.press/v80/ruff18a.html.

Rumelhart, D. E. and McClelland, J. L. *Learning Internal Representations by Error Propagation*, pp. 318–362. 1987.

Sánchez-Sáez, P., Reyes, I., Valenzuela, C., Förster, F., Eyheramendy, S., Elorrieta, F., Bauer, F. E., Cabrera-Vives, G., Estévez, P. A., Catelan, M., Pignata, G., Huijse, P., De Cicco, D., Arévalo, P., Carrasco-Davis, R., Abril, J., Kurtev, R., Borissova, J., Arredondo, J., Castillo-Navarrete, E., Rodriguez, D., Ruz-Mieres, D., Moya, A., Sabatini-Gacitúa, L., Sepúlveda-Cobo, C., and Camacho-Iñiguez, E. Alert Classification for the ALeRCE Broker System: The Light Curve Classifier. *The Astronomical Journal*, 161(3):141, March 2021. doi: 10.3847/1538-3881/abd5c1.

Schölkopf, B., Williamson, R. C., Smola, A., Shawe-Taylor, J., and Platt, J. Support vector method for novelty detection. In Solla, S., Leen, T., and Müller, K. (eds.), *Advances in Neural Information Processing Systems*, volume 12 of *Proceedings of the 12th International Conference on Neural Information Processing Systems*. MIT Press, 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/file/8725fb777f25776ffa9076e44fcfd776-Paper.pdf.

Singh, S., Luo, M., and Li, Y. Multi-class anomaly detection, 2022.

Soraisam, M. D., Saha, A., Matheson, T., Lee, C.-H., Narayan, G., Vivas, A. K., Scheidegger, C., Oppermann, N., Olszewski, E. W., Sinha, S., Desantis, S. R., and ANTARES Collaboration. A Classification Algorithm for Time-domain Novelties in Preparation for LSST Alerts. Application to Variable Stars and Transients Detected with DECam in the Galactic Bulge. *ApJ*, 892(2):112, April 2020. doi: 10.3847/1538-4357/ab7b61.

The PLAsTiCC team, Allam, Tarek, J., Bahmanyar, A., Biswas, R., Dai, M., Galbany, L., Hložek, R., Ishida, E. E. O., Jha, S. W., Jones, D. O., Kessler, R., Lochner, M., Mahabal, A. A., Malz, A. I., Mandel, K. S., Martínez-Galarza, J. R., McEwen, J. D., Muthukrishna, D., Narayan, G., Peiris, H., Peters, C. M., Ponder, K., Setzer, C. N., The LSST Dark Energy Science Collaboration, LSST Transients, T., and Variable Stars Science Collaboration. The Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC): Data set. *arXiv e-prints*, art. arXiv:1810.00001, September 2018. doi: 10.48550/arXiv.1810.00001.

Villar, V. A., Cranmer, M., Berger, E., Contardo, G., Ho, S., Hosseinzadeh, G., and Lin, J. Y.-Y. A deep-learning approach for live anomaly detection of extragalactic transients. *The Astrophysical Journal Supplement Series*, 255 (2):24, 2021.

Walmsley, M., Scaife, A. M. M., Lintott, C., Lochner, M., Etsebeth, V., Géron, T., Dickinson, H., Fortson, L., Kruk, S., Masters, K. L., Mantha, K. B., and Simmons, B. D. Practical galaxy morphology tools from deep supervised representation learning. *Monthly Notices of the Royal Astronomical Society*, 513(2):1581–1599, 02 2022. ISSN 0035-8711. doi: 10.1093/mnras/stac525. URL https://doi.org/10.1093/mnras/stac525.

Webb, S., Lochner, M., Muthukrishna, D., Cooke, J., Flynn, C., Mahabal, A., Goode, S., Andreoni, I., Pritchard, T., and Abbott, T. M. C. Unsupervised machine learning for transient discovery in deeper, wider, faster light curves. *Monthly Notices of the Royal Astronomical Society*, 498 (3):3077–3094, September 2020. doi: 10.1093/mnras/staa2395.

