Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **504,** 3084–3091 (2021) doi:10.1093/mnras/stab1082
Advance Access publication 2021 April 19

A machine learning approach for GRB detection in *AstroSat* CZTI data

Sheelu Abraham,[1,2]★ Nikhil Mukund,[2,3]★ Ajay Vibhute,[2,4] Vidushi Sharma,[2] Shabnam Iyyani,[2] Dipankar Bhattacharya,[2] A. R. Rao,[5] Santosh Vadawale[6] and Varun Bhalerao[7]

[1]*Marthoma College, Chungathara, 679334 Nilambur, Kerala*
[2]*Inter-University Center for Astronomy and Astrophysics, Post Bag 4, Ganeshkhind, 411007 Pune, India*
[3]*Max-Planck-Institut für Gravitationsphysik (Albert-Einstein-Institut) and Institut für Gravitationsphysik, Leibniz Universität Hannover, Callinstraße 38, D-30167 Hannover, Germany*
[4]*Savitribhai Phule Pune University, 411007 Pune, Maharashtra, India*
[5]*Tata Institute of Fundamental Research, 400005 Mumbai, India*
[6]*Physical Research Laboratory, Ahmedabad, 380009 Gujarat, India*
[7]*Indian Institute of Technology, 400076 Bombay, India*

Accepted 2021 March 29. Received 2021 March 27; in original form 2019 June 30

ABSTRACT

We present a machine learning (ML) based method for automated detection of Gamma-Ray Burst (GRB) candidate events in the range 60–250 keV from the *AstroSat* Cadmium Zinc Telluride Imager data. We use density-based spatial clustering to detect excess power and carry out an unsupervised hierarchical clustering across all such events to identify the different light curves present in the data. This representation helps us to understand the instrument's sensitivity to the various GRB populations and identify the major non-astrophysical noise artefacts present in the data. We use Dynamic Time Warping (DTW) to carry out template matching, which ensures the morphological similarity of the detected events with known typical GRB light curves. DTW alleviates the need for a dense template repository often required in matched filtering like searches. The use of a similarity metric facilitates outlier detection suitable for capturing previously unmodelled events. We briefly discuss the characteristics of 35 long GRB candidates detected using the pipeline and show that with minor modifications such as adaptive binning, the method is also sensitive to short GRB events. Augmenting the existing data analysis pipeline with such ML capabilities alleviates the need for extensive manual inspection, enabling quicker response to alerts received from other observatories such as the gravitational-wave detectors.

Key words: methods: data analysis – methods: statistical – gamma rays: general – X-rays: bursts.

1 INTRODUCTION

GRBs, the most energetic explosions known to occur, typically release 10^{46}–10^{52} erg s^{-1} and last from a few milliseconds to a couple of minutes. Based on the duration over which 5 per cent to 95 per cent of the total burst fluence persists (T_{90}), these events are often classified as short when T_{90} is less than 2 s and long when it is otherwise (Kouveliotou et al. 1993; Gehrels & Mészáros 2012). The long GRB events are associated with the death of massive stars (Woosley 1993; Iwamoto et al. 1998; MacFadyen & Woosley 1999), and this correlation has been confirmed with the coincident detection of a supernova 1c with the long GRB030329A (Stanek et al. 2003). The short GRBs are supposed to have a different progenitor and are likely to be produced due to the merger of compact objects like binary neutron stars or a neutron star and a black hole (Eichler et al. 1989; Narayan, Paczynski & Piran 1992). The recent discovery of gravitational waves from the binary neutron star merger GW170817 by the advanced LIGO and advanced Virgo observatories (Abbott et al. 2017a) together with the detection of a short GRB (GRB170817A) by various gamma-ray instruments such as Fermi-GBM and Integral (Goldstein et al. 2017), have confirmed this proposed mechanism, thus marking the beginning of the multimessenger astronomy era.

A GRB event can be divided into two main epochs: a prompt emission phase and a subsequent afterglow phase. The former occurs in gamma rays immediately after the initial burst trigger, while the latter is observed in multiple wavelengths from gamma rays to radio extending over a period lasting from days to months. Timely identification of prompt emission is necessary to carry out follow-up observation in multiple wavelengths by ground and space-based telescopes. This step can lead to the detection of afterglows, which is crucial in determining the GRB's redshift and various other properties. Simultaneous operation of multiple detectors capable of GRB detection would lead to improved sky coverage and constrain the event's time of occurrence to a higher degree of precision. Observing short-duration GRBs, in conjunction with a GW trigger, helps in understanding the kilonovae mechanisms. Accurate time localization of these events can also constrain the differences in speed of light and gravity and thus scrutinize various theories of gravity (Abbott et al. 2017b).

★ E-mail: sheeluabraham@gmail.com (SA); nikhil.mukund@aei.mpg.de (NM)

The onboard alert systems of the Burst Alert Telescope (BAT; [50–150 keV]) on the Neil Gehrels Swift Observatory (Gehrels et al. 2004) and of the Gamma-Ray Burst Monitor (GBM; [8 keV–40 MeV]) on the *Fermi* satellite (Meegan et al. 2009), have led to an increased number of detections along with more afterglow observations. Cadmium Zinc Telluride Imager (CZTI) onboard *AstroSat*, is a wide field hard X-ray detector, and the increased transparency of collimators and surrounding supporting structures makes it sensitive to GRBs (Rao et al. 2016). In this paper, we overcome the absence of an onboard detector using an automated ML pipeline that enables low latency event detection in CZTI data.

The paper is organized as follows: in Section 2, the various pre-processing steps involved in generating light curves from CZTI data are presented. Section 3 briefly overviews the three machine learning algorithms used in this work and the proposed detection scheme. Section 4 talks about the results from the blind search, while conclusions and prospects are presented in Section 5.

