Revista Mexicana de Astronomía y Astrofísica, 61, 25-34, (2025)

OPEN ACCESS https://doi.org/10.22201/ia.01851101p.2025.61.02.03

Deep-TAO: The Deep Learning Transient Astronomical Object data set for Astronomical Transient Event Classification

John F. Suárez-Pérez[1,2], Catalina Gómez[3], Mauricio Neira[4], Marcela Hernández Hoyos[4], Pablo Arbeláez[5] and Jaime E. Forero-Romero[2]

[1] *Tecnológico de Monterrey, Escuela de Ingeniería y Ciencias, Zapopan, México (jf.suarez@tec.mx).*
[2] *Departamento de Física, Universidad de los Andes, Bogotá, Colombia.*
[3] *Department of Computer Science, Johns Hopkins University, USA.*
[4] *Systems and Computing Engineering Department, Universidad de los Andes, Bogotá, Colombia.*
[5] *Center for Research and Formation in Artificial Intelligence, Universidad de los Andes, Bogotá, Colombia.*

Keywords: *catalogues, methods: data analysis, transients: general*

Abstract

We present the Deep-learning Transient Astronomical Object (Deep-TAO), a dataset of 1,249,079 annotated images from the Catalina Real-time Transient Survey, including 3,807 transient and 12,500 non-transient sequences. Deep-TAO has been curated to provide a clean, open-access, and user-friendly resource for benchmarking deep learning models. Deep-TAO covers transient classes such as blazars, active galactic nuclei, cataclysmic variables, supernovae, and events of an indeterminate nature. The dataset is publicly available in FITS format, with Python routines and Jupyter notebooks for easy data manipulation. Using Deep-TAO, a baseline Convolutional Neural Network outperformed traditional random forest classifiers trained on light curves, demonstrating its potential for advancing transient classification.

Resumen

Presentamos Deep-learning Transient Astronomical Object (Deep-TAO), un conjunto de 1,249,079 imágenes anotadas del Catalina Real-time Transient Survey, que incluyen 3,807 secuencias transientes y 12,500 no transientes. Deep-TAO ha sido diseñado como un recurso limpio, de acceso abierto y fácil de usar, ideal para evaluar y comparar modelos de aprendizaje profundo. Incluye eventos transientes como blazares, núcleos galácticos activos, variables cataclísmicas, supernovas y eventos de naturaleza indeterminada. El conjunto de datos está disponible públicamente en formato FITS, acompañado de rutinas en Python y cuadernos de Jupyter que facilitan su uso. Utilizando Deep-TAO, una red neuronal convolucional básica superó el desempeño de clasificadores tradicionales basados en bosques aleatorios entrenados con curvas de luz, demostrando su potencial para mejorar la clasificación de eventos transientes.

Corresponding author: John F. Suárez-Pérez *E-mail address:* jf.suarez@tec.mx
Received: January 28, 2025 **Accepted:** March 28, 2025

1. Introduction

A primary challenge in time-domain astronomy is the detection and classification of transient astronomical events. In recent years, methods for automating these processes have seen remarkable improvements in both complexity and computational efficiency, driven by the exponential growth of datasets requiring timely analysis (Kaiser, 2004; Law et al., 2009; Smartt, S. J. et al., 2015; Chambers et al., 2016; Martínez-Palomera et al., 2018; Bellm et al., 2019; Dyer et al., 2020; Nidever et al., 2021).

Machine learning (ML) (Wyrzykowski et al., 2014; D'Isanto et al., 2016; Gieseke et al., 2017; Neira et al., 2020; Sánchez-Sáez et al., 2021; Van Roestel et al., 2021) and deep learning (DL) approaches (Gieseke et al., 2017; Cabrera-Vives et al., 2017; Carrasco-Davis et al., 2019; Muthukrishna et al., 2019; Gómez et al., 2020; Sánchez-Sáez et al., 2021; Allam & McEwen, 2024; Van Roestel et al., 2021; Killestein et al., 2021) have demonstrated their capability to provide rapid and accurate solutions for transient classification tasks, offering significant advancements over traditional methods.

The further development and optimization of ML and DL algorithms critically depend on the availability of large-scale, high-quality, and representative datasets. These datasets can be constructed from real observational data (Neira et al., 2020), synthesized light curves (Carrasco-Davis et al., 2019), or image-based data derived from real (Scalzo et al., 2017) or simulated

observations (Carrasco-Davis et al., 2019). The diversity and realism of these datasets are essential for improving the generalizability and robustness of classification models in the context of astronomical transient phenomena.