Table 2. Number of transients in the training set, validation set, test set, and realistic samples (see section 4.3) for each class. All anomalous data is reserved for evaluation.

Class	Training	Validation	Test	Total	Realistic Sample[a]
SNIa	9314	1131	1142	11587	1142
SNIa-91bg	10361	1318	1321	13000	1318
SNIax	10413	1248	1339	13000	1339
SNIb	4197	507	563	5267	563
SNIc	1279	169	135	1583	135
SNIc-BL	1157	124	142	1423	142
SNII	10420	1279	1301	13000	1301
SNIIn	10323	1359	1318	13000	1318
SNIIb	9882	1233	1208	12323	1208
TDE	9078	1162	1114	11354	1114
SLSN-I	10285	1322	1273	12880	1273
AGN	8473	1046	1042	10561	1042
CaRT	0	0	10353	10353	11 ± 3
KNe	0	0	11166	11166	11 ± 3
PISN	0	0	10840	10840	11 ± 3
ILOT	0	0	11128	11128	10 ± 3
uLens-BSR	0	0	11244	11244	10 ± 3

[a] The mean number of transients across the 50 test samples is shown. The errors refer to the STD in the population size across the 50 sets. All common test data is part of every sample, hence errors are not shown.

A. Visual Comparison to other Approaches

Figure 7 is a visual representation of the results depicted in Table 1.

B. Dataset Information

A sample light curve from each class is illustrated in Figure 8. Table 2 reports the number of objects from each class in our training set and realistic sample used for evaluation in Section 4.3.

C. Classifier Results

The normalized confusion matrix in Figure 9 [left] illustrates our classifier's ability to accurately predict the correct transient class on the test data. Each cell indicates the fraction of transients from the true class that are classified into the predicted class. The high values along the diagonal, approaching 1.0, indicate strong performance. The misclassifications, indicated by the off-diagonal values, predominantly occur between subclasses of Type Ia supernovae (SNIa, SNIa-91bg and SNIax) and between the core-collapse supernova types (SNIb, SNIc, SNII subtypes), which is expected given their observational similarities. These SNe have been shown to confuse previous models (see Fig. 7 of Muthukrishna et al., 2019).

D. Real-Time Detection

Identifying anomalies in real-time is important for obtaining early-time follow-up observations, which is crucial for understanding their physical mechanisms and progenitor systems (e.g. Kasen, 2010). However, directly assessing our architecture's real-time performance is challenging due to the irregular sampling of light curves in our input format.

To assess the real-time performance of our architecture, we plot the median anomaly scores over time for a sample of 2000 common and 2000 anomalous transients in Figure 10. To construct this plot without relying on interpolation, we calculate scores at discrete times l sampled at 1-day intervals from -30 to 70 days relative to trigger, using only observations

