Machine Learning Applications for the LSST Data

Korea Astronomy and Space Science Institute

Min-Su Shin
Korea Astronomy and Space Science Institute
msshin@kasi.re.kr

Legacy Survey of Space and Time

ABSTRACT

We present examples of using machine learning (ML) algorithms in the LSST data era. First, our models of inferring photometric redshifts for LSST galaxies handle biased training spectroscopic data with methods finding out-of-distribution test data and measuring influence of training samples. Second, we also develop a machine learning method of classifying galaxies morphologically in Hubble sequence, focusing on semi-supervised approaches for the expected large number of unclassified LSST galaxies. Third, our research on asteroid taxonomy uses both semi-supervised and unsupervised learning methods to fully understand population of new asteroid taxonomy types hidden in the LSST big data on asteroids.

Why ML inference?

The scale of the LSST big data (see the LSST document LSE-163):

Prompt (Previously "Level 1" data products)	Real-time difference image analysis (DIA). A stream of ~10⁶ time-domain events per night (Alerts), detected, characterized, and distributed within 60 seconds. A catalog of orbits for ~6 million bodies in the Solar System.
Data Release (Previously "Level 2" data products)	Processed single-epoch and deep co-added images, and reprocessed DIA products. A database of ~7x10¹² detections (~30x10¹² measurements) for ~37x10⁹ objects (20x10⁹ galaxies and 17x10⁹ stars), produced annually and accessible online.
User Generated (Previously "Level 3" data products)	User-produced added-value data products, e.g., deep KBO/NEO catalogs, variable star classifications, shear maps, etc. Enabled by services and computing resources at the Data Access Centers and via the LSST Science Platform.

ML inference

- Efficient ways to handle the big data with quick inference even though there are many issues and costs of training models.
- Difficult problems of acquiring right training data and training models in a right way.
- Find the out-of-distribution (OOD) data and consider (Bayesian) statistical inference for them.

Statistical inference

- Bayesian inference with MCMC is a right way to estimate asymptotically correct posterior distributions.
- MCMC and variational inference can be adopted together in addition to ML inference.
- When requiring fast statistical inference, consider variational inference.

Extremely fast ML inference is possible in the LSST big data era!

Asteroid taxonomy

Reliable classification of asteroids and discovery of unusual objects in terms of asteroid taxonomy (Roh et al. 2022; Choi et al. 2023).