2 *AstroSat* CZTI: DATA AND PRE-PROCESSING

AstroSat (Agrawal 2006; Singh et al. 2014) is India's first multiwavelength space observatory capable of making observations in X-ray and UV bands. It carries the following five science instruments for simultaneous observations of the source of interest: Ultra-Violet Imaging Telescope (UVIT; Tandon et al. 2017), Large Area X-ray Proportional Counters (LAXPC; Yadav et al. 2016), Soft X-ray Telescope (SXT; Singh et al. 2017), Cadmium Zinc Telluride Imager (CZTI; Rao et al. 2017), and Scanning Sky Monitor (SSM; Ramadevi et al. 2017). In particular, CZTI consists of an array of Cadmium Zinc Telluride (CZT) detectors, which are pixellated such that each pixel acts as an independent photon-counting detector. CZTI has a detector area of 976 cm^2 build using CZT modules and makes use of Coded Aperture Mask (CAM) for imaging (Bhalerao et al. 2017a). The total detection area is achieved by using 64 CZT modules of area 15.25 cm^2 each. These 64 modules are arranged in four identical and independent quadrants. The collimator walls separate these modules and collimators above each detector module, restrict the field of view to $4.6^\circ \times 4.6^\circ$ (full-width at half-maximum) at photon energies below 100 keV. As the penetrating power of X-ray photons increases strongly with energy, the collimator slats, and the coded aperture mask transmits a significant fraction of photons above 100 keV, and the instrument behaves like an all-sky open detector enabling the detection of GRBs from any direction. It also carries a Caesium Iodide (Tl) based scintillator detector operating as anticoincidence with the main CZT detector and is used as a veto detector. The coded aperture telescope is sensitive to hard X-ray polarization and was recently used to measure the polarized hard X-ray emission from Crab nebula (Vadawale et al. 2018).

CZTI is configurable in 16 different modes. The default mode of operation is the event mode, denoted as Mode M0 (Normal Mode). CZTI also records accumulated spectral and housekeeping information once every 100 s and stores the recorded information when it is changed to Secondary Spectral Mode (Mode SS). Whenever the spacecraft passes through the South Atlantic Anomaly (SAA), High Voltage (HV) in the CZTI and Veto detectors are switched off, and the detector is in Mode M9 (SAA mode), during which only the housekeeping information is recorded once every second. During the normal mode, whenever a photon hits a detector, CZT records the photon's arrival time, its position on the detector plane, and the corresponding energy. The time-tagged event list is stored in an event file. The events from four quadrants are stored as four different extensions of the event file. The recorded events also contain events generated due to the interaction of charged particles with the instrument or spacecraft body. The X-ray photons, consequently generated, can also deposit their energies in the CZTI detectors. Because of the pixellated nature of CZT, one charged particle can produce events in many pixels of CZT at the same time and are referred to as 'bunches'. During pre-processing, those bunches that do not belong to any astronomical source are mostly removed from the event file. Time intervals where data are not present due to SAA passage and data transmission errors are ignored, and a Good Time Interval (GTI) file is produced. The events belonging to GTIs are filtered and passed for further processing. During the data cleaning process, events from noisy or flickering pixels are also removed. The onboard calibration source, Am-241, emits X-rays of energy 60 keV and an alpha particle simultaneously. The alpha particle is absorbed in the CsI (TI) crystal, whereas the X-ray gets detected in the CZT pixel, and the alpha flag is set to 1. The events having the alpha flag equal to 1 are thereby removed from the event list. The cleaned event list so obtained is used as the input to the GRB detection algorithm.

One event file from CZTI usually consists of multi-orbit data, which may span 6000–30 000 s. We have divided the data into small chunks of 500 s each to check for any trigger present. This step, however, limits the ability to identify a trigger that occurs between two such chunks of data. Each event file consists of data from all four quadrants and the four veto channels, and only those events with energy higher than 60 keV are considered for further analysis. We conduct the search in the count space, and the conversion from counts to flux depends on the effective area as a function of direction and energy. This conversion varies quite strongly for CZTI (Bhalerao et al. 2017a) and requires prior knowledge of the transient location. For bright transients, there has been limited success in localizing a burst (Bhalerao et al. 2017b), and in such cases, one can attempt a joint solution for the source direction and spectrum. Consequently, it is possible that the event clustering is not related to the intrinsic characteristics of the GRB but only to the location in the relative detector coordinates. We also perform pre-processing to clean the time series before feeding it to the analysis pipeline. It is never perfect, and traces of SAA could still be visible in certain data segments. However, the pipeline is configured to take care of such noise sources, and in most cases, vetoes them successfully.

3 CANDIDATE EVENT SELECTION

This section describes the framework (see Fig. 1) used for GRB candidate event detection and its resourcefulness in bringing down the time needed to issue Gamma-ray Coordinates Network (GCN) alerts to the broader astronomy community. We also discuss the different machine learning algorithms deployed to detect real GRBs candidates from false triggers that arise from artefacts such as instrumental noise, cosmic rays, or even random fluctuations.

3.1 Template bank generation

We start with creating a template bank for long GRB light curves using 87 known GRBs.[1] The key idea is to minimize the number of templates while still achieving maximal coverage of the light curves' morphologies. These templates are obtained from already identified CZTI data events using the GCN trigger information published by the currently operating space observatories. We carry out one second binning for each event and use the interval correlation optimized shifting

[1] list of GRBs available as supplementary material

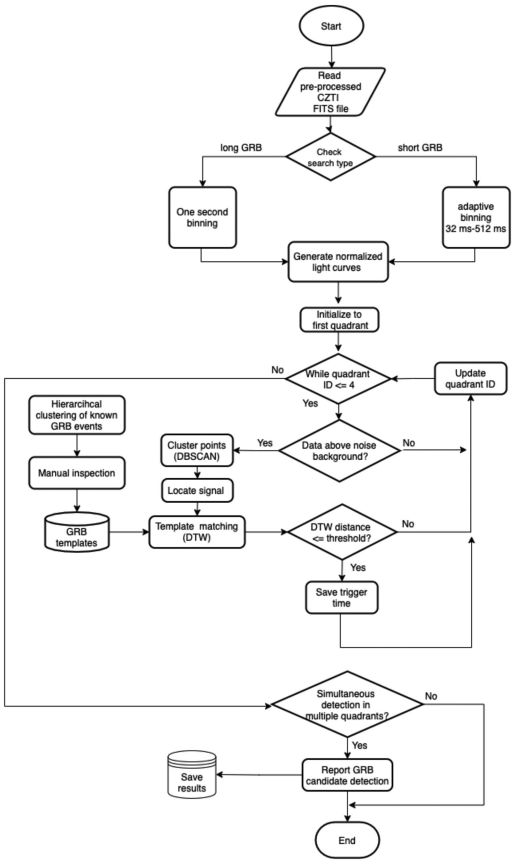

Figure 1. Schematic depicting the steps involved in *AstroSat* GRB candidate event detection pipeline.