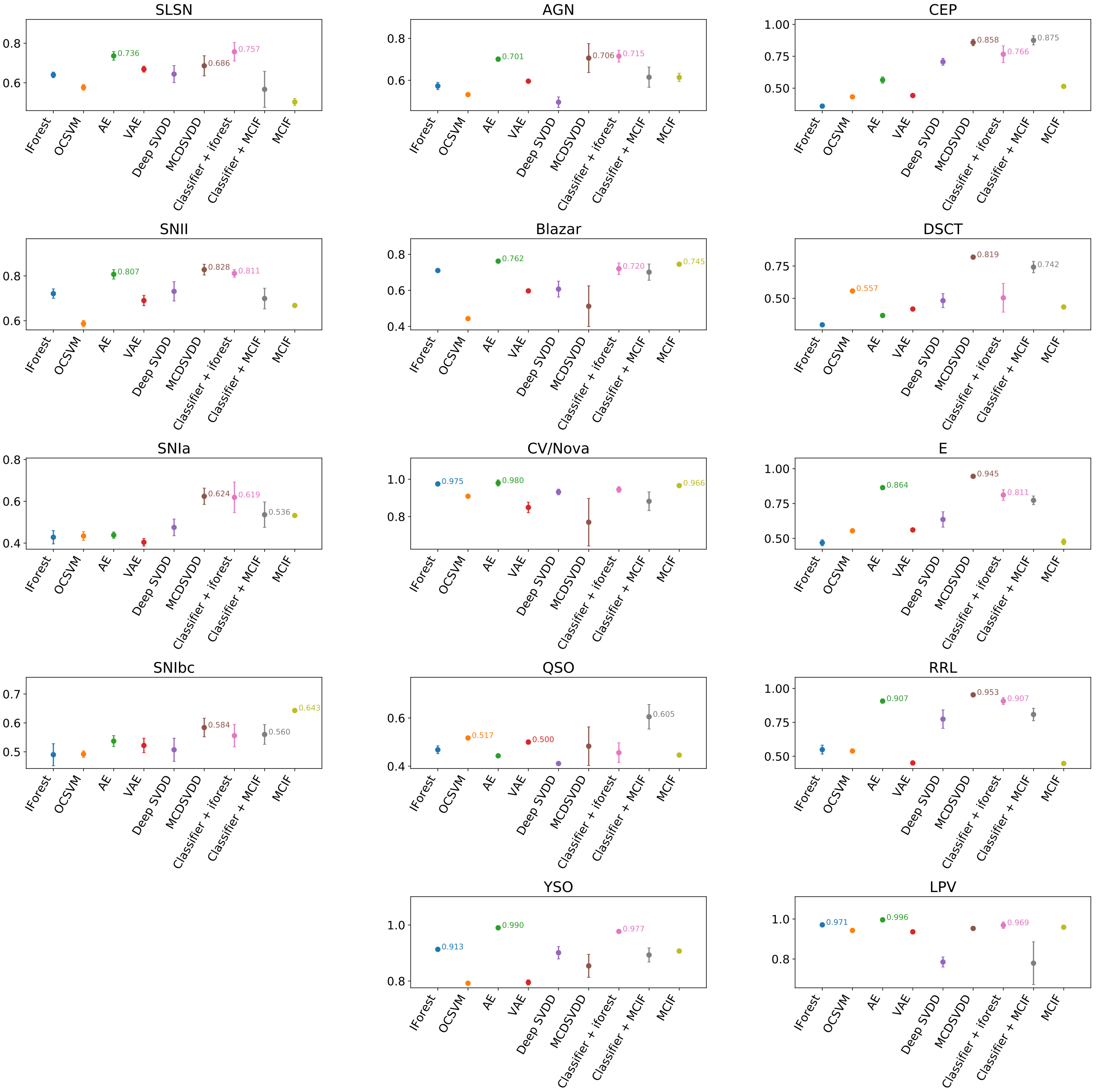


Figure 7. Visual representation of the comparative analysis depicted in Table 1. The AUROC is written for the models top 3 models for each class.

