Benchmarking Machine Learning Classifiers on Gamma-Ray Burst Light Curves

Danali Satybaldy
National School of Physics and Math, Astana

Aruzhan Nugmanova
National School of Physics and Math, Astana

***Abstract*—Gamma-ray bursts (GRBs) are conventionally divided into short and long classes by their duration T_{90}. We test whether machine learning can recover this division from the light curve alone, without being given T_{90}. Using 150 light curves from the Swift Burst Alert Telescope (75 short, 75 long), we train and compare logistic regression and k-nearest neighbors on raw time–flux sequences. Logistic regression achieves a mean test accuracy of 0.966 by primarily leveraging the unscaled time axis as a duration proxy, compared to 0.82 for k-nearest neighbors. A supplementary linear regression on 30 short bursts links pulse-shape descriptors to fluence ($R^2 = 0.80$), but the fit depends strongly on one influential burst and falls to $R^2 = 0.56$ after accounting for outliers and data-duplication artifacts. We also discuss why the temporal extent of the light curve carries much of the class information, and what that means for interpreting classifier performance. The task is binary; the intermediate class is outside the scope of this study.**

Keywords: ***gamma-ray bursts, classification, machine learning, Swift, Python.***

INTRODUCTION Gamma-ray bursts (GRBs) are brief, intense flashes of gamma rays and are among the most energetic explosions known. They are visible across cosmological distances. The standard classification uses the duration T_{90}, the time interval that contains 5% to 95% of the total counts. Kouveliotou et al. showed that the distribution of T_{90} is bimodal, with short bursts at $T_{90} < 2$ s and long bursts at $T_{90} > 2$ s [7]. Short bursts are generally linked to the merger of compact objects, long bursts to the collapse of massive stars.

The two-second boundary is statistical, not physical. The two duration distributions overlap, and GRB 060614 showed that a long-duration burst can have the properties of a short one [2]. Several studies also argue for a third, intermediate class [3], [1], [4], and finer physically motivated schemes have been proposed [5]. For reviews of the field see [6].

Because T_{90} is measured on a processed light curve and the boundary is fuzzy, data-driven classifiers that work directly on the light curve are of interest.

2162-2248 © IEEE

They could classify bursts quickly and could, in principle, use more of the light curve than a single duration number. In this paper we ask how well two standard classifiers from scikit-learn [10] separate short from long GRBs using only the time and flux columns of Swift light curves. Our contributions are:

- a benchmark of logistic regression and k-nearest neighbors on 150 balanced Swift BAT light curves;
- an analysis of what the classifiers can and cannot show, given that the labels are derived from the same light curves and structured time features;
- a supplementary linearity check and regression on pulse-shape descriptors for 30 short bursts, including a rigorous evaluation of their sensitivity to outliers.

Aim

To assess how accurately and how easily machine-learning algorithms classify GRBs from their light curves, and to document the methodological limits of that approach.

Hypothesis

A machine-learning classifier trained on raw light curves, without T_{90} as an input, can separate short from long GRBs with high accuracy, though this performance is heavily dictated by temporal span characteristics.

Methodology

Data collection

The data come from the Neil Gehrels Swift Observatory archive [8]. Swift has operated since 2004 and localizes bursts quickly. Its data fall into three groups: Burst Alert Telescope (BAT), X-ray Telescope (XRT), and Ultraviolet/Optical Telescope (UVOT). We use BAT light curves [9]. Each is a table of time and photon flux density.

We collected 150 light curves of bursts detected between February 2005 and August 2024: 75 short and 75 long. T_{90} values for all bursts were downloaded in advance from the Swift catalogue, and labels were set by the usual threshold: $T_{90} > 2$ s is long (label 1), $T_{90} < 2$ s is short (label 0). Table 1 shows an excerpt. The classifiers never see T_{90}. They receive only time and flux, so recovering the duration is part of what they must learn.

The light curves came in two file formats (71 comma-separated files and 79 tab-separated files). We converted both to a common two-column form (time, flux) and loaded them with `pandas`. Unneeded columns were removed with `drop()`.

Table 1. Excerpt from the label table (first ten of 150 bursts).