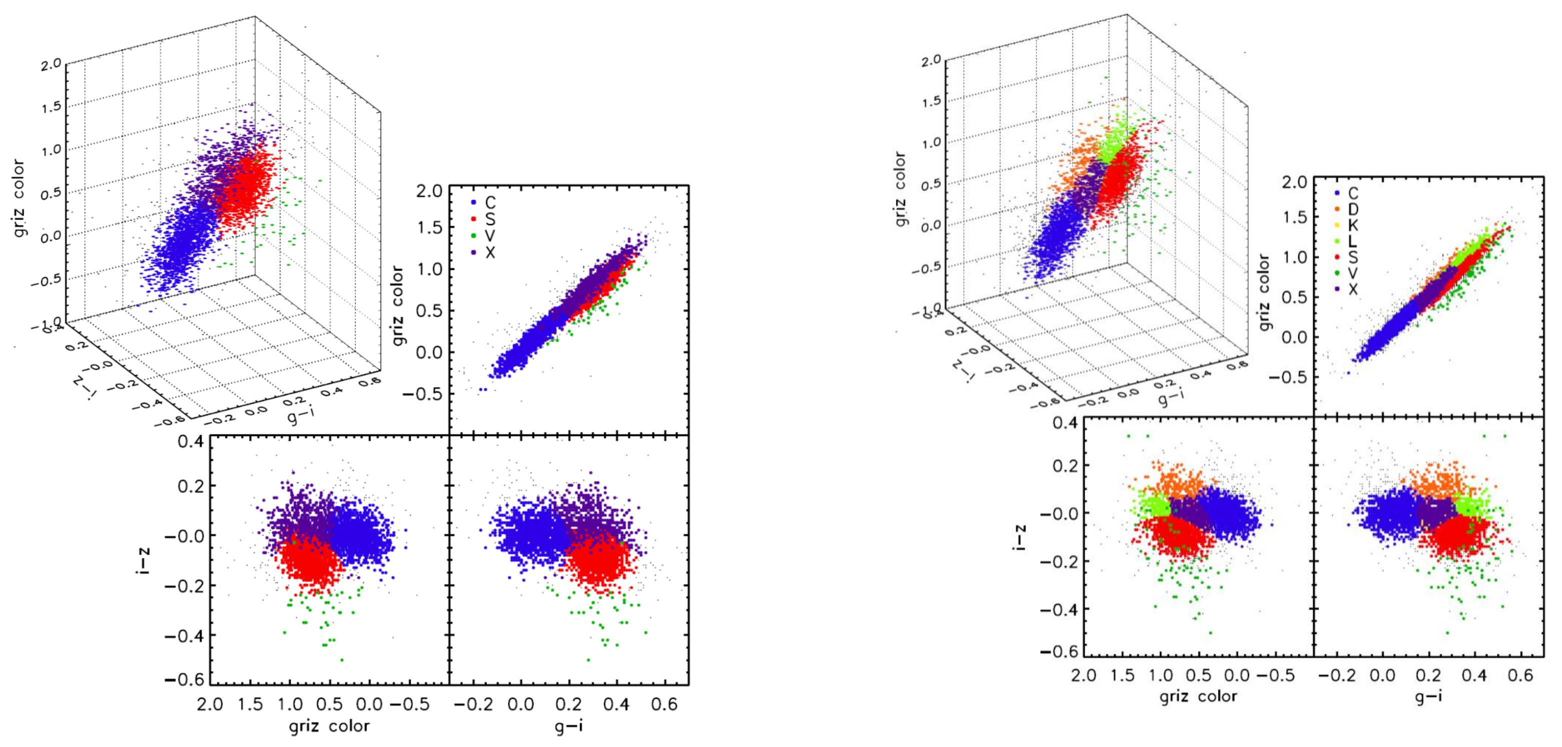
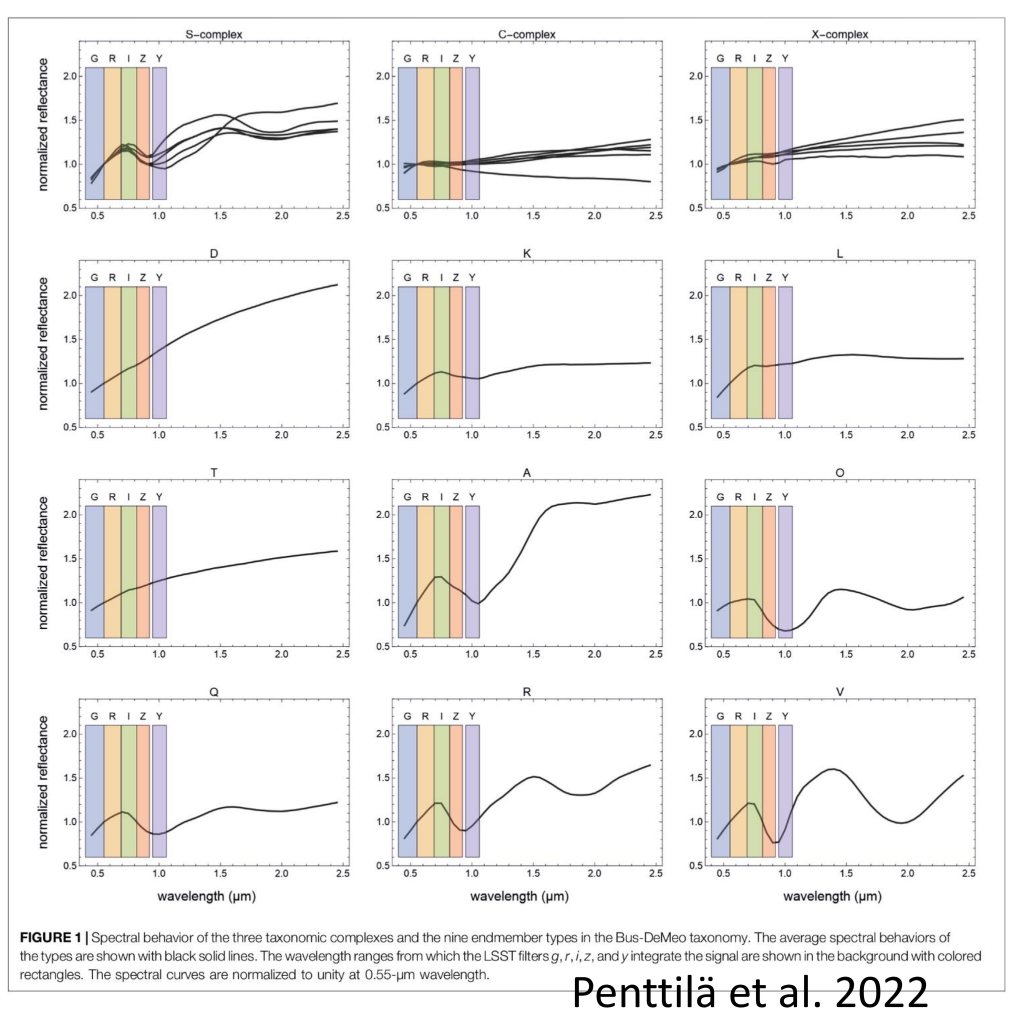


