Score-Based Log-Density Estimation for Anomaly Detection in Astronomical Surveys

Sebastián Guzmán-Olave
Dept. of Electrical Engineering
University of Chile
Santiago, Chile
sebastianguzman@ug.uchile.cl

Pablo A. Estévez
Dept. of Electrical Engineering
University of Chile
Santiago, Chile
pestevez@cec.uchile.cl

***Abstract*—Astronomical brokers must filter millions of alerts to prioritize follow-up in surveys such as the Zwicky Transient Facility (ZTF) and the Vera C. Rubin Observatory's Legacy Survey of Space and Time (LSST). To address this challenge we present Score-based Log-Density Network (SLDNet), an unsupervised generative model that estimates the gradient of the data log-density in a multiscale manner. Trained with *denoising score matching*, SLDNet defines an anomaly score using a conservative potential as a proxy for the negative log-density, avoiding the calculation of the normalization constant.**

We evaluate SLDNet on two complementary protocols: (i) *leave-one-class-out* on 152 tabular features extracted by the Automatic Learning for the Rapid Classification of Events (ALeRCE) broker for ZTF, where SLDNet achieves the best area under the receiver operating characteristic curve in most transient and stochastic subclasses; and (ii) Out-of-distribution discovery and Rare-in-distribution scenarios following the Distance Multi-Metric Anomaly Detection (DiMMAD) evaluation protocol on the Extended LSST Astronomical Time-series Classification Challenge (ELAsTiCC) and ZTF using Supernova Parametric Model features, reporting area under the precision–recall curve and purity/diversity vs. follow-up budget. To our knowledge, this is the first study validating multiscale log-density estimation based on *denoising score matching* for anomaly detection in the astronomical domain.

***Index Terms*—anomaly detection, score matching, generative models, tabular data, astronomy.**

I. Introduction

Large-scale surveys such as the Zwicky Transient Facility (ZTF) [1] and the upcoming Legacy Survey of Space and Time (LSST) of the Vera C. Rubin Observatory [2] generate massive alert streams, making anomaly detection—the identification of sources that deviate from known patterns—a key step for discovery.

Astronomical brokers ingest alert streams, enrich them with features and contextual information, and rank candidates for follow-up (e.g., the Automatic Learning for the Rapid Classification of Events (ALeRCE) broker [3]). We propose SLDNet (*Score-based Log-Density Network*), a score-based log-density estimator for anomaly detection on the tabular feature space extracted from ZTF light curves [1]. Unlike methods based on direct likelihood, *score matching* [4] learns the gradient of the log-density (the *score*) without computing the normalization constant, simplifying training and scaling to high dimensions. Inspired by multiscale log-density estimators such as MULDE [5], SLDNet learns a conservative potential that yields an anomaly score with probabilistic interpretation across variability families of stellar objects.

To situate our work within the state of the art, we compare SLDNet with the ALeRCE anomaly detector, the so-called Multi-Class Deep Support Vector Data Description (MCSVDD), which models *inlier* categories using a hierarchy of hyperspheres [6], [7]. We also compare our approach against the Distance Multi-Metric Anomaly Detection (DiMMAD) [8], a discovery approach based on distance ensembles evaluated on the Extended LSST Astronomical Time-series Classification Challenge (ELAsTiCC) and ZTF. In addition to reporting the area under the receiver operating characteristic curve (AUROC) per subclass, we adopt DiMMAD's discovery metrics, such as the area under the precision–recall curve (AUPRC) and purity/diversity curves against follow-up budget, to evaluate high-confidence regimes critical for resource allocation.

To the best of our knowledge, this is the first work adapting multiscale *score*-based log-density estimation to anomaly detection in astronomy. Our main contributions are:

- An adaptation to the tabular domain. Unlike MULDE, which utilizes a pre-trained encoder to regularize the input space for video anomaly detection, SLDNet operates directly on the astronomical tabular data, avoiding the need for auxiliary pre-trained networks.
- An inference strategy that combines normalized scores across noise scales, simplifying the Gaussian mixture model (GMM)-based aggregation used in previous works while maintaining estimator robustness.
- Achieve state-of-the-art results on anomaly detection (AUROC) in complex astronomical families such as transients and stochastics events, and offering the best cost-benefit trade-off (AUPRC) in discovery scenarios.

This paper is structured as follows: Section II describes the related work; Section III presents the proposed method; Section IV describes the experimental setup; Section V presents

This work was partially funded by the National Agency for Research and Development (ANID) Human Capital Subdirection National Master's Scholarship Program 2024-22240913, Millennium Institute of Astrophysics (AIM23_0001) and FONDECYT Project 1220829.

the experimental results; Section VI analyzes the findings and limitations; and Section VII concludes the work[1].