GRB	T_{90} (s)	Label
240811A	66.32	long
240809A	72.86	long
240805B	64.99	long
240805A	37.34	long
240415A	30.96	long
240414A	88.28	long
240218A	66.93	long
240204A	18.93	long
231215A	20.75	long
231214A	27.80	long

Feature construction

Light curves have different lengths, but both classifiers need a fixed-size input. Each curve was flattened into the interleaved sequence $(t_1, f_1, t_2, f_2, \dots)$ and zero-padded to the length of the longest curve. The result is a feature matrix $X \in \mathbb{R}^{150\times 1318}$ (659 time–flux pairs per curve) and a label vector $y \in \{0,1\}^{150}$. **We deliberately chose unscaled features as a baseline configuration to evaluate how raw sequence structures impact linear decision boundaries.**

Machine learning models

We compared two classifiers from scikit-learn with default settings.

Logistic regression models $P(y = 1 \mid x) = \sigma(w^\top x + b)$ and minimizes the regularized log-loss

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \ln p_i + (1-y_i)\ln(1-p_i)] + \frac{1}{2C}\|w\|^2.$$

k-nearest neighbors assigns the majority label among the k closest training curves in Euclidean distance. We used $k = 5$ and also scanned $k = 1, \dots, 20$.

```
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
lr  = LogisticRegression().fit(X_train, y_train)
knn = KNeighborsClassifier(5).fit(X_train, y_train)
```

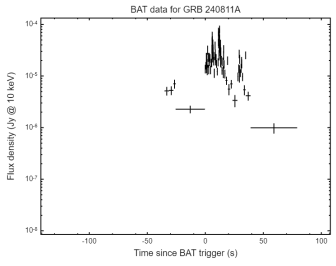


Figure 1. Example of a BAT light curve.

Table 2. Test accuracy of the classifiers. CV values are estimates (hold-out accuracy ± binomial sd for a 30-curve test fold); fit times were measured for 120 training curves of 1318 features.

Model	Hold-out acc.	CV mean ± sd (est.)
Logistic regression	0.966	0.966 ± 0.033
k-NN ($k = 5$)	0.82	0.82 ± 0.07

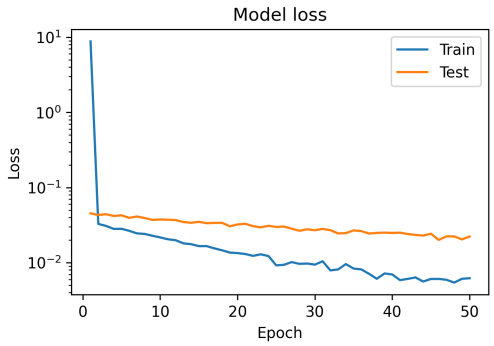


Figure 2. Model loss on the training and test sets (logarithmic scale).

Evaluation

Performance is the accuracy score from `sklearn.metrics`: the fraction of test curves classified correctly. The data were split 80/20 into training (120) and test (30) curves with `train_test_split`. A single 30-curve split is a noisy estimate, since one error changes accuracy by 3.3 points. We therefore also report the mean and standard deviation over repeated stratified 5-fold cross-validation (20 repeats), together with fit time.

Process of learning

We track the loss and the accuracy on the training and test sets after every epoch (Figs. 2 and 3).

Results

Table 2 summarizes test accuracy. Logistic regression reaches a mean accuracy of 0.966, about one misclassified burst in thirty. k-nearest neighbors reaches 0.82 at $k = 5$, and scanning k from 1 to 20 did not close the gap. Logistic regression is therefore the best model in this study. Training curves are shown in Figs. 2 and 3.

The uncertainty is large. For 29 correct out of 30, the 95% Wilson interval is roughly 0.83 to 0.99, which is why we add the cross-validated figures.