Figure. Taxonomy classification in unsupervised (left) and semi-supervised learning (right) (Roh et al. 2022).



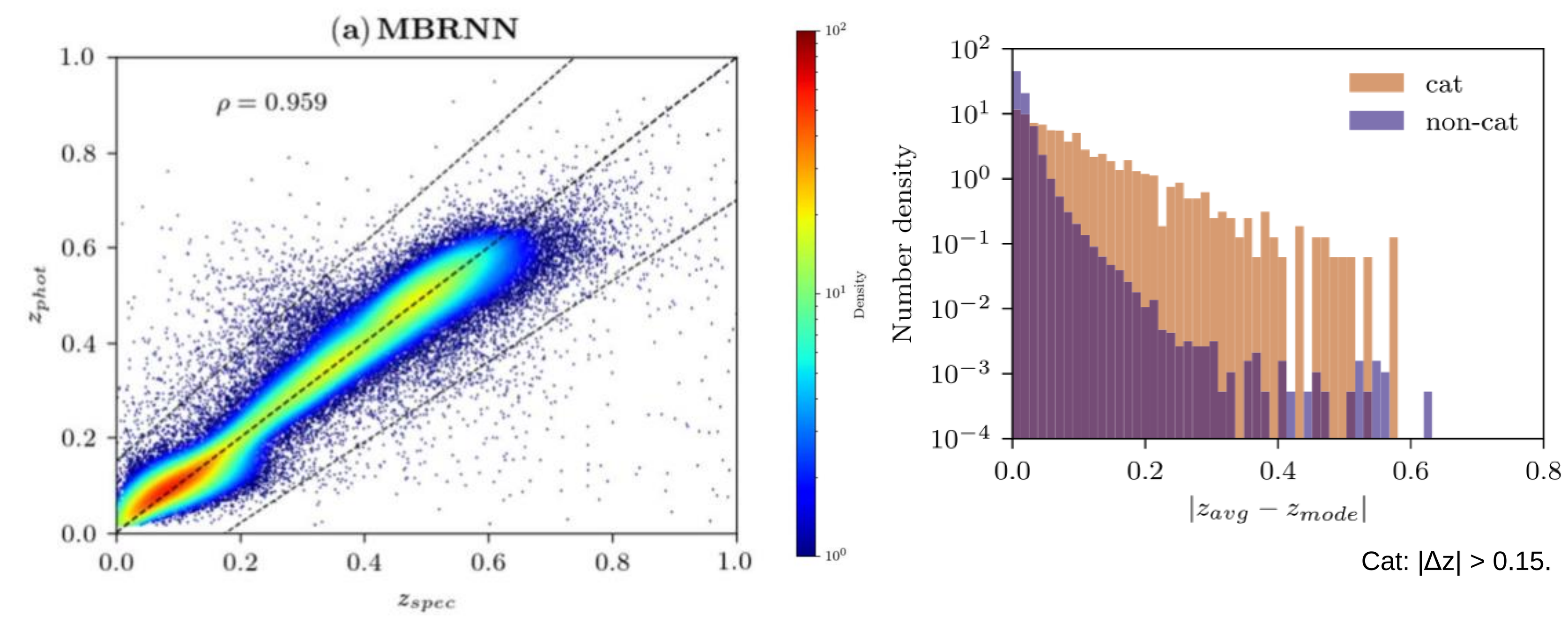
Penttilä et al. 2022

Expected increase of unlabeled data in the LSST.
→ **Importance of semi-supervised and unsupervised learning**.