(icoshift) technique (Savorani, Tomasi & Engelsen 2010; Tomasi, Savorani & Engelsen 2011) to correct for any delays among the light curves observed within the four quadrants. We then normalize every light curve individually by rescaling them to be between 0 and 1. This process also brings down the mean DC background close to zero for all GRB-like events. These curves are further stacked together to get a temporal sequence with a higher signal-to-noise ratio (SNR) for each of the 87 known GRB events. We then carry out a hierarchical clustering using Dynamic Time Warping (DTW; see Section 3.2) and identify the significant morphologies present within the data (see Fig. 2). The mean profile within each such cluster is then used to generate the GRB template bank. We use a bottom-up agglomerative clustering where the objects start as individual clusters, which are then hierarchically combined to form a dendrogram. The technique allows the user to choose any valid distance metric to compare the similarity between the objects. By maximizing the sum of similarities among the adjacent clusters, we can achieve optimal leaf ordering within the dendrogram (Bar-Joseph, Gifford & Jaakkola 2001). This ordering allows observing the progressively changing morphology within the given data samples. Hierarchical clustering has previously successfully identified the dominant groups among the short-duration transients, such as those observed in gravitational-wave observatories (Mukund et al. 2017). We construct a template bank consisting of 52 GRB light-curve templates based on the hierarchical clustering analysis results. We carefully choose these templates to guarantee adequate representation of all the probable morphologies of GRB events.

3.2 Detection scheme

We perform one second binning for the 500s data chunks independently for each of the four quadrants. We identify data significantly above the background noise by setting a threshold level three times the median absolute deviation above the median noise level. We then perform clustering using the DBSCAN (Ester et al. 1996) algorithm on these samples and identify the significant temporal sequences that are later used in template matching analysis. DBSCAN, which stands for Density-based spatial clustering of applications with noise, groups together data sets with similar features and identifies outliers automatically. As compared to K-Means like clustering algorithms (Hartigan & Wong 1979), there is no need to specify the number

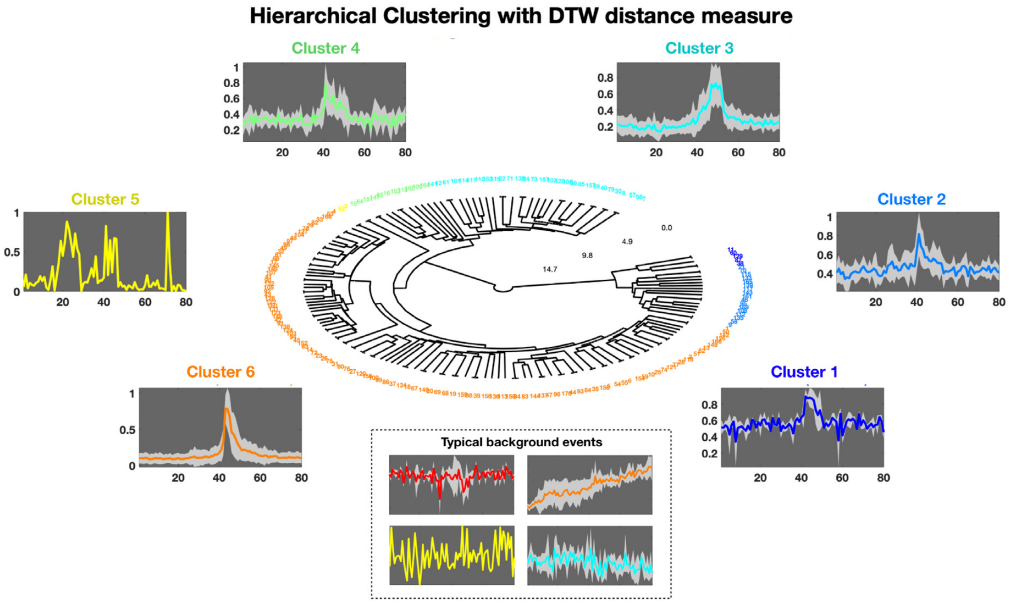

Figure 2. Hierarchical clustering of known GRB light curves using DTW as the distance measure. Mean curve identified in each cluster is further utilized in the search for new events via DTW based template matching. Events within the box depict the typical background events seen in the CZTI data.

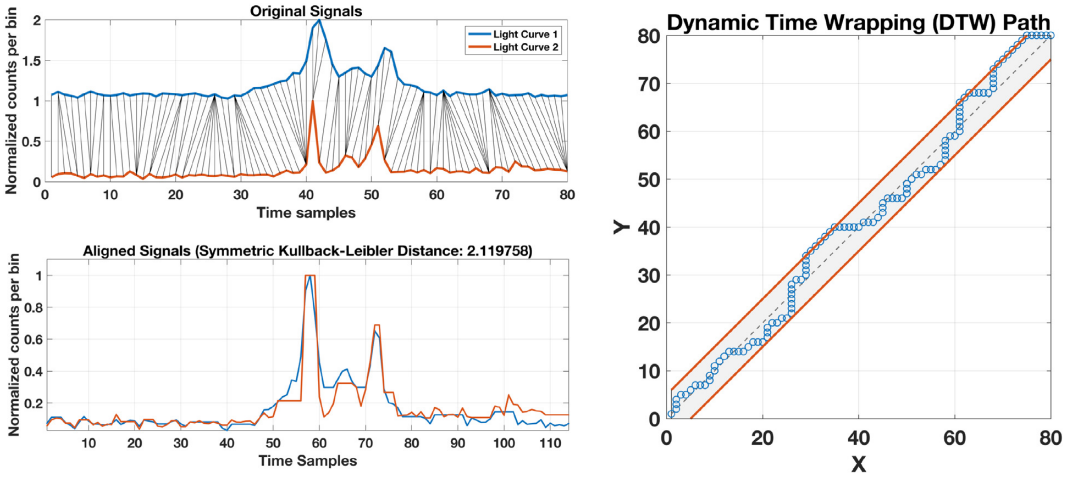

Figure 3. Distance between two GRB light curves estimated using DTW with symmetric Kullback–Leibler metric and an adjustment window size of five samples. Each curve is individually normalized based on its peak height.