II. Related Work

Anomaly detection is a well-established field in machine learning; classic reviews discuss definitions, assumptions, and metrics in imbalanced scenarios [9]. In astronomy, the goal is to identify celestial objects exhibiting unusual behavior.

Unsupervised approaches include space partitioning-based methods such as Isolation Forest [10], kernel-based one-class models such as One-Class SVM (OC-SVM) [11], and related formulations such as Support Vector Data Description (SVDD) [12], which seeks a minimal hypersphere containing the majority of normal data.

With the advent of deep learning, reconstruction-based methods such as autoencoders [13] and variational generative models [14] have been widely used. Combining SVDD with deep learning has led to methods like Deep SVDD [15] and multi-class extensions like MCSVDD [7], which model distinct *inlier* categories using multiple hyperspheres.

In time-domain astronomy, the ALeRCE broker implements anomaly detection on the ZTF alert stream using tabular features computed from light curves (brightness of an astronomical object over time) together with contextual information from alert packets and cross-matches. In [6], it is reported that for transient and periodic sources the best performance is obtained with a Deep SVDD variant, while for stochastic sources it is obtained with autoencoders, suggesting that the variability family impacts the suitability of the anomaly score.

In the context of scientific discovery, DiMMAD [8] proposes an ensemble of distance metrics for anomaly detection and evaluates two complementary scenarios: out-of-distribution (OOD) discovery (new classes) and rare-in-distribution (RID) discovery (rare subtypes). In that work, training is performed on features derived from the Supernova Parametric Model (SPM): 25 features for ELAsTiCC and 14 features for ZTF, chosen for their low failure rate, in contrast to the 152 standard tabular features used in the ALeRCE detector [8]. This difference is important when interpreting results across configurations.

Recent literature emphasizes that performance in discovery scenarios depends on both the model and the evaluation metric; thus, log-density estimation offers an alternative that avoids explicit selection of distance metrics [5], [8].

III. Proposed Method

SLDNet is a variant of the family of *score*-based generative models, which adopts a multiscale log-density estimation formulation like MULDE [5]. In this work, we use this formulation to model the gradient of the log-density (the *score*) of the data distribution of sources considered normal, without focusing on sample generation.

[1]Code is available at https://github.com/M4thinking/sldnet

A. Score Matching and Denoising Score Matching

Let $\mathbf{x} \in \mathbb{R}^d$ be a vector of tabular features derived from a light curve. The *score* is defined as $\nabla_{\mathbf{x}} \log p(\mathbf{x})$. In *score matching*, a model $s_\theta(\mathbf{x})$ is trained to approximate the gradient without requiring a closed form for $p(\mathbf{x})$ [4]. In practice, training is usually stabilized using *denoising score matching* (DSM) [16], where samples are perturbed with Gaussian noise $\tilde{\mathbf{x}} = \mathbf{x} + \sigma\boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and the *score* of the smoothed distribution $q_\sigma(\tilde{\mathbf{x}})$ is learned at multiple noise levels.

B. SLDNet as a Multiscale Potential

In SLDNet, we model a scalar potential $f_\theta(\tilde{\mathbf{x}}, \sigma)$ that approximates $-\log q_\sigma(\tilde{\mathbf{x}})$ (up to an additive constant dependent on σ). The induced *score* is obtained as a conservative field $s_\theta(\tilde{\mathbf{x}}, \sigma) = -\nabla_{\tilde{\mathbf{x}}} f_\theta(\tilde{\mathbf{x}}, \sigma)$. The use of multiple noise levels allows capturing structure at different scales and regularizing training analogously to multiscale approaches in log-density estimation [5]. Here, $\sigma > 0$ denotes the noise level. We minimize a DSM objective with variance-preserving weighting $\lambda(\sigma) = \sigma^2$ and a potential-regularization coefficient $\beta = 10^{-3}$, motivated by the framework of MULDE [5]:

$$\min_\theta \ \mathbb{E}_{\substack{\mathbf{x}\sim p(\mathbf{x}) \\ \tilde{\mathbf{x}}\sim\mathcal{N}(\mathbf{x},\sigma^2\mathbf{I}) \\ \sigma\sim\mathcal{LU}(\sigma_{\text{low}},\sigma_{\text{high}})}} \left[\lambda(\sigma) \left\| \nabla_{\tilde{\mathbf{x}}} f_\theta(\tilde{\mathbf{x}}, \sigma) - \frac{\tilde{\mathbf{x}} - \mathbf{x}}{\sigma^2} \right\|_2^2 + \beta f_\theta(\mathbf{x}, \sigma)^2 \right]. \tag{1}$$