The image-based datasets that can be used to test and train new DL applications usually present some limitations.

1) Restricted access. Some datasets are private, and only survey collaborators can access them. This limits the possibilities for a broader group of scientists to use the dataset to improve DL techniques.
2) Inconvenient access. Some surveys have set up public websites to access their data. However, sometimes the system is designed to retrieve information about individual objects (Drake et al., 2009; Scalzo et al., 2017; Nidever et al., 2021) and not large samples. This makes it inconvenient to compile the full dataset required for DL training.
3) Unrealistic images. Although other public, open access datasets exist, they are based on simulated images (Carrasco-Davis et al., 2019). This limits the realism required to optimally train DL architectures.
4) Incomplete labels. There are public, easy-to-gather, and realistic datasets that do not have labels on their data (Smartt, S. J. et al., 2015). These labels are required to train the supervised DL architectures.

To date, no dataset for DL transient classification has been made easily accessible to the public in the form of a fully labeled catalog based on observations.

The purpose of this study is to present a dataset to fill this gap. We denominate this dataset Deep-TAO, for Deep-Learning Transient Astronomical Objects. Deep-TAO was built using public data from the Catalina Real-Time Transient Survey (CRTS) (Drake et al., 2009), an astronomical survey searching for transient and highly variable objects. We developed a procedure for extraction and transformation from CRTS into a homogeneous data set of thousands of objects that can be used to train DL algorithms and establish benchmarks.

The remainder of this paper is organized as follows. In § 2 we describe the CRTS, together with the selection and compilation procedures. In § 3 we describe the main features of Deep-TAO, including its structure. In § 4 we describe how to connect our dataset with MANTRA (Neira et al., 2020), a light curve-based dataset built from the CRTS. Finally, in § 5 we demonstrate how Deep-TAO can be used in deep learning-based classification tasks, and then we provide a brief discussion and summary.

2. Observational Inputs to build Deep-TAO

2.1. The Catalina Real-Time Transient Survey and the Catalina Sky Survey

We retrieved the images for Deep-TAO from the public catalogs of the Catalina Real-Time Transient Survey (CTRS Drake et al., 2009; Mahabal et al., 2011), an astronomical survey for transients and highly variable objects. The area covered by the CRTS is 33,000 square degrees, and it has been observing the sky since 2007 with three telescopes: Mt. Lemmon Survey (MLS), Catalina Sky Survey (CSS), and Siding Spring Survey (SSS). We used data from the CSS telescope, an $f/1.8$ Schmidt catadioptric equipped with a 111-megapixel CCD detector. The CSS telescope and detector have a scale of 2.5 arcseconds per pixel, providing an 8 square degrees field of view. Observations were made on a grid of adjacent fields. The survey covered 4,000 square degrees per night with a limiting magnitude of 19.5 in the V-band. Each observation consist of one image obtained using an exposure time of 30 seconds.

2.2. Transient catalogs from the CRTS and the CSS

We built Deep-TAO from the public transient catalog published by CRTS. The data reports five classes: blazars (BZ), active galactic nuclei (AGN), cataclysmic variables (CV), supernovae (SN), high proper motion stars (HPM), and other events of unknown nature (Drake et al., 2009). The transient catalog lists Right Ascension (RA), Declination (Dec), V-band magnitude, discovery date, classification class, and light curve points.

The CSS catalog contains observations from 2003 to 2012. The selected fields were typically visited four times at night, and the median total number of visits over 10 years was 20. Each CSS image (of size $4,110 \times 4,096$ pixels covering an area of 29,500 square arcminutes) is divided into 1,156 smaller images called cutouts stored in the Flexible Image Transport System (FITS) format. Each cutout is about 120×120 pixels and represents an area of 5×5 arcminutes. Each cutout file stores the pixel intensities, the date on which the image was captured, a field identifier, and a number identifying the order of the image in the sequence of observations taken on a given night.

2.3. Building Regions of Interest

We used the cutouts to build a Region of Interest (RoI) centered on an object of interest. We designed RoIs to be squares of 64×64 pixels size centered on a RA/Dec coordinate of interest. This requires downloading the cutouts, assembling them into a single image, and finally cutting out the RoI around the RA/Dec of interest.

