Improving and Assessing Astronomical Light Curve Classifiers with Classification Histories

Zhuoyang Zhou
Department of Statistics & Data Science
Carnegie Mellon University
grantz@andrew.cmu.edu

Chad M. Schafer
Department of Statistics & Data Science
McWilliams Center for Cosmology and Astrophysics
Carnegie Mellon University

Alex I. Malz
Space Telescope Science Institute and
Department of Physics, McWilliams Center for Cosmology and Astrophysics
Carnegie Mellon University

Abstract

The Legacy Survey of Space and Time (LSST) will generate a massive collection of time series (light curves) of the measured flux of transient and variable astronomical objects and classification metadata. Each new observation of a detected source will lead to an updated probability distribution over candidate classes, which will then be provided to the global community for the purpose of identifying interesting targets for prompt follow-up observations as well as less time-sensitive analyses. Using the synthetic light curves and classification results from the Extended LSST Astronomical Time-series Classification Challenge (ELAsTiCC), we propose a novel framework to enhance existing classifiers by incorporating the temporal evolution of their historical classifications. To demonstrate this, we investigate a new classifier that combines a Long Short-Term Memory and an additive attention module. The new classifier shows higher classification accuracy and more balanced precision-recall performance across all classes compared to existing classifiers in the challenge. We also propose new performance metrics that can better evaluate the model's classification stability and early classification ability with partial observations. This offers a more comprehensive perspective for model assessment by supplementing classical methods such as the confusion matrix and precision-recall.

1 Introduction

The Legacy Survey of Space and Time (LSST) [1] on the Vera C. Rubin Observatory will repeatedly scan the night sky for ten years and generate time series (light curves) of measured flux observed in six passbands for detected transient and variable objects [2, 3]. Accurate, reliable, and timely classification of these light curves is necessary so that scientists worldwide can select specific targets of interest and conduct detailed follow-up observations.

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

Existing classifiers either directly use the raw time series, as light curves or images, for deep learning models [4, 5, 6, 7] or extract informative features for training other machine learning models such as random forest or boosted decision trees [8, 9, 10, 11]. The typical classification scheme can be characterized as follows: Whenever a new observation of a detected variable or transient object is made, the classifiers refit the enlarged full light curve and update the classification PMF. Such classification approaches ignore the entire historical classification results with partial light curves, which may contain important features that can help us further differentiate easily-confused object types and rectify systematic bias from the final classification results.

Using the synthetic data and participating classifiers from the Extended LSST Astronomical Time-series Classification Challenge (ELAsTiCC) [12], we propose a new model based on Long Short-Term Memory [13] and an additive attention mechanism [14] to improve the existing classifiers by incorporating both the raw light curves and the classifiers' historical classifications. We evaluate the new model by selecting three representative classifiers from ELAsTiCC, focusing on the five common supernova classes. The new models show higher classification accuracy and much more balanced precision-recall performance compared with all selected classifiers.

As for model evaluation, while we have classifications for each new observation, most, if not all, of the existing classifiers are mainly evaluated by the final classification probabilities obtained with full light curves using classical metrics like the precision-recall, accuracy, and confusion matrix. Such measures fail to evaluate the model's early classification performance and classification stability, which can be crucial for quick identification of objects of interest and reliable follow-up decision-making. To address such limitations, we propose two novel Early-Stable Classification Metrics based on Wasserstein distances between classification PMFs and the temporal evolution of the classification PMFs. We demonstrate some potential applications of the metrics by applying them to both the new and baseline models.

2 Dataset and Analysis Set-up

Synthetic data and classifications are obtained from the ELAsTiCC2 challenge, which contains simulated light curves that mimic the real LSST observation scheme for 4.1 million transient and variable stars [12]. More detailed documentation on the data generation, classification taxonomy, and original training data can be found in [12]. We select the five common supernova classes from the original 25 classes of objects, including SN Ia, SN Ib/c, SN II, SN Iax, and SN 91bg. Four alert brokers, which are automated software systems and classifiers designed to process, characterize, and prioritize alerts from the LSST and other surveys for follow-up, participated in ELAsTiCC2, including ALeRCE [11], Pitt-Google [15], ANTARES [10], and AMPEL [16]. We chose three representative classifiers with relatively high classification coverage and reasonably good classification performance compared with the remaining candidates. Due to anonymity requirements, we cannot provide identification information for the classifiers and simply refer to them as classifiers A, B, and C, where classifiers A and B are from the same alert broker and classifier C is from another group.