1) Architecture: In our implementation, f_θ is a Multilayer Perceptron (MLP) for tabular data that produces a scalar output $f_\theta(\tilde{\mathbf{x}}, \sigma) \in \mathbb{R}$ which approximates $-\log q_\sigma(\tilde{\mathbf{x}})$. We use two dense layers of 1024 and 512 neurons with Gaussian error linear unit (GELU) activation and *layer normalization*. The *score* $s_\theta(\tilde{\mathbf{x}}, \sigma)$ is obtained via automatic differentiation.

2) Noise Sampling and Anomaly Score: During training, we sample noise levels σ from a log-uniform distribution $\sigma \sim \mathrm{LogUniform}(\sigma_{\text{low}}, \sigma_{\text{high}})$, with $\sigma_{\text{low}} = 10^{-3}$ and $\sigma_{\text{high}} = 3$. The upper limit $\sigma_{\text{high}} = 3$ is selected to fully cover the dynamic range of the standardized data (with unit variance), ensuring that at the highest scale the data distribution structure diffuses towards an isotropic Gaussian.

In inference, we perturb the test point with the lowest noise level to regularize evaluation in the data manifold region:

$$\tilde{\mathbf{x}}_{\text{low}} = \mathbf{x} + \sigma_{\text{low}}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{2}$$

Fig. 1 illustrates the relationship between the (smoothed) density and the learned potential, and clarifies why DSM provides a tractable surrogate for score estimation without evaluating normalization constants. We define a scale-wise anomaly detection score $\mathrm{SAD}_\sigma(\mathbf{x})$ as the potential evaluated at the perturbed point (higher SAD indicates lower density):

$$\mathrm{SAD}_\sigma(\mathbf{x}) = f_\theta(\mathbf{x} + \sigma\boldsymbol{\epsilon}, \sigma), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{3}$$

In experiments, we report two variants, identified by L. For $L = 0.001$, we use a single inference scale at $\sigma_{\text{low}} = 10^{-3}$ and report $\mathrm{SAD}_{\sigma_{\text{low}}}(\mathbf{x})$. For $L = 4$, we use four scales in a

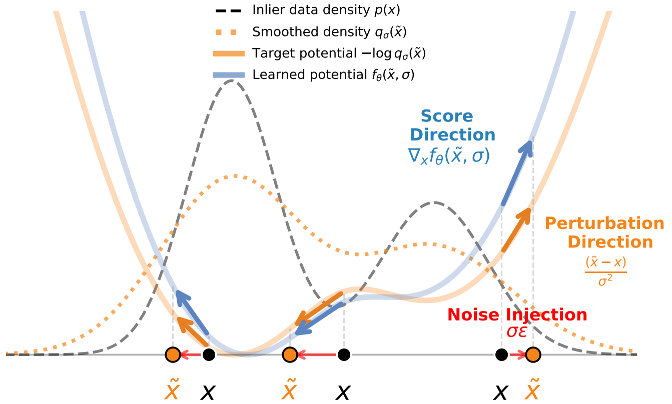


Figure 1: One-dimensional intuition for *denoising score matching*. Noise maps $x \sim p(x)$ to $\tilde{x} = x + \sigma\epsilon$, inducing a smoothed density $q_\sigma(\tilde{x})$. Training enforces $\nabla_{\tilde{x}} f_\theta(\tilde{x}, \sigma) \approx (\tilde{x} - x)/\sigma^2$, hence $s_\theta(\tilde{x}, \sigma) = -\nabla_{\tilde{x}} f_\theta(\tilde{x}, \sigma)$ approximates $\nabla_{\tilde{x}} \log q_\sigma(\tilde{x})$.

`linspace` within $[\sigma_{\text{low}}, \sigma_{\text{high}}]$, calculate $\text{SAD}_{\sigma_\ell}(\mathbf{x})$ at each scale, normalize each score using its mean and standard deviation in training, and average the normalized scores to obtain an aggregated score. This scheme differs from MULDE, which combines scales using a Gaussian Mixture Model (GMM) [5].

IV. Experimental Setup

A. Datasets

We evaluate SLDNet on two datasets commonly used in time-domain discovery benchmarks: real sources from the Zwicky Transient Facility (ZTF) [1] using forced photometry, and the simulated Extended LSST Astronomical Time-series Classification Challenge (ELAsTiCC), designed to mimic Rubin/LSST light curves [8].