We refer to the time-ordered set of RoIs around the same coordinates as *a RoI sequence*. We built RoI sequences over a three-year interval, where the second year always included the date of maximum brightness. During this period, the time spacing between images was not uniform. The intervals ranged from days to months.

We queried the RoI sequences using web scraping techniques to automatically access and download the images using the desired RA/Dec position as an input. This process comprises five steps:

1) Download all the available cutouts that overlap with the input RA/Dec in a time span of three years for each object.

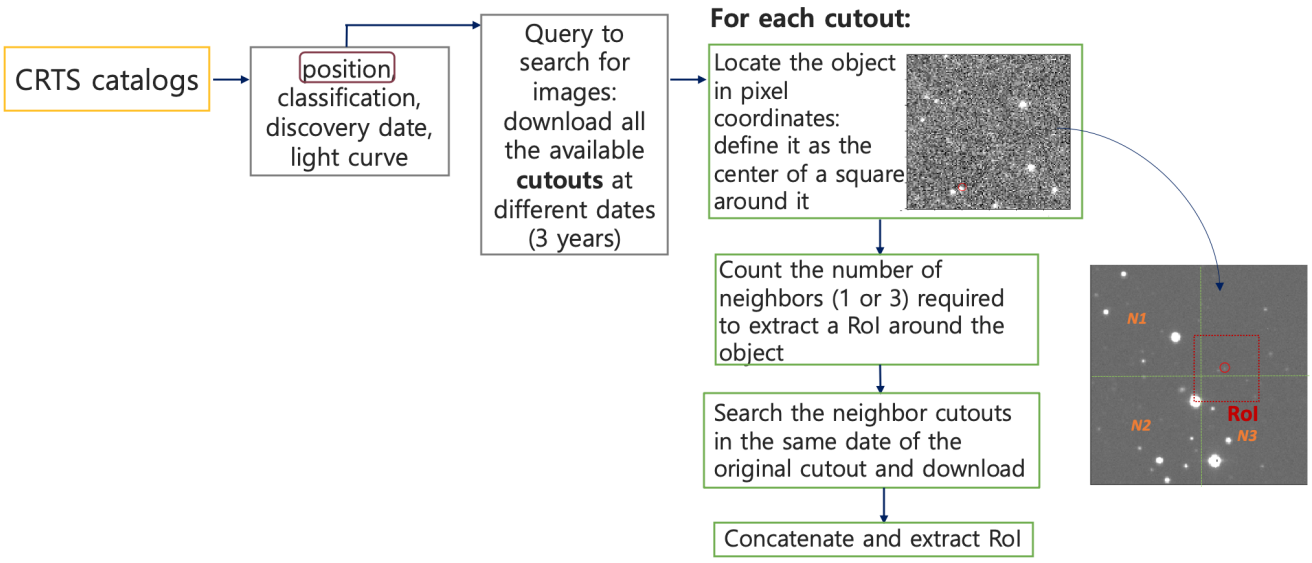

Figure 1. Overview of the search procedure to acquire the image sequences of transient objects.

2) For each cutout, locate the RA/Dec location to define a region of interest (RoI) around that coordinate.
3) Count the number of neighboring cutouts (one or three) required to build the RoI.
4) Query for the neighboring cutouts. If any of those does not exist, the RoI is not built.
5) Concatenate all cutouts to extract and store the RoI.

Figure 1 illustrates these steps. It took 11,000 CPU hours to query the CRTS/CSS database to build the full Deep-TAO dataset.

Transient objects are available in the CRTS catalogs. However, a catalog of non-transient objects is not available. To define the Non-Transient RA/Dec locations, we used transient source cutouts. All sources in the cutout of a transient at any date were detected. Then, sources at a distance greater than a threshold of 33 pixels from the transient are considered as a possible non-transient candidate.

This threshold ensures that the transient object does not appear in the RoI of the non-transient candidate. For each non-transient candidate, we computed its RA/Dec coordinates to build all the RoIs on the same dates as the parent transient sequence. Using this procedure, we compiled a total of 12,500 non-transient locations.

3. Deep-TAO Description

Figure 2 shows a grid of illustrative examples for different transients and Non-Transients in Deep-TAO. The images in that figure are a subset of the full RoI sequence for each object, the temporal spacing between images is uneven, and the time stamps are not uniform across different objects. To ease visualization, the pixel values were renormalized to have the same range across all images.