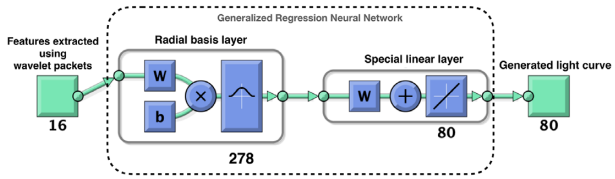

Figure 4. Shallow two-layer generalized regression neural network with Gaussian kernels used to synthesize simulated light curves from a limited amount of training data. **W** represents the Weight, and **b** is the bias used in the respective neural network hidden layer. The number of radial basis layers (278) equals the number of training data sets. We use 80 per cent of light curves from 87 verified GRB events and use the data from all four quadrants.

of clusters present in the data. The only required parameters for this algorithm are the minimum number of points in each cluster (**minPts**) and the maximum separation between samples (**eps**) for them to part of the same cluster. For this search looking for long GRB candidate events, we set to a value of five to both **minPts** and **eps**. We have used python implementation of DBSCAN from scikit-learn (Pedregosa et al. 2011).

Once the clusters are identified, the corresponding temporal sequences are normalized and checked for similarity to known astrophysical signals using the above-mentioned DTW technique. It is a general method developed for time-series alignment for speech and handwriting recognition. It can be applied to detect similar temporal sequences that are relatively stretched or squeezed with the template (Sakoe & Chiba 1978). The method eliminates the need for feature extraction and can be easily extended to carry out the similarity search using a template bank of known sequences. DTW finds the optimal alignment between time-series data, which allows a non-linear mapping of one signal to another, minimizing the distance between the two. To overcome the quadratic time and space complexity associated with the original DTW algorithm, we use FastDTW (Salvador & Chan 2007) implementation, an approximation to DTW whose complexity is linear, thus speeding up the computation time. Pursuing alternative methods like cross-correlation or matched filtering in this scenario would require a dense template bank, making them computationally challenging for rapid detection. DTW has previously been demonstrated to be helpful in the similarity study between light curves from both GRBs and their X-ray flares (Zhang, Zhang & Castro-Tirado 2016).

Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two vectors of lengths M and N, respectively. To create a mapping between the two vectors, we need to define a path. The aim is to find the path of minimum distance. The optimal path starts from (0,0), ends at (M, N), and in between maps the vectors on to a common set of indices i_x & i_y such that the total sum of distances, d

$$d = \sum_{\substack{m \in i_x \\ n \in i_y}} d_{m,n}(X, Y) \tag{1}$$

is minimized where the distance $d_{m,n}$ is expressed in terms of symmetric Kullback–Leibler (KL) metric (Kullback & Leibler 1951),

$$d_{m,n}(X, Y) = (x_m - y_n)(\log x_m - \log y_n)\ . \tag{2}$$

The KL divergence is widely used in Bayesian inference and provides information about how well an approximate probability distribution represents the real underlying model.

The DTW path is constrained to move close to the diagonal by specifying a window around the main diagonal to minimize the effect of outliers. Additionally to ensure alignment of the complete signal and not just segments as well to prevent sample skipping, only the following transitions are permitted while the path proceeds from (0,0) to (M, N),

$$(m, n) \rightarrow (m + 1, n)$$
$$(m, n) \rightarrow (m, n + 1)$$
$$(m, n) \rightarrow (m + 1, n + 1)\ .$$

Fig. 3 shows one such instance of DTW-based alignment of two GRB light curves. We claim a detection if a trigger matches any of the template GRB models within a specified DTW distance and is coincidentally present in at least three quadrants channels.

3.3 Performance evaluation

In general, we can assess the performance of the detection scheme described above from its receiver operator characteristic (ROC) curve, which compares the rate of detected actual events to the false triggers at varying detection threshold levels. Based on the available comparatively small data set, we carry out non-parametric modelling of both the GRB-like events and the expected noise sources and construct the ROC curves mentioned above. Combining

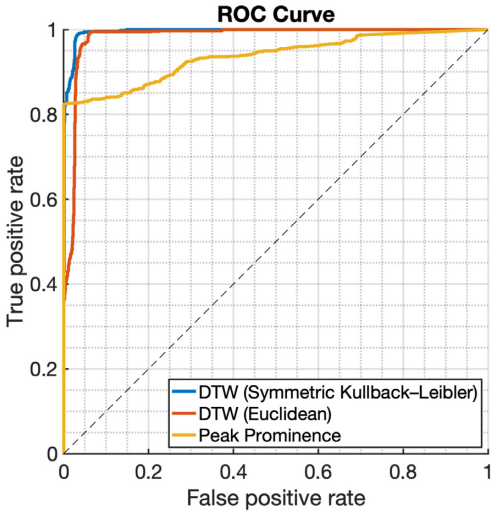

Figure 5. Receiver operator characteristic (ROC) curve depicting the true positive rate versus the false positive rate for both the DTW template matching scheme and a traditional peak detection algorithm. Curves are generated by varying the respective threshold parameter, DTW distance, and the peak prominence.

GRNN with discrete wavelet transform has been previously shown to model temporal sequences effectively (Kişi 2011). The usual discrete wavelet decomposition (DWT) splits the signal into approximation and detail coefficients where the detail coefficients record the information lost between successive lower frequency approximations. However, wavelet packets decompose the details coefficients further into approximation and detail coefficients, thus improving the coarse resolution DWT and not being as computationally expensive as the continuous wavelet transform (CWT). We carry out such wavelet packet decomposition (Laine & Fan 1993; Ta 1994; Walczak, Van Den Bogaert & Massart 1996) of the light curves using Daubechies wavelets and train a generalized regression neural network (GRNN) with these extracted features to generate the synthetic light curves.