Since classifiers only made classifications on a subset of detected sources due to the different classification scheme chosen by classifiers, we only include objects for which the classifier made a valid classification. For each object, we have a time series of synthetic flux, classification PMFs over the five classes, and observation timestamps. Other features, including the corresponding passband of the flux and flux-to-error ratio, will not be used for modeling. Classifier C has different coverage than that of classifiers A and B, resulting in classifications on a different set of objects. The number of observations per object ranges from a few to more than a thousand. To enable the study of temporal evolutions and to prevent objects in the same training batch from differing drastically in length, we only select objects with at least 15 classifications and apply truncation with a maximum length of 100. One could use different lower and upper cuts for a more holistic view.

For most objects, observations are irregularly spaced, including some with large gaps with no measurements, due to the survey's observation plan. To account for the highly irregular time series, we compute four absolute and relative time encodings based on the observation history for each object: the raw timestamps, time since the first classification, time since the previous classification, and time to the next classification. These time encodings are normalized to the range $[0, 1]$ and directly concatenated to the existing time series. The resulting data take the form of a 10-dimensional multivariate time series, with five channels of classification probabilities, one channel for the raw flux, and four time encoding channels at each timestamp. We use these multivariate time series directly without further feature extraction or augmentation, aiming to retain the full raw information.

3 Methods

3.1 Model Architecture

We propose a three-stage model that combines a recurrent network and attention mechanisms. In the first stage, the varying-length multivariate time series of raw flux, classification PMFs, and time encodings are preprocessed through masking and padding per batch. The inputs are fed into a single-layer LSTM with the standard implementation by PyTorch [17], generating hidden state representations per timestamp that encode temporal dependencies across the observations. For the second stage, we apply an additive attention mechanism to the sequence of hidden states, allowing the model to selectively weights the importance of different time steps and aggregate them for the final classifications. Instead of using the more standard dot-product attention [18], we use an additive attention mechanism (Bahdanau attention) [14], which computes the scoring functions using a feed-forward network with a single hidden layer and a nonlinear $\tanh$ activation function. The main goal is to capture non-linear and more complicated relationships among hidden states. Finally, the attention-weighted hidden states are passed through three-layer fully connected networks with ReLU activation to produce the final classifications.

We acknowledge that the combination of standard LSTM and an additive attention does not represent the state-of-the-art models handling irregular time series. More advanced approaches ([6, 19, 7]) have been developed for irregular temporal data. However, the primary goal of our work is not to propose a novel model architecture, but rather to demonstrate the value of incorporating historical classifications into light curve classification. Future work could explore integrating our approach with more advanced architectures.

3.2 New Metric for Model Evaluation

We propose two performance metrics that take into account classification accuracy, stability, and early classification performance with the 1-Wasserstein distance ([20, 21, 22]) for stability quantification. We name the two metrics as Early-Stable Classification Fraction (ESC-f) and the Early-Stable Classification Score (ESC-s). We define the Early-Stable Classification Fraction (ESC-f) as follows:

Definition 1 (Early-Stable Classification Fraction) *Given a sequence of classification PMFs for a n-class classification task, $\{p_t\}_{t=1}^T$, where T is the total number of observations, the classification converges (ϵ, ρ, k)-fast with convergence time τ, $1 \leq \tau \leq T$, if both conditions are satisfied:*

Stability Condition*: There exist k consecutive observations, $k \leq T$ and $k \geq 2$, such that*

$$W_1(p_i, p_{i+1}) \leq \epsilon \quad \forall i \in \{\tau - k + 1, t - k + 2, \ldots, \tau - 1\},$$

for some fixed $\epsilon \in (0, 1)$ and $W_1(\cdot, \cdot)$ is the 1-Wasserstein distance computed for the two PMFs with a cost matrix that is selected to have a unit cost of moving unit mass between any distinct classes.

Accuracy Condition*:*

$$p_{i,*} \geq \rho \quad \forall i \in \{\tau - k + 1, t - k + 2, \ldots, \tau\},$$

for some fixed $\rho \in (0, 1)$, where $p_{i,*}$ denotes the classification probability for the true class at time i. Then, we define the Early-Stable Classification Fraction (ESC-f) for the object as:

$$ESC_f = \tau / T \textit{ if converged, or } 1 \textit{ otherwise} \tag{1}$$

Ideal classifiers should have a higher proportion of objects that converged per class and a smaller convergence fraction per object. We purposefully select the cost matrix to have a unit transport cost of moving a unit mass between classes for standardization. Thus, for randomly generated PMFs, the Wasserstein distances are between $[0, 1]$. Choices for ϵ could be 0.05 for a stricter requirement and 0.1 for a relatively relaxed tolerance. The choice of ρ and k is more application-specific, with larger values corresponding to stricter requirements.