In all the cases shown in Figure 2, the variability of the central source was easy to spot by eye. This illustrative example also shows features (i.e. trails at the end of the Cataclysmic Variable sequence, overall brightness change in the first half of the Other Objects class) that might come from fluctuating observational and instrumental conditions, representing the realism of Deep-TAO.

In the following sections, we describe the overall Deep-TAO statistics, the data model used to store the information in the public repositories, and the Python-based tools to interact with Deep-TAO files.

3.1. General Statistics

Table 1 summarizes the global statistics for the Deep-TAO dataset. The first row shows the total number of targets in the original CRTS catalog. The second row indicates the number of targets for which we managed to recover a RoI sequence. Some transients in the original catalog were not included in Deep-TAO out due to the impossibility of having the transient centered in the cutout. The third row indicates the total number of RoI extracted for each class.

Figures 3, 4, 5 present some cumulative statistics computed over the RoI sequences for each class. Figure 3 shows the cumulative distribution of the number of images by sequence. The left panel shows all the transient classes, and the right panel compares transients and non-transients. This figure shows that the median value is approximately 100 RoIs per sequence. The shortest sequence had five RoIs, and the longest had approximately 300 RoIs. For non-transient sequences, there was a median of 70 RoIs, whereas for transients, the median was 100 RoIs per sequence.

Figure 4 shows the results for the average RoI signal. Here, we define the signal as the sum of all CCD counts