Usage of rich information in the more LSST bands than the SDSS bands.
→ Requirements for better data and models.
→ Identification of important training samples requiring labeling (i.e., spectroscopy).

Photo-z estimation of galaxies

Estimating photometric redshifts of Pan-STARRS galaxies with multiple-bin regression with neural networks (MBRNN) for potential applications in the LSST era (see Lee & Shin 2021).

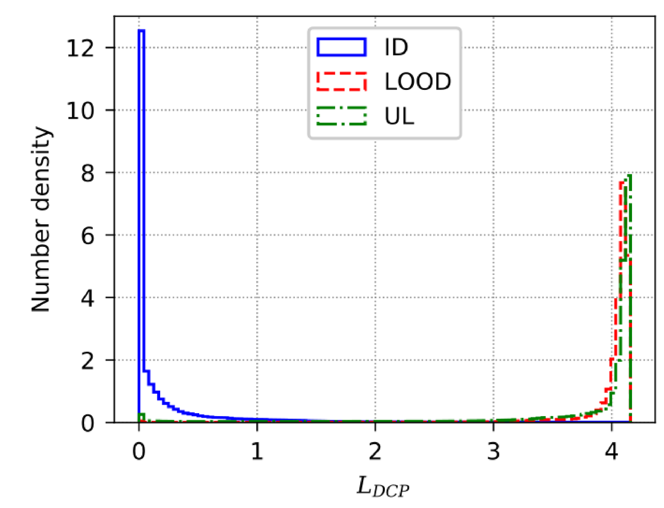


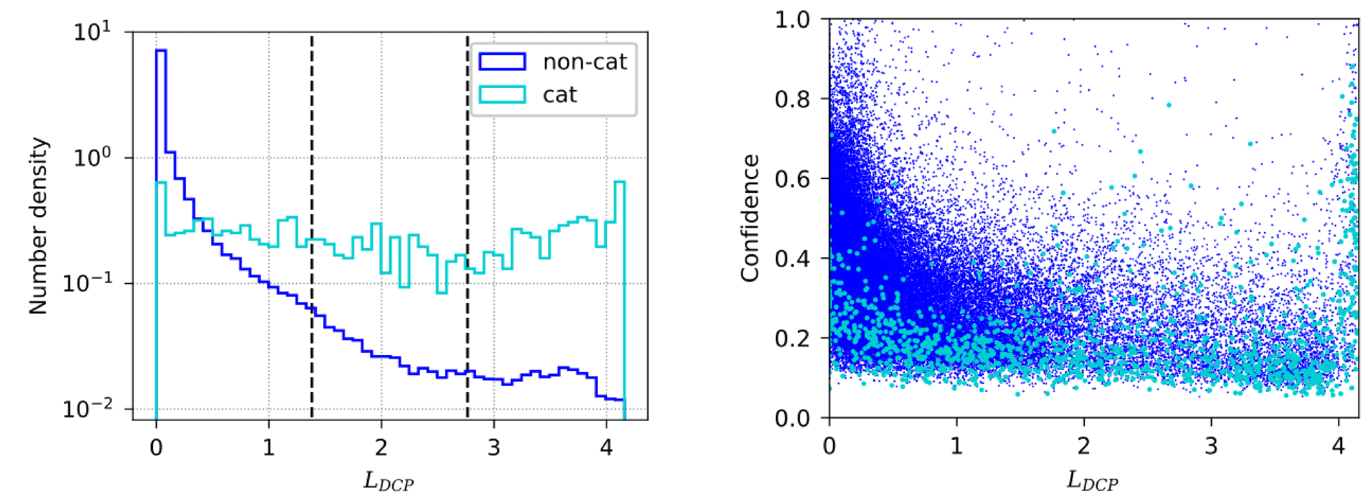
Scoring the OOD likelihood of test galaxies to find objects requiring time-consuming statistical analysis (see Lee & Shin 2022).

L_{DCP}: scores of OOD likelihood.

ID: in-distribution data,
LOOD: labeled OOD (stars and quasars),
UL: unlabeled data.