As for the Early-Stable Classification Score (ESC-s), we define it as follows:

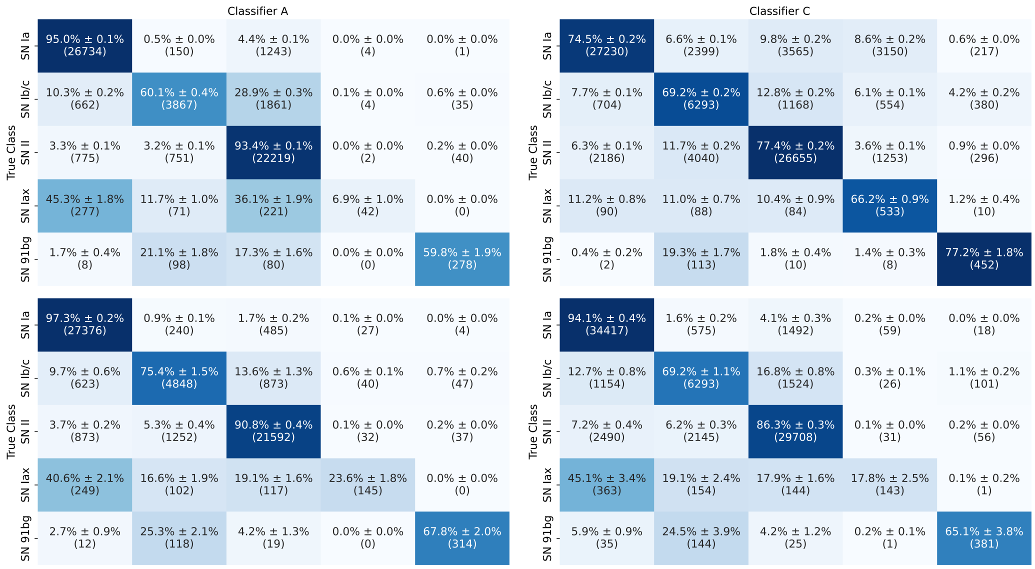

Figure 1: The comparisons between the baseline (upper) and new (bottom) models for classifier A and C. The confusion matrix is normalized per row and annotated with average absolute counts. The new models show improvements in overall accuracy and more balanced precision-recall.

Definition 2 (Early-Stable Classification Score) *For the same sequence defined above* $\{p_t\}_{t=1}^{T}$, *we define a weight function* $w : [0,1] \to [0,1]$. *Then, we have:*

$$ESC_s = -\frac{1}{Z \cdot (T-1)} \sum_{t=2}^{T} w\left(\frac{t}{T}\right) \cdot W_1(p_{t-1}, p_t) \cdot \ln(p_{t,*}), \tag{2}$$

where $Z = \sum_{t=2}^{T} w(\frac{t}{T}) \cdot W_1(p_{t-1}, p_t)$ *is the sum of weights for normalization and* $p_{t,*}$ *is the classification probability for the true class at time* t.

4 Experiments and Results

The proposed model is implemented with PyTorch and trained with a cross-entropy loss and an Adam optimizer [17, 23]. We fixed the hyperparameters to 256 hidden state dimensions for the single-layer LSTM and 64 dimensions for the additive attention module. A dropout of 0.4 is used for the final classification layers. We apply a 70-30 train-test split, and 30% of the training set is held out as the validation set. Models were trained for up to 200 epochs with early stopping (patience=20) based on classification accuracy on the validation set. The final model was selected to be the one with the highest validation accuracy. As for the baseline model, the classification results on the test set are obtained by using the most recent classification PMFs for each object, with the final label assigned to be the class with the highest classification probability. To ensure a more robust comparison, we repeated the training process 10 times using different random seeds for the train-test split. All reported evaluation metrics and confusion matrices represent the mean values across these 10 iterations, with $\pm$ 1 standard deviation.