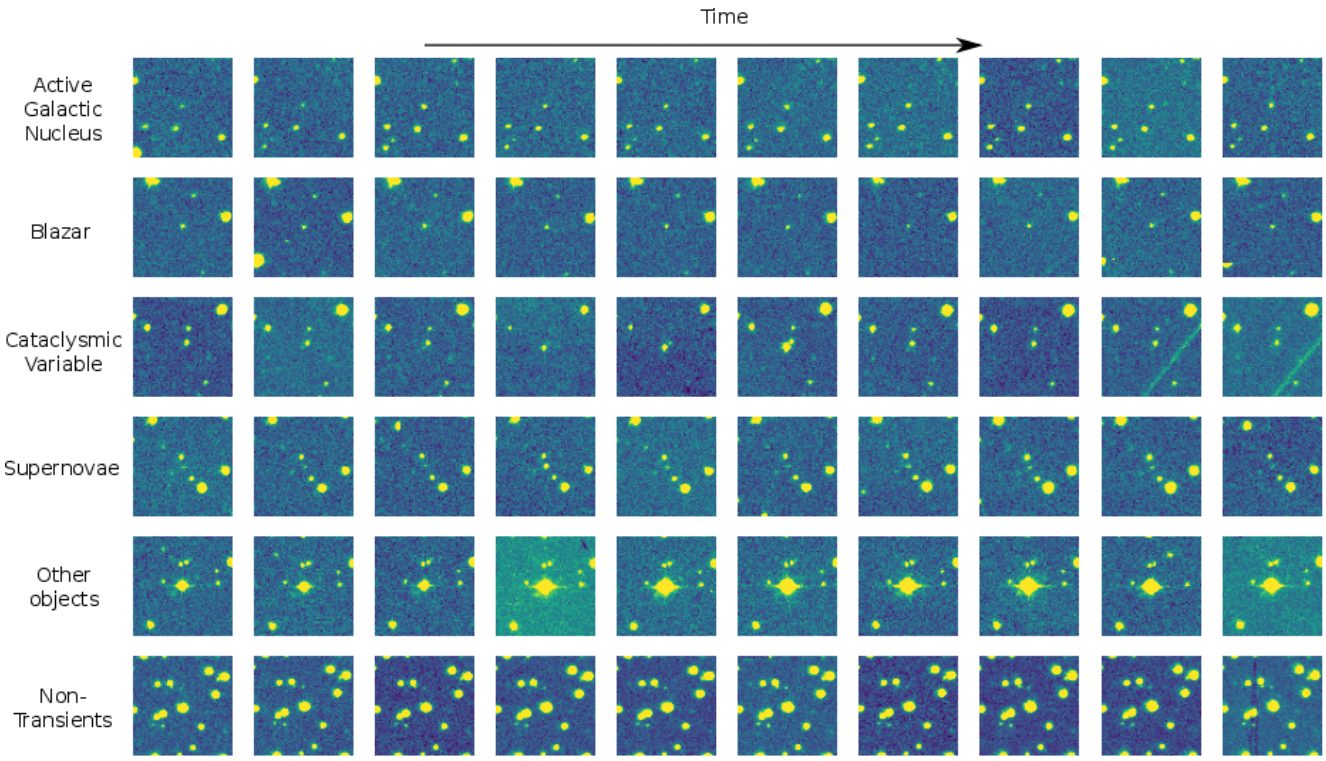

Figure 2. Sample images in Deep-TAO. Each row corresponds to a sample from a different class. The temporal spacing between consecutive images varied for each example. Images were normalized for visualization.

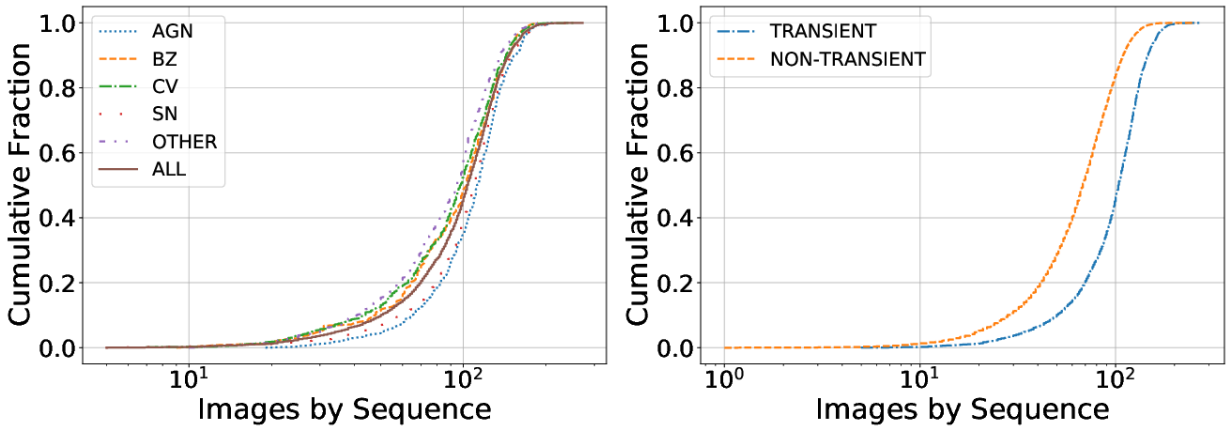

Figure 3. Cumulative distribution of RoIs per sequence. (Left) The distributions are split across the transient classes. The median was approximately 100 images per sequence. (Right) Distributions are split between transient and non-transient objects. The median for non-transients is around 70 images per sequence.

Table 1. General statistics of Deep-TAO data set*

	BZ	AGN	CV	OTHER	SN	Total Transients	Non-Transients	Total
Targets in CRTS	270	651	987	1,054	1,723	4,712	-	4,712
Targets in Deep-TAO	239	606	772	818	1,372	3,807	12,500	16,307
Total RoIs	23,429	66,998	73,739	74,536	146,847	385,549	863,530	1,249,079

*The first row corresponds to the transients included in the public CRTS transient catalog. The second row represents the number of objects for which a sequence of RoIs can be retrieved over a three-year observation period. The last row is the total number of RoIs included for each class.