ZTF (real, g/r bands). We use the ALeRCE-labeled sample and taxonomy [3], [17]. We report results for 14 subclasses grouped into three top-level families: transients (SLSN, SNII, SNIa, SNIbc), where SN stands for Supernova, stochastics (Active Galactic Nucleus (AGN), Blazar, CV/Nova, Quasi-Stellar object (QSO), Young Stellar Object (YSO)), and periodics (Cepheids (CEP), Delta Scuti (DSCT), Eclipsing Binaries (E), RR Lyrae (RRL), Long Period Variable (LPV)) [6]. For each object, 152 tabular features are computed from light curves for sources with at least six detections in either the g or r band, together with contextual cross-matches [6].

ELAsTiCC ELAsTiCC is a simulated dataset designed to mimic Rubin/LSST six-band ($ugrizy$) light curves and includes diverse variable and transient classes, such as CEP, RRL, DSCT and E, as well as rare explosive phenomena (e.g., kilonovae, superluminous supernovae (SLSN), strongly lensed Type Ia supernovae, and microlensing events) [8].

B. Preprocessing

We apply the same preprocessing to all tabular inputs using a `QuantileTransformer` with uniform distribution, fitted on the training split to mitigate heavy tails and balance feature contributions, in line with the preprocessing used in the Astronomical Transformer for Time-Series and Tabular Data (ATAT) [18]. We replace missing values (NaN) with -1 to encode measurement absence explicitly. Finally, we standardize features using training statistics (zero mean, unit variance) to define the input data space for SLDNet.

C. Protocols

1) Leave-One-Class-Out (ALeRCE Protocol): We follow the training/evaluation methodology of the ALeRCE anomaly detector [6]. Data are split into 80% train and 20% held-out test (TS1) using a stratified split to preserve class proportions. The training set is divided into five stratified folds for model selection. For each trial, one subclass within a top-level family (transient, stochastic, periodic) is treated as the outlier class, removed from training, and used for evaluation. A second test set (TS2) is constructed as a realistic mixture with 90% inliers and 10% outliers: TS2 contains all inlier objects from TS1 and all objects from the selected outlier subclass taken from both TS1 and the training pool [6]. AUROC performance is reported for detecting the removed subclass with $L = 0.001$.

Example. In the transient family, if SLSN is selected as the outlier subclass, the detector is trained on $\{\text{SNIa}, \text{SNIbc}, \text{SNII}\}$ and evaluated on TS2 where 10% of instances are SLSN and 90% are transients from the remaining subclasses [6].

2) OOD/RID Discovery (DiMMAD Protocol): We adopt the OOD/RID discovery setup of DiMMAD [8]. The follow-up budget is defined as the number N of top-ranked sources selected for follow-up (up to $N = 300$). We report purity at budget N (precision among the top-N candidates), diversity vs. budget (cumulative number of distinct anomaly classes recovered), and AUPRC as a summary metric focused on the high-confidence regime [8]. In the discovery experiments, we represent ELAsTiCC and ZTF using features derived from the Supernova Parametric Model (SPM) fit to light curves with the ALeRCE pipeline software [8]. These features are low-dimensional, computationally inexpensive, and have low failure rate ($< 5\%$): we use 25 SPM features for ELAsTiCC (from grizy-band fits) and 14 SPM features for ZTF [8].

We use the same Monte Carlo cross-validation with 20 runs (different random seeds), evaluating all methods on identical splits [8]. For each run, the training set is balanced by down-sampling all inlier classes to the size of the smallest class. The test set is composed of unseen inliers plus a diverse, stratified sample of unknown outliers, leading to a high anomaly fraction ($\sim 47\%$) to create a statistically stable benchmark.

Examples. For ELAsTiCC OOD, models are trained on four variable-star classes (CEP, RRL, DSCT, E) and evaluated on 34 unseen transient classes as outliers. For ZTF OOD, models are trained on 11 common periodic and stochastic classes and evaluated on four supernova classes as unseen outliers. For RID, models are trained on 16 known supernova subtypes and evaluated on three rare pair-instability supernova models [8].

D. Baselines

For *leave-one-class-out* evaluation (AUROC), we compare against classic methods: Isolation Forest [10], OC-SVM [11],

Table I: AUROC per subclass in ZTF (average per class), following ALeRCE's *leave-one-class-out* protocol [6], where each class reported is treated as an anomaly within their respective group (transient, stochastic, periodic). The best mean value is shown in bold and the second best is underlined.