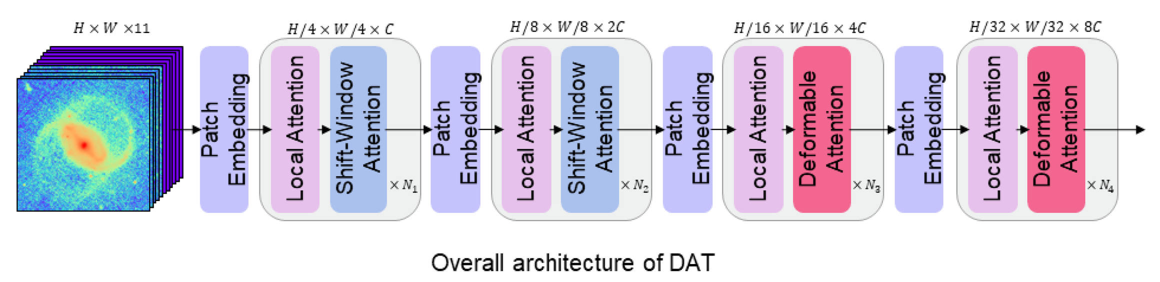
cat: cases of catastrophic error.





Morphological classification of galaxies

Fine-level morphological classification in terms of the Hubble sequence by using deformable attention transformer (DAT) (Kang et al., ML4PS, NeurIPS, 2022).



- DAT[1] is implemented based on Swin-Transformer.
- DAT use deformable attention which captures more important tokens.
- Use 11-dimensional high dynamic range image, which consists of 5 Raw Galaxy Images, 1 Galaxy Mask, 5 Nan Value Masks

Figure. DAT model using multiple 2D data made of 480 x 480 pixels: galaxy images in five bands, single object mask per galaxy, and nan-value mask per band.

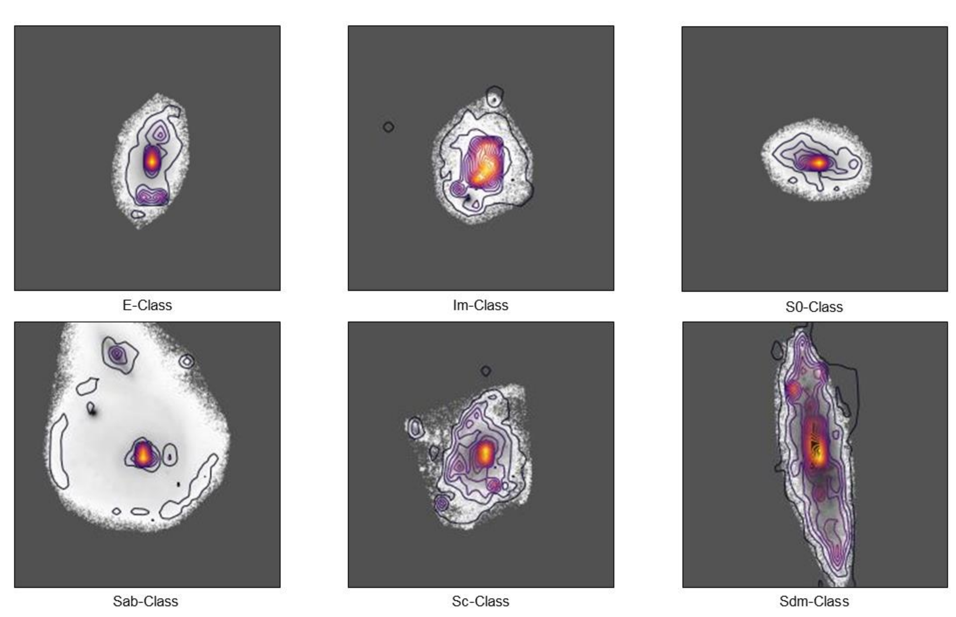


Figure. Distribution of the last layer's attention as contours in the example galaxy for each class with r-band images.

Expected increase of unlabeled data
+
Imbalance of labeled training samples and covariate shift between training data and test data in the LSST era.
→ **Importance of semi-supervised learning and difficulty-based learning strategy.**

If you are interested in ML applications with the LSST data, please, contact M.-S. Shin for possible collaboration.

VERA C. RUBIN OBSERVATORY