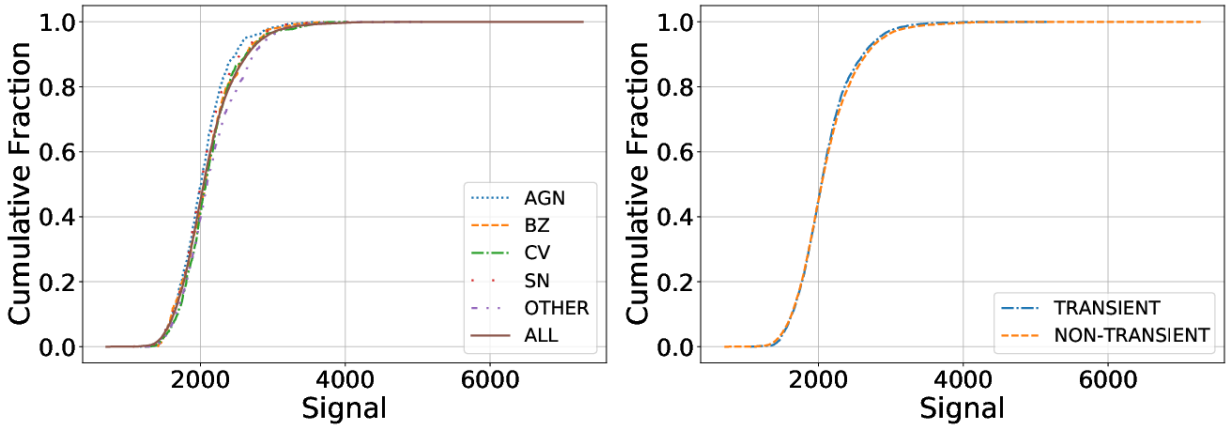

Figure 4. (Left) Cumulative fraction as a function of the median signal for the objects in each transient class and all transients objects (continuous line). (Right) Cumulative fraction between transient and non-transient objects. The shapes of these classes were similar. In both figures the media of the signal is around 2000.

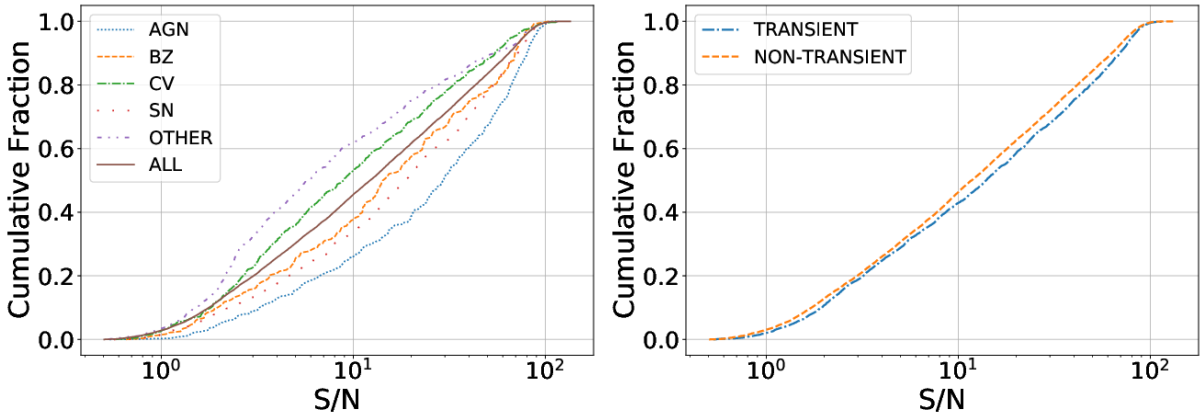

Figure 5. (Left) Cumulative distribution of the average signal-to-noise. for the object in each transient class and all transient objects (continuous line). The media of the signal/noise for all transient objects is approximately 10. (Right) Cumulative fraction between transients and non-transients objects. In both cases the media of signal/noise is around 20.

across the RoI. The left panel corresponds to all transient classes, whereas the right panel compares transients and non-transients. This figure shows that all transient classes and non-transients have similar intensity distributions.

Figure 5 shows a comparison of the average signal-to-noise (S/N) distribution for all transient classes (left) and transients versus non-transients (right). We estimated the signal-to-noise ratio for an RoI as the ratio between the sum of all CCD counts and the standard deviation of the CCD counts.

We found that the average S/N spans almost two orders of magnitude, ranging from 1 to 100. For transients, the median of the average S/N ranged between 6 and 20 across all classes, with some differences between classes. In contrast, the distributions of Transients and Non-Transients were virtually the same.

3.2. Data Model

The Deep-TAO data set is allocated on GitHub into two different repositories, one for transients objects[1] and other for non-transients[2]. The transient repository contains three main folders: `data`, `paper`, and `mantra`.

The `data` folder contains all the transient sequences separated in subfolders by class (AGN, BZ, CV, OTHERS, and SN), each subfolder contains the sequences stored in FITS files. A single FITS file stores all the RoIs associated with a transient event, and the file name is the CRTS identifier. Each file contains a header, and the FITS header in each file has minimal identifying information, such as the CRTS_ID unique identifier, the J2000 RA/Dec coordinates, the number of RoIs (`N_Images`) in the sequence, and the Universal Time UT_Date associated with the discovery date. The full list of fields included in the header is presented in Table 2.