In Table 1, the new models outperform the baseline models across all three classifiers in terms of accuracy, weighted precision, and F1 score. For classifiers A and B, the new models achieve accuracy improvements of around 2% and 4.5% over the baseline, respectively. The new model shows a more substantial improvement in accuracy on classifier C, with accuracy increasing from 0.751 (baseline) to 0.871 (new). We next examine per-class performance with confusion matrices presented in Figure 1. Since classifier B originates from the same alert broker as classifier A and exhibits similar characteristics, we mainly focus on classifiers A and C. For the baseline classifier A, we observe higher recall for SN Ia and SN II, with relatively lower recall for SN Ib/c and SN 91bg, and substantially poorer performance for SN Iax. While the two minority classes achieve relatively high precision, their recall remains low. The new model improves recall across all classes, with a

Table 1: Comparisons of Classical and Early-Stable Classification Metrics

Metric	Classifier	Models: Baseline	Models: New
Accuracy	A	0.894 ± 0.00059	**0.913 ± 0.0015**
	B	0.869 ± 0.0066	**0.914 ± 0.00098**
	C	0.751 ± 0.0014	**0.871 ± 0.0012**
Weighted Precision	A	0.891 ± 0.00071	**0.911 ± 0.0017**
	B	0.886 ± 0.0066	**0.913 ± 0.001**
	C	0.820 ± 0.0013	**0.869 ± 0.0013**
Weighted F1 Score	A	0.888 ± 0.00058	**0.911 ± 0.0015**
	B	0.874 ± 0.0015	**0.913 ± 0.001**
	C	0.777 ± 0.0013	**0.868 ± 0.0013**
Convergence Proportion	A	0.828 ± 0.001	**0.873 ± 0.0016**
	B	0.792 ± 0.0017	**0.872 ± 0.0031**
	C	0.398 ± 0.0016	**0.751 ± 0.0048**
Early-Stable Classification Fraction	A	0.483 ± 0.00073	**0.451 ± 0.0076**
	B	0.618 ± 0.00078	**0.474 ± 0.011**
	C	0.843 ± 0.00077	**0.655 ± 0.0093**

small reduction for SN II. However, despite achieving a higher overall recall, the baseline classifier A exhibits a substantial false positive rate for SN II, misclassifying 28.9% of SN Ia objects and 4.4% of SN Ib/c objects as SN II, which makes it the dominant misclassification for both majority classes. Such poor precision poses practical concerns, potentially leading to misallocation of observational resources and undermining the model's reliability in real-world applications. The proposed model effectively addresses these limitations, achieving an improved recall-precision trade-off that enhances overall recall without substantial loss of precision.

For classifier C, the baseline model exhibits relatively poorer baseline performance compared to classifiers A and B. The recall is more uniform across all classes, including the two minority classes. The proposed model improves or maintains recall for all majority classes (with SN Ib/c remaining unchanged), achieving gains of approximately 20 and 10 percentage points for SN Ia and SN II, respectively, while maintaining satisfactory precision. In contrast, recall for both minority classes decreases under the proposed model. However, this follows the same recall-precision trade-off, analogous to the SN II behavior observed in classifier A. More specifically, while the baseline classifier C achieves 66.2% recall (533 true positives) for SN Iax, the cost is 4,965 false positives, making the high recall practically meaningless due to severely degraded precision. The high recall becomes less meaningful under such low precision. A similar pattern is observed in the SN 91bg class, which has 452 true positives but incurs 903 false positives. The new model substantially reduces the number of false positives and achieves a more balanced precision-recall performance.

As for the assessment with the new metrics, the convergence parameters are fixed at $\epsilon = 0.1, \rho = 0.5, k = 5$ for all evaluations. For baseline models, we directly use the existing historical classification PMFs for evaluation. For the new models, we construct classification histories of test objects through a sequential truncation procedure: for each time step $t \in \{2, \dots, T\}$, we truncate the input time series of classification PMFs and raw flux to the first t observations and apply the model to obtain the classification PMF at time t. This yields a complete temporal sequence of classification PMFs for each object. For each object class, we compute the proportion of objects achieving convergence under the specified criteria and the mean ESC-f across all objects within the class. These class-level metrics are then aggregated via a weighted average, with weights proportional to class sample sizes. We summarize the results in Table 1 for the classifiers A, B, and C and all three model types. As shown in Table 1, the proposed model achieves higher convergence proportions compared to the baseline models. The improvements range from approximately 5% – 8% for classifiers A and B, and a substantial 35% for classifier C, indicating that the new model yields more objects with stable classifications under the accuracy threshold $\rho = 0.5$. As suggested by the Early-Stable Classification fractions, the new model achieves stable and accurate classifications at earlier stages of the observation compared with the baseline models. The improvement is relatively small for classifier A and more substantial for classifiers B and C.