	Transient				Stochastic					Periodic				
Method	**SLSN**	**SNII**	**SNIa**	**SNIbc**	**AGN**	**Blazar**	**CV/Nova**	**QSO**	**YSO**	**CEP**	**DSCT**	**E**	**RRL**	**LPV**
SLDNet	**0.759 ±0.015**	0.814 ±0.014	**0.716 ±0.021**	**0.588 ±0.017**	**0.737 ±0.005**	**0.796 ±0.006**	0.977 ±0.001	**0.543 ±0.008**	0.974 ±0.001	0.760 ±0.009	0.726 ±0.045	0.934 ±0.004	0.943 ±0.011	0.995 ±0.002
MCSVDD	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	**0.887 ±0.010**	**0.819 ±0.015**	**0.945 ±0.006**	**0.953 ±0.003**	0.953 ±0.008
AE	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 ±0.004	0.367 ±0.015	0.864 ±0.009	0.907 ±0.015	**0.996 ±0.000**
Deep SVDD	0.644 ±0.043	0.690 ±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 ±0.020	0.482 ±0.054	0.636 ±0.055	0.774 ±0.068	0.785 ±0.025
VAE	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 ±0.006	0.417 ±0.007	0.561 ±0.010	0.451 ±0.006	0.936 ±0.007
IForest	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
OC-SVM	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

deep methods (autoencoder (AE) [13], variational autoencoder (VAE) [14], Deep SVDD [15], and MCSVDD [7]), following ALeRCE's protocol [6]. For OOD/RID schemes (AUPRC/purity/diversity), we incorporate baselines reported by DiMMAD (iForest, Local Outlier Factor (LOF), OC-SVM, autoencoder, and Multi-Class Deep Support Vector Data Description (MCSVDD)) and DiMMAD in two aggregation variants (min-med and med-med) [8]. In each case, we reuse splits and configurations described in reference works to maintain comparability.

V. Experimental Results

Table I summarizes AUROC per subclass under ALeRCE's *leave-one-class-out* protocol [6]. In transients, SLDNet achieves the best performance in three out of four classes, reaching 0.759 for SLSN and 0.716 for SNIa. Among stochastic sources, it performs best for AGN (0.737), Blazar (0.796), and QSO (0.543), while remaining highly competitive for CV/Nova and YSO (0.977 and 0.974). For periodic objects, although MCSVDD outperforms it in several subclasses, SLDNet remains competitive for E, RRL, and LPV.

In OOD/RID schemes (DiMMAD-style evaluation), we report AUPRC in four configurations (ELAsTiCC/ZTF $\times$ OOD/RID) to prioritize the high-confidence regime, as shown in Table II. To synthesize comparisons, we report *wins* counts based on paired permutation tests per fold (Table III). In OOD, the highest number of wins is obtained by DiMMAD (min-med), while in RID the best counts correspond to SLDNet ($L = 0.001$) and iForest. In specific results, DiMMAD (med-med) leads in ELAsTiCC OOD (0.659), while SLDNet ($L = 4$) leads in ZTF OOD (0.980). In RID, the best value in ELAsTiCC corresponds to iForest (0.367), and in ZTF the best performance is reached by SLDNet ($L = 0.001$) (0.478).

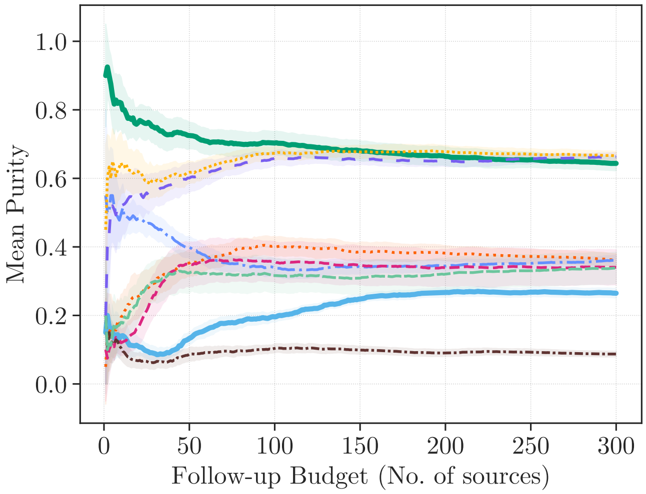
Fig. 2 shows purity vs. budget curves in ELAsTiCC and ZTF, following the evaluation scheme proposed by DiMMAD [8]. In ZTF OOD, the multiscale aggregation in SLDNet ($L = 4$) reaches near-perfect purity across budgets, while the single-scale variant ($L = 0.001$) provides a stronger purity–diversity trade-off by recovering OOD classes earlier (Fig. 3). In ELAsTiCC OOD, SLDNet ($L = 0.001$) attains the highest purity for the smallest budgets, whereas DiMMAD variants yield the strongest diversity and AUPRC. In RID, where diversity saturates quickly, differences are dominated by purity.