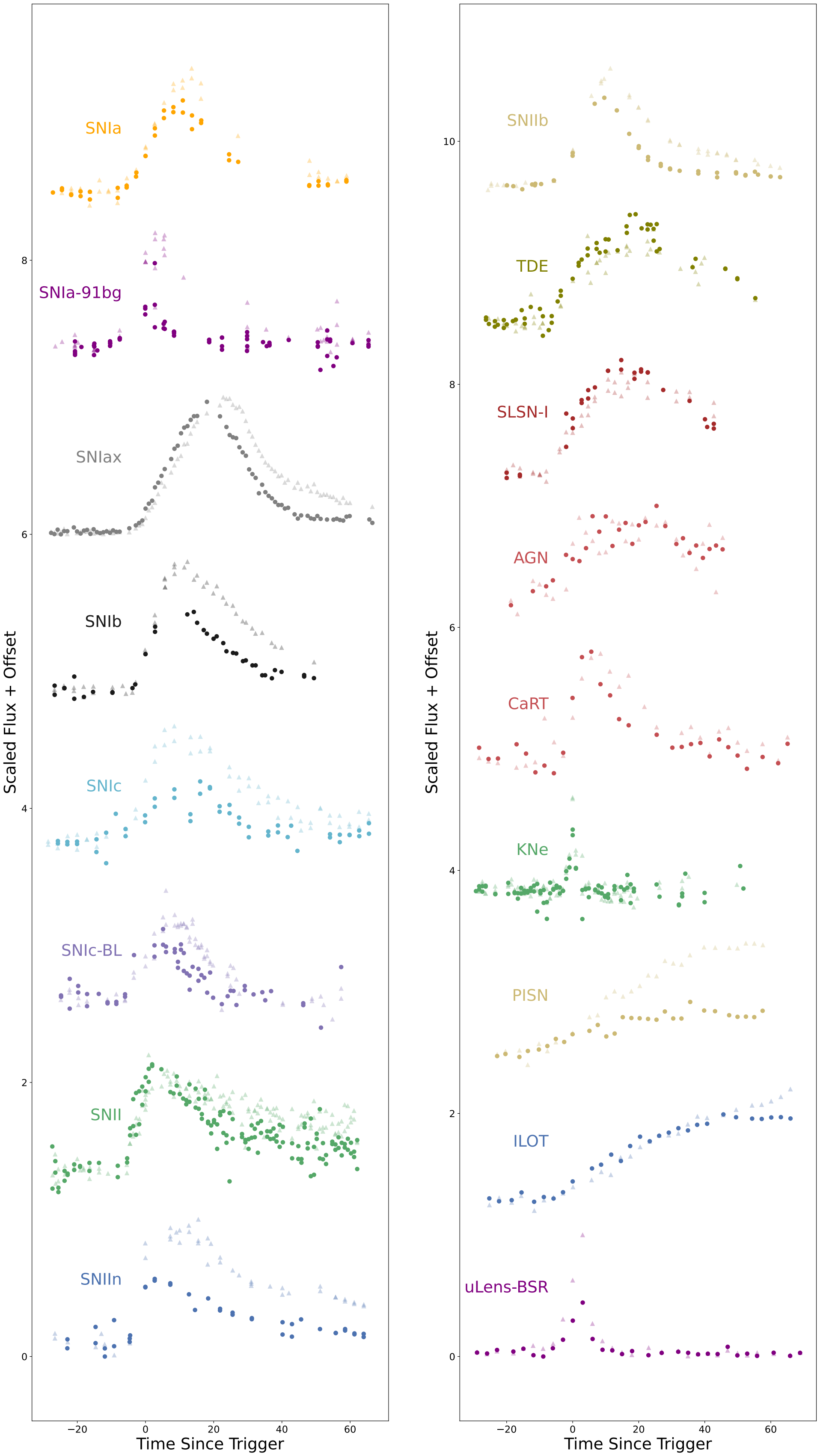
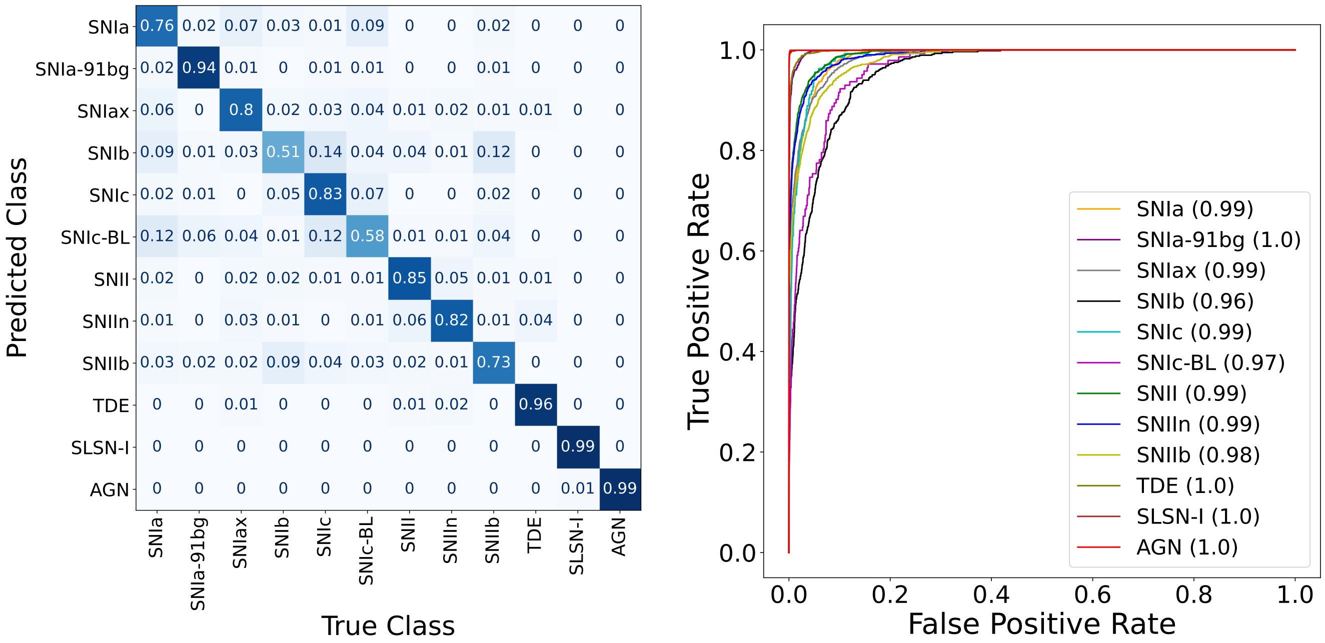