With their feed-forward shallow network architecture (see Fig. 4), GRNNs avoid back-propagation and carry out a single pass learning with as many neurons as the number of data sets (Specht 1991). These networks use normalized radial basis function in their hidden layer, memorize all input–output sequences, and generalize them for newer inputs. These characteristics considerably decrease the overall training time and make them well suited for problems where training data availability is limited (Sarshar, Kabiri & Barkeshli 2001). We use the manually verified 87 GRB events and 36 non-astrophysical artefacts that include instances of SAA and cosmic rays to create a training data set where 20 per cent is kept aside for validation. GRNN learning is, in general, sensitive to the variance of the involved radial basis function. We compare the network predictions against the validation data and optimize this parameter by minimizing the normalized mean-squared error between the actual and predicted light curves. To generate synthetic events (1000 samples each for source and background events), we introduce random jitter at a few percent levels in the extracted wavelet parameters, draw samples from their distribution, and feed them to the respective trained GRNNs.

To access the relative improvement in performance, we compare the DTW classifier with a traditional peak finding algorithm on the synthetic data set and depict the obtained ROC curves in Fig. 5. These curves are constructed by varying the respective threshold parameter, DTW distance, and the peak prominence and calculating the true positive rate (TPR) and the false positive rate (FPR) at each of these points. As compared to the Euclidean metric, the DTW distance calculated using the symmetric KL metric provides better performance. We set the permissible FPR to 1 per cent (TPR = 98 per cent) and accordingly get an upper limit on the DTW

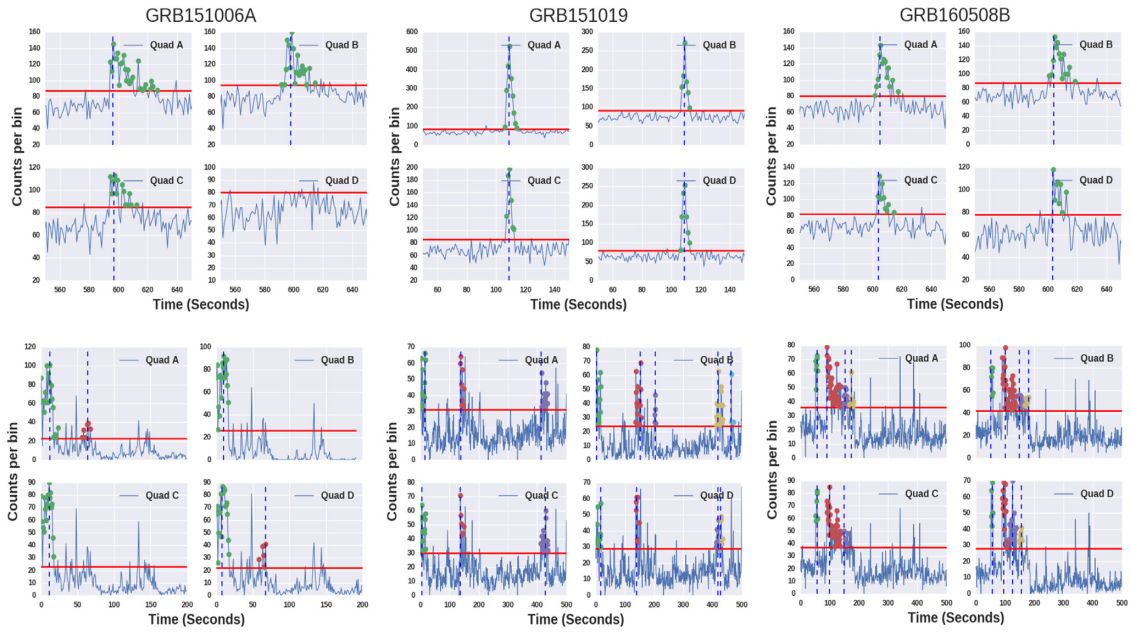

Figure 6. Few detection scenarios encountered while searching for long GRBs with one-second binning. The upper panel shows the correctly identified events, while the lower panel highlights instances of contamination from non-GRB transients. The horizontal red line determines the threshold caused by the noise background level, and significant points above this are clustered using the DBSCAN algorithm. Each cluster is uniquely coloured, with the vertical line providing the cluster centre.

Table 1. GRBs candidate events detected with the machine learning algorithm described in this paper. GCN circulars have been issued for the highlighted events.