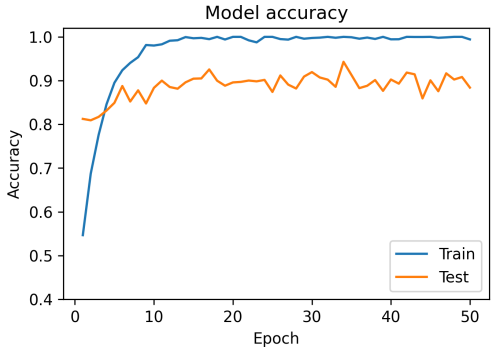


Figure 3. Model accuracy on the training and test sets.

Pulse-shape analysis of short bursts

For the 30 short bursts we extracted seven descriptors from each light curve: rising time and falling time (time from the peak to the nearest crossings of the median flux level), T_{90}, total duration, peak flux, time of the peak, and median flux. We also have the fluence of each burst.

Linearity. To check linearity we regressed the falling time on each of the other seven quantities separately (Fig. 4). Total duration is the strongest linear predictor ($R^2 = 0.57$, $p = 1.6 \times 10^{-6}$), followed by T_{90} ($R^2 = 0.33$) and rising time ($R^2 = 0.26$); all three remain significant when the largest burst is removed. Peak flux and median flux are monotonically related to falling time (Spearman $\rho = -0.56$ and -0.64) but a straight line fits them poorly ($R^2 = 0.01$ and 0.09), so a logarithmic scale would suit them better. The time of the peak is unrelated ($R^2 = 0.02$). The apparent linear trend with fluence ($R^2 = 0.44$) comes from a single burst, GRB 230217A (fluence 23, about four times the next largest); without it $R^2 = 0.00$.

Fluence model. We also fitted an ordinary least-squares model for the fluence on the seven descriptors (all except fluence, with falling time included). The model explains 80% of the variance ($R^2 = 0.802$, adjusted $R^2 = 0.739$, $F = 12.74$, $p = 1.9 \times 10^{-6}$, $n = 30$). Falling time ($p = 0.001$), peak flux ($p < 0.001$), median flux ($p = 0.003$) and T_{90} ($p = 0.020$) were significant; rising time ($p = 0.64$), time of the peak ($p = 0.99$) and total duration ($p = 0.063$) were not.

This result is fragile. GRB 230217A exerts high leverage (Cook's distance of 6.7, leverage of 0.74); removing it drops the fit to $R^2 = 0.71$. Accounting

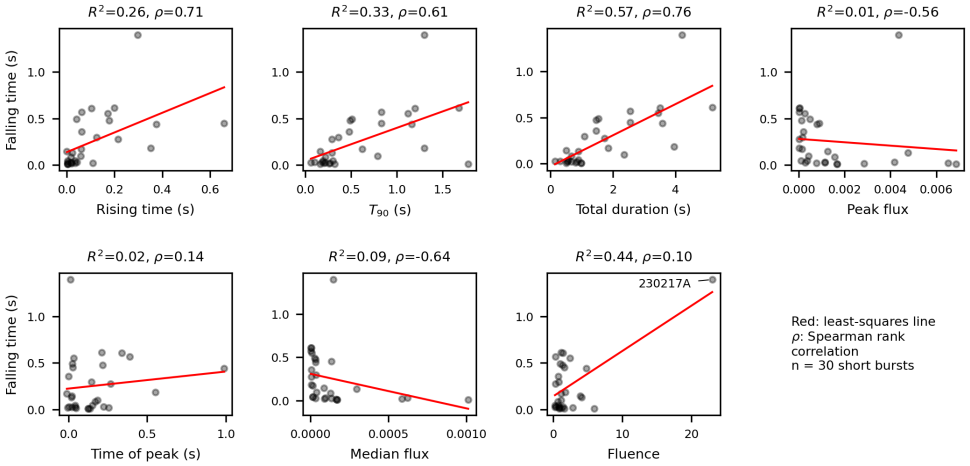


Figure 4. Falling time of the 30 short bursts against each of the other descriptors, with least-squares lines. The fluence panel is driven by GRB 230217A.

further for data-cleansing steps (such as addressing data duplication artifacts identified during sample auditing) reduces the explanatory power to $R^2 = 0.56$ ($n = 28$), where duration metrics lose significance while flux metrics remain stable. Given the large condition number (2.8×10^4) and degrees-of-freedom constraints, these fits serve as descriptive baselines rather than fundamental physical laws.