5 Acknowledgements

This work was supported by Schmidt Sciences. We thank Konstantin Malanchev for his help with processing the Extended LSST Astronomical Time-series Classification Challenge (ELAsTiCC) data. This work used Bridges-2 at Pittsburgh Supercomputing Center through allocation [PHY250104, PHY210095] from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program [24], which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. We thank the reviewer for helpful comments and suggestions.

References

[1] Paul A Abell, Julius Allison, Scott F Anderson, John R Andrew, J Roger P Angel, Lee Armus, David Arnett, SJ Asztalos, Tim S Axelrod, Stephen Bailey, et al. Lsst science book, version 2.0. 2009.

[2] Željko Ivezić, Steven M Kahn, J Anthony Tyson, Bob Abel, Emily Acosta, Robyn Allsman, David Alonso, Yusra AlSayyad, Scott F Anderson, John Andrew, et al. Lsst: from science drivers to reference design and anticipated data products. *The Astrophysical Journal*, 873(2):111, 2019.

[3] R Hložek, AI Malz, KA Ponder, M Dai, G Narayan, EEO Ishida, T Allam Jr, A Bahmanyar, X Bi, R Biswas, et al. Results of the photometric lsst astronomical time-series classification challenge (plasticc). *The Astrophysical Journal Supplement Series*, 267(2):25, 2023.

[4] Kaiming Cui, DJ Armstrong, and Fabo Feng. Identifying light-curve signals with a deep-learning-based object detection algorithm. ii. a general light-curve classification framework. *The Astrophysical Journal Supplement Series*, 274(2):29, 2024.

[5] Johanna Pasquet, Jérôme Pasquet, Marc Chaumont, and Dominique Fouchez. Pelican: deep architecture for the light curve analysis. *Astronomy & Astrophysics*, 627:A21, 2019.

[6] Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. *Advances in neural information processing systems*, 33:6696–6707, 2020.

[7] Satya Narayan Shukla and Benjamin M Marlin. Multi-time attention networks for irregularly sampled time series. *arXiv preprint arXiv:2101.10318*, 2021.

[8] Kyle Boone. Avocado: Photometric classification of astronomical transients with gaussian process augmentation. *The Astronomical Journal*, 158(6):257, 2019.

[9] Kaylee M de Soto, V Ashley Villar, Edo Berger, Sebastian Gomez, Griffin Hosseinzadeh, Doug Branton, Sandro Campos, Melissa DeLucchi, Jeremy Kubica, Olivia Lynn, et al. Superphot+: Real-time fitting and classification of supernova light curves. *The Astrophysical Journal*, 974(2):169, 2024.

[10] Thomas Matheson, Carl Stubens, Nicholas Wolf, Chien-Hsiu Lee, Gautham Narayan, Abhijit Saha, Adam Scott, Monika Soraisam, Adam S Bolton, Benjamin Hauger, et al. The antares astronomical time-domain event broker. *The Astronomical Journal*, 161(3):107, 2021.

[11] P Sánchez-Sáez, I Reyes, C Valenzuela, F Förster, S Eyheramendy, F Elorrieta, FE Bauer, G Cabrera-Vives, PA Estévez, Márcio Catelan, et al. Alert classification for the alerce broker system: The light curve classifier. *The Astronomical Journal*, 161(3):141, 2021.

[12] ELAsTiCC Team, et al. The desc elasticc challenge. `https://portal.nersc.gov/cfs/lsst/DESC_TD_PUBLIC/ELASTICC/`, 2023.

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[14] Dzmitry Bahdanau. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[15] Pitt-Google Broker. Pitt-google broker. `https://pitt-broker.readthedocs.io/en/latest/`, 2025.

[16] Jakob Nordin, Valery Brinnel, Jakob van Santen, Simeon Reusch, and Marek Kowalski. Ampel workflows for lsst: Modular and reproducible real-time photometric classification. *Astronomy & Astrophysics*, 698:A13, 2025.

[17] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[19] Yuan Zhang. Attain: Attention-based time-aware lstm networks for disease progression modeling. In *In Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI-2019), pp. 4369-4375, Macao, China.*, 2019.

[20] Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

[21] Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE signal processing magazine*, 34(4):43–59, 2017.

[22] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[24] Timothy J Boerner, Stephen Deems, Thomas R Furlani, Shelley L Knuth, and John Towns. Access: Advancing innovation: Nsf's advanced cyberinfrastructure coordination ecosystem: Services & support. In *Practice and experience in advanced research computing 2023: Computing for the common good*, pages 173–176. 2023.