Figure 8. Sample light curves from each transient class used in this work. We only plot transients with low signal-to-noise to help visually compare shapes. The dark circular markers represent the r band while the light triangular markers represent the g band. Flux errors are not plotted.



	SNIa	SNIa-91bg	SNIax	SNIb	SNIc	SNIc-BL	SNII	SNIIn	SNIIb	TDE	SLSN-I	AGN
SNIa	0.76	0.02	0.07	0.03	0.01	0.09	0	0	0.02	0	0	0
SNIa-91bg	0.02	0.94	0.01	0	0.01	0.01	0	0	0.01	0	0	0
SNIax	0.06	0	0.8	0.02	0.03	0.04	0.01	0.02	0.01	0.01	0	0
SNIb	0.09	0.01	0.03	0.51	0.14	0.04	0.04	0.01	0.12	0	0	0
SNIc	0.02	0.01	0	0.05	0.83	0.07	0	0	0.02	0	0	0
SNIc-BL	0.12	0.06	0.04	0.01	0.12	0.58	0.01	0.01	0.04	0	0	0
SNII	0.02	0	0.02	0.02	0.01	0.01	0.85	0.05	0.01	0.01	0	0
SNIIn	0.01	0	0.03	0.01	0	0.01	0.06	0.82	0.01	0.04	0	0
SNIIb	0.03	0.02	0.02	0.09	0.04	0.03	0.02	0.01	0.73	0	0	0
TDE	0	0	0.01	0	0	0	0.01	0.02	0	0.96	0	0
SLSN-I	0	0	0	0	0	0	0	0	0	0	0.99	0
AGN	0	0	0	0	0	0	0	0	0	0	0.01	0.99

Figure 9. The normalized confusion matrix [left] and ROC curve [right] of the 12 common transient classes used for training given full light curve data. Each cell in the confusion matrix signifies the fraction of transients from each *True Class* that was classified into the *Predicted Class*. The ROC curve illustrates the True Positive Rate against the False Positive Rate across various threshold probabilities for each class, with the Area Under ROC curve (AUROC) in parenthesis. The model's evaluation is conducted on the test set consisting of 10% of the data from the common classes.