Table II: Average AUPRC in OOD and RID scenarios (E = ELAsTiCC, Z = ZTF). The best mean value is shown in bold and the second best is underlined.

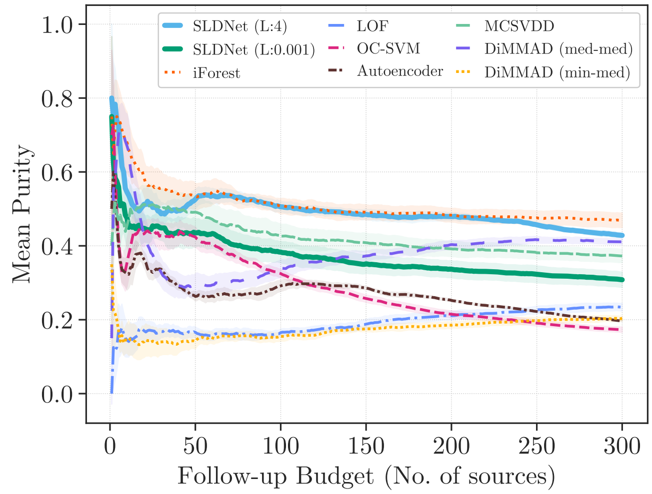
Model	**E-OOD**	**E-RID**	**Z-OOD**	**Z-RID**
SLDNet ($L = 0.001$)	0.545 ±0.016	0.173 ±0.020	0.893 ±0.019	**0.478 ±0.006**
DiMMAD (med-med)	**0.659 ±0.007**	0.337 ±0.007	0.844 ±0.019	0.422 ±0.006
DiMMAD (min-med)	0.553 ±0.018	0.227 ±0.008	0.969 ±0.006	0.434 ±0.005
SLDNet ($L = 4$)	0.412 ±0.010	0.326 ±0.009	**0.980 ±0.008**	0.453 ±0.005
iForest	0.447 ±0.011	**0.367 ±0.022**	0.942 ±0.009	0.436 ±0.007
MCSVDD	0.464 ±0.051	0.312 ±0.027	0.675 ±0.168	0.459 ±0.028
LOF	0.486 ±0.009	0.145 ±0.005	0.504 ±0.039	0.451 ±0.008
OC-SVM	0.481 ±0.043	0.090 ±0.004	0.880 ±0.057	0.412 ±0.004
Autoencoder	0.317 ±0.001	0.219 ±0.004	0.288 ±0.000	0.455 ±0.004

VI. Analysis and Limitations

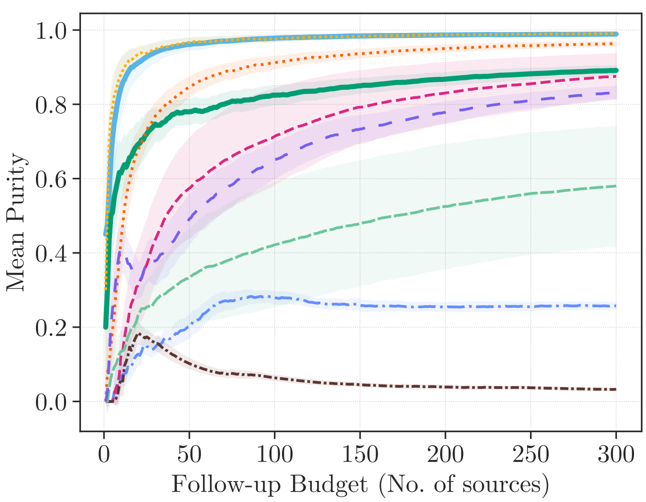
Results are consistent with the expected behavior of one-class methods under heterogeneous variability families. In transients and stochastic objects, learning a multiscale potential allows SLDNet to capture dispersed modes and low-density regions flexibly. This capability is crucial for modeling complex manifolds where the boundary between normality and