GRB ID	Trigger time (UTC)	T_{90} (s)	Peak count rate (s^{-1})	Total counts	Mean background count rate (s^{-1})	Detection significance
GRB151019A	Oct 19 2015 8:05:25	6.1 ± 0.16	1370 ± 79.7	4877 ± 112.4	453 ± 25.2	64.4 ± 4.2
GRB151217A	Dec 17 2015 4:58:21	29.6 ± 0.07	108 ± 30.0	2030 ± 137.1	503 ± 22.6	4.8 ± 1.3
GRB151219B	Dec 19 2015 9:11:18	16.5 ± 0.15	300 ± 37.8	1952 ± 92.8	473 ± 22.9	13.8 ± 1.8
GRB151224A	Dec 24 2015 2:26:35	7.1 ± 0.19	117 ± 30.3	440 ± 41.2	477 ± 23.5	5.4 ± 1.4
GRB160119C	Jan 19 2016 3:08:26	31.2 ± 0.12	233 ± 31.8	4819 ± 147.2	423 ± 21.5	11.3 ± 1.6
GRB160128A	Jan 28 2016 12:59:42	33.4 ± 017	712 ± 61.0	9614 ± 189.1	469 ± 23.7	32.9 ± 2.9
GRB160214A	Feb 14 2016 9:17:26	26.6 ± 0.1	526 ± 55.7	3840 ± 139.1	463 ± 30.8	24.4 ± 2.7
GRB160221B	Feb 21 2016 18:56:43	7.8 ± 0.05	378 ± 49.4	1319 ± 58.1	491 ± 24.0	17.1 ± 2.3
GRB160223B	Feb 23 2016 9:59:03	13.5 ± 0.15	239 ± 38.5	1985 ± 85.3	480 ± 28.8	10.9 ± 1.8
GRB160310C	Mar 10 2016 16:27:13	16.7 ± 0.01	1481 ± 82.9	13327 ± 154.3	503 ± 23.4	66.0 ± 4.0
GRB160325B	Mar 25 2016 6:59:23	43.6 ± 0.02	1467 ± 81.9	22263 ± 279.8	474 ± 22.8	67.4 ± 4.1
GRB160418A	Apr 18 2016 18:08:44	31.0 ± 0.37	532 ± 48.0	7247 ± 175.9	425 ± 22.0	25.8 ± 2.4
GRB160720A	Jul 20 2016 18:25:23	14.5 ± 0.06	430 ± 41.8	4244 ± 99.9	459 ± 23.7	20.1 ± 2.0
GRB160805A	Aug 05 2016 22:26:18	20.1 ± 0.02	132 ± 30.9	218 ± 60.3	428 ± 22.3	6.4 ± 1.5
GRB160824B	Aug 24 2016 13:51:28	17.8 ± 0.06	172 ± 35.5	2556 ± 98.9	590 ± 29.5	7.1 ± 1.5
GRB160829B	Aug 29 2016 14:18:47	18.0 ± 0.05	1652 ± 85.5	5438 ± 131.5	504 ± 25.0	73.6 ± 4.2
GRB170210B	Feb 10 2017 2:48:13	34.3 ± 0.16	985 ± 70.4	13444 ± 249.0	551 ± 26.3	42.0 ± 3.2
GRB170216A	Feb 16 2017 16:39:33	14.8 ± 0.19	301 ± 41.0	2634 ± 91.0	508 ± 24.0	13.4 ± 1.8
GRB170228A	Feb 28 2017 19:03:01	13.4 ± 0.02	728 ± 62.1	4356 ± 103.5	485 ± 23.8	33.1 ± 2.9
GRB170311C	Mar 11 2017 13:45:10	7.4 ± 0.05	486 ± 53.9	2152 ± 59.1	517 ± 23.8	21.4 ± 2.4
GRB170316A	Mar 16 2017 17:02:22	14.2 ± 1.22	261 ± 41.9	1861 ± 106.7	486 ± 25.2	11.8 ± 1.9
GRB170423B	Apr 23 2017 20:55:22	12.9 ± 0.05	441 ± 46.6	2413 ± 77.1	474 ± 23.6	20.3 ± 2.2
GRB170614A	Jun 14 2017 11:40:01	15.6 ± 0.27	666 ± 54.6	6181 ± 122.1	485 ± 23.4	30.2 ± 2.6
GRB170808B	Aug 08 2017 22:27:47	12.1 ± 0.05	1292 ± 77.6	3679 ± 79.3	460 ± 22.4	60.2 ± 3.9
GRB170825B	Aug 25 2017 12:00:06	5.9 ± 0.02	449 ± 52.4	1543 ± 54.2	501 ± 24.0	20.1 ± 2.4
GRB170901B	Sep 01 2017 11:59:57	11.4 ± 0.04	310 ± 43.3	2138 ± 66.7	495 ± 23.3	13.9 ± 2.0
GRB170915A	Sep 15 2017 3:51:28	10.2 ± 0.07	246 ± 38.7	1708 ± 72.9	543 ± 25.4	10.6 ± 1.7
GRB180401A	Apr 01 2018 20:17:35	19.5 ± 0.03	806 ± 65.3	6177 ± 110.5	531 ± 24.2	35.0 ± 2.9
GRB180403A	Apr 03 2018 13:32:52	7.0 ± 0.06	186 ± 38.0	881 ± 39.4	497 ± 23.3	8.3 ± 1.7
GRB180411C	Apr 11 2018 12:28:32	75.0 ± 0.02	360 ± 47.9	4603 ± 222.6	487 ± 24.1	16.3 ± 2.2
GRB180416C	Apr 16 2018 8:10:52	9.2 ± 0.05	429 ± 46.3	2595 ± 67.0	508 ± 23.6	19.0 ± 2.1
GRB180426A	Apr 26 2018 13:10:59	12.4 ± 0.03	434 ± 49.7	1944 ± 61.6	498 ± 23.3	19.4 ± 2.3
GRB180504B	May 04 2018 3:15:57	13.3 ± 0.03	251 ± 51.0	1870 ± 89.0	501 ± 47.9	11.2 ± 2.3
GRB180526A	May 26 2018 11:04:18	57.8 ± 0.15	541 ± 55.6	6504 ± 213.9	497 ± 24.6	24.3 ± 2.6
GRB180603A	Jun 03 2018 16:22:57	31.1 ± 0.07	372 ± 42.8	6124 ± 139.7	486 ± 23.4	16.9 ± 2.0

threshold value to be 12 for the KL metric-based distance estimation. We carried out initial prototyping of the algorithms in MATLAB. The final detection pipeline is written in python for better integration with the rest of the satellite data analysis tools. The pipeline is currently configured to alert the CZTI *AstroSat* support team about the most probable GRB candidates, who then makes the final decision on issuing GCN alerts.

4 RESULTS

Some of the detection scenarios involving true detection of long GRBs along with typical false-positive candidates are shown in Fig. 6. The upper panels show the correctly identified events, while the lower panels depicted certain instances when the events picked up were due to noise artefacts. After the testing and validation steps mentioned in the previous section, we carried out a blind search targeting long GRB events on the CZTI data collected from 2015 October 8 to 2018 November 7. The pipeline detected 223 probable candidates, out of which 170 were already known to be GRBs. Detailed analysis of the rest led to the discovery of 35 long GRBs candidate events along with 18 false positives. Table 1 lists these newly discovered events along with their trigger time in UTC, T_{90},[2] peak count rate, total count rate, mean background counts, and detection significance with their uncertainties. Of these, GCN alerts were issued for two such events, namely GRB180526A (Sharma, Vibhute & Bhattacharya 2018a) and GRB180603A (Sharma, Vibhute & Bhattacharya 2018b). The observed number of false triggers are higher than the expected as shown in Fig. 5. While the origin of many of these is not well understood, some of them seem to be related to the scenario shown in Fig. 6 (lower right), where we see multiple short-duration transients all occurring within tens of seconds.