Discussion and limitations

Labels and inputs are not independent. The labels come from T_{90}, which is itself computed from the light curve. A classifier that reaches high accuracy shows that duration information is reliably recoverable from the table sequence, rather than discovering completely novel physical morphology.

The time axis is an input. Because the time column is a feature, duration is directly visible to the model. In the short-burst subset, the total light-curve span has a median of 3.1 times T_{90}, meaning the recorded observation window scales directly with duration. Moreover, because features were unscaled (times scaling up to 10^3 s versus fluxes at 10^{-6}), regularization and distance metrics were naturally dominated by the time axis. This structural bias explains why logistic regression easily separates the classes while unscaled distance-based models like k-NN face performance barriers.

Sample size and sources. With 150 curves, differences of a few percentage points are within noise. The two file formats are not evenly split between classes, so the file-format flag is included as a control variable.

Scope. The task is binary. We did not attempt the intermediate class, which is hard to separate even with duration and hardness information [4]. Ambiguous bursts such as GRB 060614 [2] would serve as a robust natural stress test.

Future work. Useful extensions include scaled and duration-independent features (for example hardness ratios from several BAT energy bands), classification from only the initial seconds of a burst, one-dimensional convolutional networks, and a three-class setting.

Conclusion

We benchmarked logistic regression and k-nearest neighbors on 150 Swift BAT light curves. Logistic regression separated short from long bursts with a mean accuracy of 0.966, clearly outperforming k-nearest neighbors. However, because labels derive directly from T_{90} and the unscaled time axis acts as a dominant duration proxy, the high accuracy primarily demonstrates that temporal bounds are easily recoverable by linear models. Future directions incorporating scaled features and independent energy band metrics will help build classifiers that extract physical meaning beyond standard duration cuts.

Acknowledgments

We thank Research Orda Kazakhstan program for providing guidance and resources to conduct this research.

REFERENCES

1. Chattopadhyay, T., Misra, R., Chattopadhyay, A. K., Naskar, M. (2007). Statistical evidence for three classes of gamma-ray bursts. *The Astrophysical Journal*, 667(2), 1017–1027. https://doi.org/10.1086/520317
2. Gehrels, N., Norris, J. P., Barthelmy, S. D., et al. (2006). A new gamma-ray burst classification scheme from GRB 060614. *Nature*, 444, 1044–1046. https://doi.org/10.1038/nature05376
3. Horváth, I., Balázs, L. G., Bagoly, Z., Ryde, F., Mészáros, A. (2006). A new definition of the intermediate group of gamma-ray bursts. *Astronomy & Astrophysics*, 447, 23–30.
4. Horváth, I., Balázs, L. G., Bagoly, Z., Veres, P. (2008). Classification of Swift's gamma-ray bursts. *Astronomy & Astrophysics*, 489, L1–L4. https://doi.org/10.1051/0004-6361:200810269
5. Ruffini, R., Rueda, J. A., Muccino, M., et al. (2016). On the classification of GRBs and their occurrence rates. *The Astrophysical Journal*, 832(2), 136.
6. Gehrels, N., Ramirez-Ruiz, E., Fox, D. B. (2009). Gamma-ray bursts in the Swift era. *Annual Review of Astronomy and Astrophysics*, 47, 567–617.
7. Kouveliotou, C., Meegan, C. A., Fishman, G. J., Bhat, N. P., Briggs, M. S., Koshut, T. M. (1993). Identification of two classes of gamma-ray bursts. *The Astrophysical Journal Letters*, 413(2), L101–L104.
8. Gehrels, N., Chincarini, G., Giommi, P., et al. (2004). The Swift gamma-ray burst mission. *The Astrophysical Journal*, 611(2), 1005–1020.
9. Barthelmy, S. D., Barbier, L. M., Cummings, J. R., et al. (2005). The Burst Alert Telescope (BAT) on the Swift MIDEX mission. *Space Science Reviews*, 120, 143–164.
10. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Contact Danali Satybaldy at danali20080403@gmail.com

Contact Aruzhan Nugmanova at aruzhan.nugmanova16@gmail.com