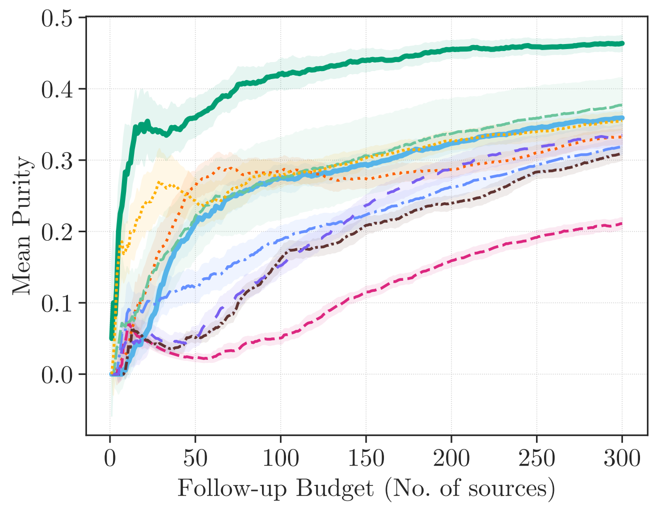
(a) ELAsTiCC OOD



(b) ELAsTiCC RID



(c) ZTF OOD



(d) ZTF RID

Figure 2: Purity vs. follow-up budget N (number of top-ranked sources selected for follow-up) for OOD and RID discovery on ELAsTiCC and ZTF. SLDNet is compared with baselines and DiMMAD [8]. Higher purity at low budgets indicates better prioritization under limited resources.

anomaly is not rigid, outperforming partition-based methods (iForest) or simple geometric boundaries (OC-SVM) in the presence of photometric noise (Table I).

In the case of periodic objects, tabular features usually form compact and well-separated structures in feature space. Here, methods explicitly imposing a compact cluster structure, like MCSVDD's multiple hyperspheres, may present inductive advantages over a smooth density estimator like SLDNet (Table I). However, our method maintains notable competitiveness, suggesting that the learned potential manages to fit these topologies adequately without collapsing.

The contrast between AUPRC performance for OOD and RID reinforces the complementarity of approaches (See Table II and Table III; Fig. 2 and Fig. 3). While DiMMAD, with its robust distance ensemble, positions itself as a solid reference for discovering new classes (OOD) [8], SLDNet offers superior performance in detecting rare variants (RID) in the real ZTF data. This robustness in ZTF can be attributed to the fact that training with noise (*denoising*) confers the model certain immunity against observational artifacts that could confuse rigid distance metrics. From a methodological perspective, SLDNet avoids dependence on the underlying distance metric (Euclidean, Cosine, etc.) by learning the intrinsic metric of the data density. This property is theoretically attractive for *domain shift* scenarios, although validating it will require future experiments with simulated versus real data.

Among limitations, it is worth noting that SLDNet operates on precomputed tabular features. While this allows fair com-

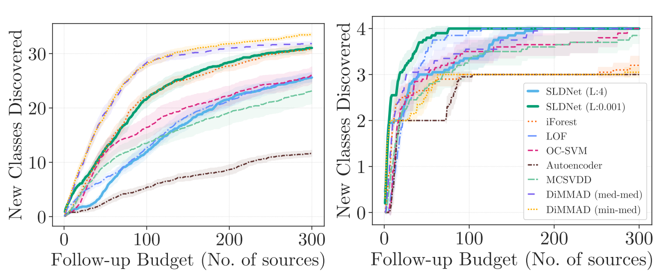
Table III: Count of AUPRC *wins* per OOD/RID scheme (E = ELAsTiCC, Z = ZTF). For each scheme, all methods are compared against each other via paired permutation test on AUPRC per fold. If $p \leq 0.05$, the method with higher average AUPRC receives a win; if $p > 0.05$ or tie, no wins are assigned. OOD and RID are the sum of the two corresponding schemes; Total corresponds to the sum of all four.

Model	E-OOD	E-RID	Z-OOD	Z-RID	OOD	RID	Total
SLDNet ($L = 0.001$)	6	2	4	**8**	10	**10**	**20**
DiMMAD (med-med)	**8**	7	3	1	11	8	19
DiMMAD (min-med)	6	4	7	2	**13**	6	19
SLDNet ($L = 4$)	1	5	**8**	4	9	9	18
iForest	2	**8**	6	2	8	**10**	18
MCSVDD	2	5	2	4	4	9	13
LOF	3	1	1	4	4	5	9
OC-SVM	3	0	4	0	7	0	7
Autoencoder	0	3	0	4	0	7	7

parison with the current state of the art (ALeRCE, DiMMAD), it ignores richness of sequential temporal information that models like recurrent networks or Transformers could exploit directly. Additionally, while we used a simple MLP potential to keep inference lightweight, exploring alternative architectures for tabular data (e.g., ATAT [18]) could further improve representation capacity. Finally, inference computational cost, though mitigated by the MLP architecture, scales with the number of gradient evaluations, that can be a bottleneck for real-time applications of more complex models.