Certain minor modifications were necessary to make the pipeline sensitive to GRBs occurring at shorter time-scales. As these events showed a higher sensitivity to the time bins' size, it was necessary to vary the bin size across time-scales ranging from 32 to 512 ms for the entire data chunk and select the value that maximizes the peak count in at least three of the four quadrants. The time bin's optimal value was determined using the differential evolution (Storn 1996) global

[2]The calculation of T_{90} and the other column values are described in Appendix A

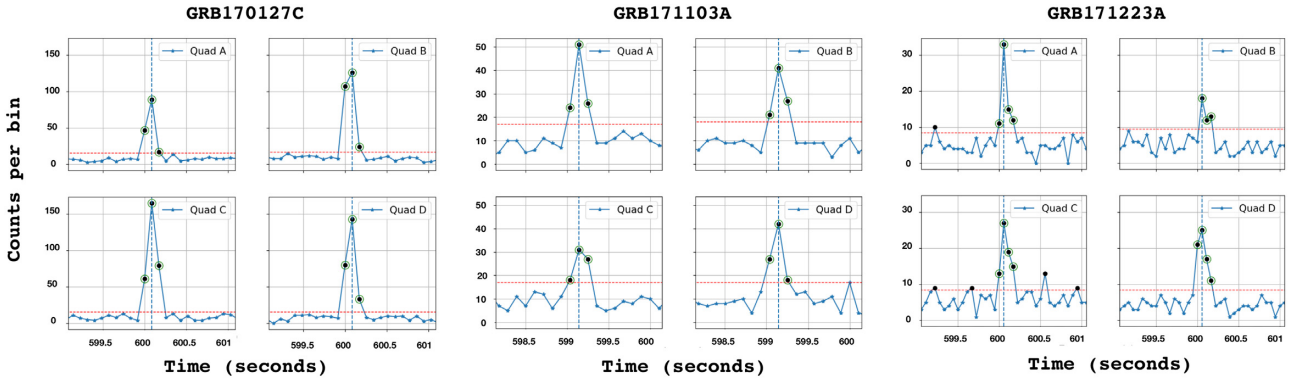

Figure 7. Instances of short GRBs detected in CZTI data. In each case, we optimize the time binning to maximize the peak counts in multiple quadrants. Time bins used for GRB170127C, GRB171103A, and GRB171223A are respectively 85, 106, and 55 ms. The DBSCAN algorithm identifies significant clusters denoted by the green circles, followed by the DTW technique, which cross-matches them with the known GRB templates.

optimizer, which performed a gradient-free direct search across this continuous parameter space. Besides, the 3σ detection threshold used for longer GRBs had to be increased to 4.5σ to minimize contamination from the background noise. In Fig. 7 we report the triggers seen for GRB170127C (Ajello et al. 2019), GRB171103A, and GRB171223A, which were respectively constructed using a bin size of 85, 106, and 55 ms.

5 CONCLUSIONS AND OUTLOOK

We demonstrated various machine learning algorithms' resourcefulness for robust GRB detection using *AstroSat* CZTI data. Automating such tasks can bring down the response time leading to efficient follow-up studies related to multimessenger astronomy. Compared to conventional peak detection algorithms, incorporating morphology decreases the false detection rate from instrumental artefacts and non-GRB phenomena. The newly developed scheme has been tested on both short and long-duration GRB events and is now part of the *AstroSat* CZTI data analysis pipeline. In the future, we would like to focus more on improving the detection efficiency for short GRBs through better time localization and a reduction in the number of false positives. One natural way to achieve this would be to extend the template bank to include more models for real and spurious events. The techniques presented in this work are very well applicable to astronomical data sets such as stellar spectra or temporal sequences that are transient hence like the gravitational-wave transient signals. The feasibility of embedding such ML algorithms in FPGA-based hardware for low latency onboard trigger detection is also worth exploring in the context of next-generation detectors and would be part of future studies.

ACKNOWLEDGEMENTS

The authors would like to thank the referees for their valuable comments, which helped improve the manuscript. This publication uses the data from the *AstroSat* mission of the Indian Space Research Organisation (ISRO), archived at the Indian Space Science Data Centre (ISSDC). CZT-Imager is built by a consortium of institutes across India, including Tata Institute of Fundamental Research, Mumbai, Vikram Sarabhai Space Centre, Thiruvananthapuram, ISRO Satellite Centre, Bengaluru, Inter-University Centre for Astronomy and Astrophysics, Pune, Physical Research Laboratory, Ahmedabad, Space Application Centre, Ahmedabad: contributions from the vast technical team from all these institutes are gratefully acknowledged. The authors express thanks to the CZTI *AstroSat* support cell at IUCAA for their help in data curation and pre-processing. NM acknowledges Council for Scientific and Industrial Research (CSIR), India, for providing financial support as Senior Research Fellow. The authors also express thanks to Ninan Sajeeth Philip for his valuable comments and suggestions.

DATA AVAILABILITY

This publication uses the data from the *AstroSat* mission of the Indian Space Research Organisation (ISRO) and the data are available in Indian Space Science Data Center (ISSDC) at https://astrobrowse.issdc.gov.in/astro_archive/archive/Home.jsp.