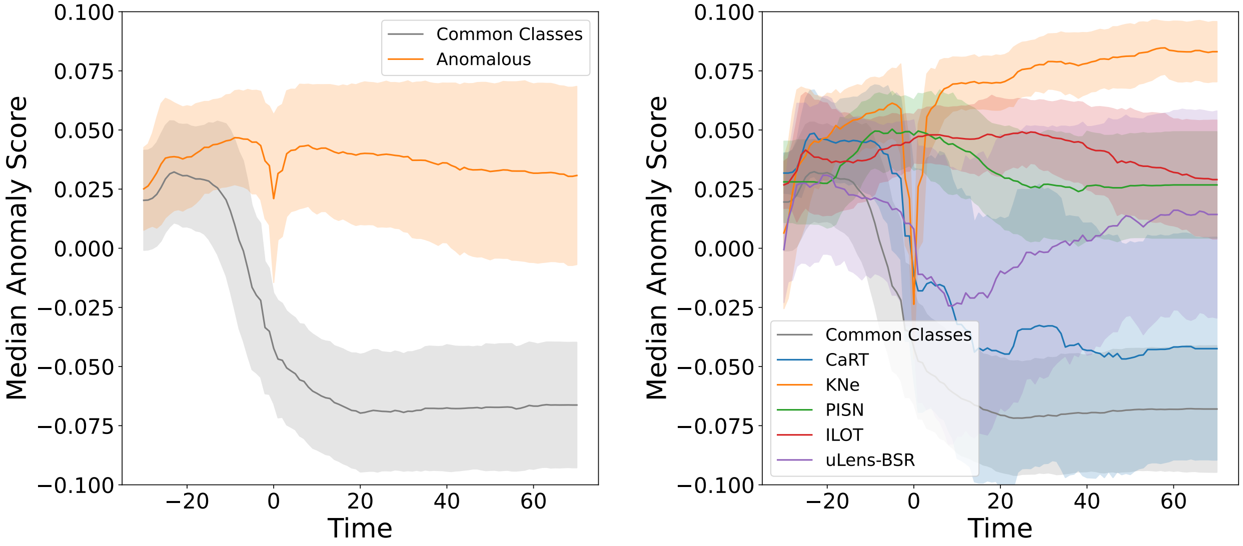


Figure 10. Median `MCIF` anomaly score over time for a sample of transients from the test set. Real-time anomaly scores are calculated at intervals of 1 day for a sample of 2000 common and 2000 total anomalous light curves. The left plot shows the scores for the common and anomalous transients as a whole, while the right plot shows each anomalous class individually. The anomaly scores for the common transients decline before the trigger, while the anomalous transients remain at high scores throughout most of the transient's evolution.

occurring before each time l to mimic a real-time scenario. To ensure sufficient information for robust scoring, we only consider transients where the final observation was recorded after time $l - 5$. The results show a clear divergence where common transient scores tend to decline around trigger, while anomalous transient scores remain consistently high.

Figure 10 reveals two notable irregularities. Firstly, the anomaly scores for common transients decline before trigger, which is unexpected given that the pre-trigger phase of most transient classes should primarily consist of background noise. Further analysis of the pre-trigger classification results reveals that certain transients, most notably SLSN-I and AGN, are almost all classified before trigger, thereby lowering the average anomaly score for common transients. This can be attributed to the fact that redshift and pre-trigger information such as host galaxy color and some AGN pre-trigger variability are particularly useful for classifying these transients before trigger (see Figure 16 of Muthukrishna et al., 2019).

Secondly, KNe exhibit a significant dip around the time of trigger. Upon further analysis, we found that certain common transient classes also experienced a similar dip around trigger; however, unlike KNe, they do not rebound back to higher anomaly scores. This dip is related to the inherent nature of the trigger of a light curve, which often marks the first *real* observation of the transient phase of a light curve, and serves as a reset for the anomaly score. A more detailed analysis of this phenomenon is omitted for brevity.

These preliminary findings suggest the potential for enabling real-time identification of anomalous transients. While some known rare classes can be difficult to distinguish from the common classes without a significant amount of data, others can be detected surprisingly soon after trigger. The ability to flag unusual events early in their evolution could prove invaluable for optimizing the allocation of follow-up resources and maximizing the scientific returns from rare transient discoveries.