VII. Conclusions

We proposed SLDNet, an anomaly detector based on *score matching* that learns a multiscale conservative potential as a proxy for negative log-density, delivering an anomaly score interpretable in probabilistic terms. Evaluated on ZTF tabular features calculated by ALeRCE, SLDNet outperforms classic baselines and is competitive against deep methods, with notable improvements in transients and stochastic objects, and competitive results in various periodic subclasses. In DiMMAD OOD/RID discovery scenarios, SLDNet achieves competitive AUPRC in OOD and strong performance in RID, yielding high purity at low follow-up budgets.



(a) ELAsTiCC OOD (b) ZTF OOD

Figure 3: Diversity (unique anomalous classes recovered) curves as a function of follow-up budget N (up to $N = 300$) for ELAsTiCC and ZTF OOD scenarios. We do not show RID cases because diversity saturates quickly and space is limited.

As future work, it is relevant to (i) report full comparisons per subclass and sensitivity analysis to hyperparameters/scale, (ii) evaluate robustness to domain shift towards simulated datasets or future surveys, (iii) explore the use of classifier latent space for anomaly detection under the same evaluation schemes, and (iv) study alternative tabular architectures (e.g., ATAT [18]) as a replacement for the MLP.

Acknowledgment

The authors acknowledge the ALeRCE team for providing the ZTF labeled dataset, which enabled this work. We also thank the authors of DiMMAD for providing the data used to evaluate their benchmarks.

References

[1] E. C. Bellm *et al.*, "The zwicky transient facility: System overview, performance, and first results," *Publications of the Astronomical Society of the Pacific*, vol. 131, no. 995, p. 018002, Jan. 2019.
[2] Ž. Ivezić *et al.*, "Lsst: From science drivers to reference design and anticipated data products," *The Astrophysical Journal*, vol. 873, no. 2, p. 111, Mar. 2019.
[3] F. Forster, G. Cabrera-Vives, E. Castillo-Navarrete *et al.*, "The automatic learning for the rapid classification of events (alerce) alert broker," *The Astronomical Journal*, vol. 161, no. 5, p. 242, May 2021.
[4] A. Hyvärinen, "Estimation of non-normalized statistical models by score matching," *Journal of Machine Learning Research*, vol. 6, pp. 695–709, 2005.
[5] J. Micorek, H. Possegger, D. Narnhofer *et al.*, "Mulde: Multiscale log-density estimation via denoising score matching for video anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 18 868–18 877.
[6] M. Pérez-Carrasco *et al.*, "Alert classification for the alerce broker system: The anomaly detector," *The Astronomical Journal*, vol. 166, no. 4, p. 151, Oct. 2023.
[7] ——, "Multi-class deep svdd: Anomaly detection approach in astronomy with distinct inlier categories," in *2nd ICML Workshop on Machine Learning for Astrophysics*, ser. Proceedings of Machine Learning Research, vol. 202, 2023. [Online]. Available: https://ml4astro.github.io/icml2023/assets/19.pdf
[8] S. Chaini, F. B. Bianco, and A. Mahabal, "In search of the unknown unknowns: A multi-metric distance ensemble for out of distribution anomaly detection in astronomical surveys," in *NeurIPS 2025 Workshop on Machine Learning and the Physical Sciences*, 2025.
[9] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 15:1–15:58, 2009.
[10] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 Eighth IEEE International Conference on Data Mining*, Pisa, Italy, 2008, pp. 413–422.
[11] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* Cambridge, MA, USA: MIT Press, 2001.
[12] D. M. J. Tax and R. P. W. Duin, "Support vector data description," *Machine Learning*, vol. 54, no. 1, pp. 45–66, 2004.
[13] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.
[14] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations, ICLR 2014*, Banff, AB, Canada, 2014.
[15] L. Ruff, R. Vandermeulen, N. Goernitz *et al.*, "Deep one-class classification," in *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweden, 2018, pp. 4393–4402.
[16] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural Computation*, vol. 23, no. 7, pp. 1661–1674, 2011.
[17] P. Sánchez-Sáez *et al.*, "Alert classification for the alerce broker system: The light curve classifier," *The Astronomical Journal*, vol. 161, no. 3, p. 141, Mar. 2021.
[18] G. Cabrera-Vives, D. Moreno-Cartagena, N. Astorga *et al.*, "Atat: Astronomical transformer for time-series and tabular data," *Astronomy & Astrophysics*, vol. 689, p. A289, 2024.