REFERENCES

Abbott B. P. et al., 2017a, Phys. Rev. Lett., 119, 161101
Abbott B. P. et al., 2017b, ApJ, 848, L13
Agrawal P., 2006, Adv. Space Res., 38, 2989
Ajello M. et al., 2019, ApJ, 878, 52
Bar-Joseph Z., Gifford D. K., Jaakkola T. S., 2001, Bioinformatics, 17, S22
Bhalerao V. et al., 2017a, J. Astrophys. Astron., 38, 31
Bhalerao V. et al., 2017b, ApJ, 845, 152
Eichler D., Livio M., Piran T., Schramm D. N., 1989, Nature, 340, 126
Ester M., Kriegel H.-P., Sander J., Xu X., 1996, in Evangelos S., Jiawei H., Usama F., eds, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. KDD'96. AAAI Press, Portland, Oregon, USA, p. 226
Gehrels N., Mészáros P., 2012, Science, 337, 932
Gehrels N. et al., 2004, ApJ, 611, 1005
Goldstein A. et al., 2017, ApJ, 848, L14
Hartigan J. A., Wong M. A., 1979, JSTOR: Appl. Stat., 28, 100
Iwamoto K. et al., 1998, Nature, 395, 672
Kişi Ö., 2011, KSCE J. Civil Eng., 15, 1469
Kouveliotou C., Meegan C. A., Fishman G. J., Bhat N. P., Briggs M. S., Koshut T. M., Paciesas W. S., Pendleton G. N., 1993, ApJ, 413, L101
Kullback S., Leibler R. A., 1951, Ann. Math. Stat., 22, 79
Laine A., Fan J., 1993, IEEE Trans. Pattern Anal. Mach. Intell., 15, 1186
MacFadyen A., Woosley S., 1999, ApJ, 524, 262
Meegan C. et al., 2009, ApJ, 702, 791
Mukund N., Abraham S., Kandhasamy S., Mitra S., Philip N. S., 2017, Phys. Rev. D, 95, 104059
Narayan R., Paczynski B., Piran T., 1992, ApJ, 395, L83
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825

Ramadevi M. C. et al., 2017, Exp. Astron., 44, 11
Rao A. R. et al., 2016, ApJ, 833, 86
Rao A. R., Bhattacharya D., Bhalerao V. B., Vadawale S. V., Sreekumar S., 2017, Current Science, 113, 595
Sakoe H., Chiba S., 1978, IEEE Trans. Acoust. Speech Signal Process., 26, 43
Salvador S., Chan P., 2007, Intell. Data Anal., 11, 561
Sarshar N., Kabiri A., Barkeshli K., 2001, IEEE Antennas and Propagation Society International Symposium. 2001 Digest. Held in conjunction with: USNC/URSI National Radio Science Meeting (Cat. No.01CH37229, vol. 2). IEEE, Boston, MA, USA, p. 690
Savorani F., Tomasi G., Engelsen S. B., 2010, J. Magn. Reson., 202, 190
Sharma V., Vibhute A., Bhattacharya D., 2018a, GRB 180526A: AstroSat CZTI detection Available at: https://gcn.gsfc.nasa.gov/other/180526A.gcn3
Sharma V., Vibhute A., Bhattacharya D., 2018b, GRB 180603A: AstroSat CZTI detection. Available at: https://gcn.gsfc.nasa.gov/other/180603A.gcn3
Singh K. P. et al., 2014, in Takahashi T., ed., Space Telescopes and Instrumentation 2014: Ultraviolet to Gamma Ray. SPIE - International Society for Optics and Photonics Montreal, Quebec, Canada, p. 91441S
Singh K. P. et al., 2017, J. Astrophys. Astron., 38, 29
Specht D. F., 1991, IEEE Trans. Neural Netw., 2, 568
Stanek K. Z. et al., 2003, ApJ, 591, L17
Storn R., 1996, in Smith M. H., Lee M. A., Keller J., Yen J., eds, Proceedings of North American Fuzzy Information Processing. IEEE, Berkeley, California, USA, p. 5
Tandon S. N. et al., 2017, AJ, 154, 128
Ta N. P., 1994, Proceedings of IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis. IEEE, Philadelphia, PA, USA, p. 508
Tomasi G., Savorani F., Engelsen S. B., 2011, J. Chromatogr. A., 1218, 7832
Vadawale S. V. et al., 2018, Nat. Astron., 2, 50
Walczak B., Van Den Bogaert B., Massart D. L., 1996, Anal. Chem., 68, 1742
Woosley S. E., 1993, ApJ, 405, 273
Yadav J. S. et al., 2016, in den Herder J.-W. A., Takahashi T., Bautz M., eds, Proc. SPIE 9905, Space Telescopes and Instrumentation 2016: Ultraviolet to Gamma Ray. IEEE, Edinburgh, UK, p. 99051D
Zhang B.-B., Zhang B., Castro-Tirado A. J., 2016, ApJ, 820, L32

SUPPORTING INFORMATION

Supplementary data are available at *MNRAS* online.

Template_Gen_Evnt.pdf

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

APPENDIX: LIGHT CURVE PARAMETERS

The *AstroSat* CZTI method of estimating the T_{90} is based on the accumulation of counts. A similar process is used in BeppoSAX, HETE-2, CGRO/BATSE, and INTEGRAL observatories. At first, the light curve is generated with counts per time bin. We coarsely choose and store the information of the GRB that includes both pre- and post-background regions from the light curve. The count estimated in each time bin in the light curve has a Poisson error associated with it. We simulate 50 000 such light curves by randomly drawing each bin count from the corresponding Poissonian distribution.

For each simulated light curve, the parameters: T_{90}, peak count rate (PCR; which is the maximum count rate observed in the light curve of the GRB), accumulated total counts, and mean background rate are calculated as follows:

(i) The background is modelled by fitting the selected pre- and post-GRB background regions by a polynomial. The local mean background rate (MBR) is obtained by averaging the count rates found in these regions, subsequently subtracted from the light curve. The peak count rate and the corresponding time found in the resultant light curve are noted.

(ii) Using the background-subtracted light curve, the cumulative counts per bin are plotted with time. The duration, T_{90} is calculated as $T_{90} = T_{95} - T_5$, where T_{95} and T_5 are the times when 95 per cent and 5 per cent of the total GRB event counts are obtained, respectively. The accumulated total counts in T_{90} interval is also calculated. We obtain the distribution for each parameter with the above steps, and its mean and standard deviation are used for reporting the parameter value and its uncertainty. In the case of count rate, we note that the standard deviation obtained from the distribution only reflects the variation arising from different simulations of the light curve, and therefore, the total error on the reported count rate value (N s^{-1}) is the sum of the standard deviation of the distribution and the Poisson error ($\sqrt{N}$).

The detection significance is calculated as PCR/$\sqrt{\text{MBR}}$ and the error on it is obtained by the standard method of error propagation using the uncertainities reported for PCR and MBR.

This paper has been typeset from a TEX/LATEX file prepared by the author.