[1] https://github.com/MachineLearningUniandes/TAO_transients
[2] https://github.com/MachineLearningUniandes/TAO_non-transients

Table 2. FITS header of the transient files

Header Dict	Description	Type
CRTS_ID	Catalina Real-time Transient Survey ID	str
RA_(J2000)	Right Ascension (degrees)	float
Dec_(J2000)	Declination (degrees)	float
N_Images	Total number of images for CRTS ID	int
UT_Date	UT Discovery Date (YYYYMMDD)	float
Mag	Unfiltered CSS magnitude	float
CSS_Images	Pre and post-discovery images ID	int
SDSS	Covered by SDSS DR-12 (yes/no)	str
Others	ID to other image data at the location (PQ, DSS, 2MASS, SDSS)	int
Followed	P60 follow up (yes/no)	str
Last	Last Observation date	str
LC	Current CSS lightcurve	int
FC	Finding chart (yes/no)	str
Class	Transient classification	str

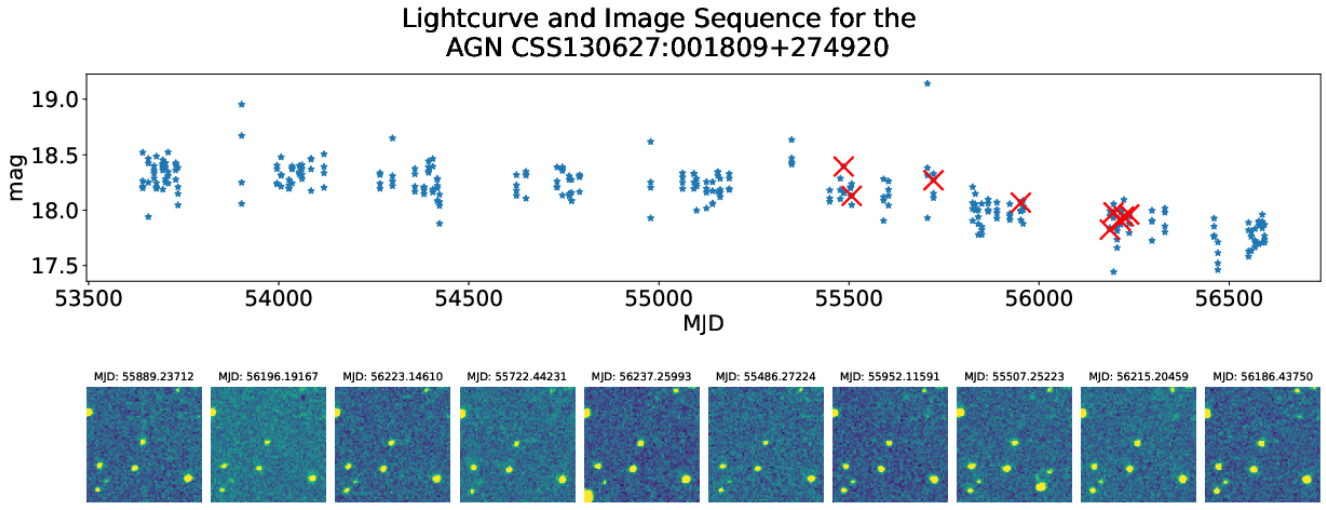

Figure 6. Lightcurve and examples of the image sequence for the AGN CSS130627:001809+274920 from MANTRA and Deep-TAO obtained using the `Connection_MANTRA` Jupyter notebook. The red cross correspond to the images plotted bottom in the figure.

The first HDU (extension 1) in the FITS files is a 2D array with the columns listed in Table 3. This array contains information for each RoI in the sequence, such as the HDU extension for each RoI and the observation date. Starting from HDU 2 onward to HDU `N_Images+1`, each HDU contains an RoI as an integer array of size 64×64.

The second main folder is `paper`, which contains Figures 3, 4, 5 which describe the general statistics of Deep-TAO. This folder also contains a Python-based tool to reproduce these results. This tool is explained in the next subsection.

The `mantra` folder contains the Figure 6, which shows an example of how to connect Deep-TAO with MANTRA (Many ANnotated TRAnsients), an annotated Machine-learning Reference lightcurve dataset in the V-band also built from the CRTS (Neira et al., 2020). More details are provided in Section 4.

Finally, the non-transient repository only contains the `data` folder with the FITS files of the non-transient objects. In contrast to the header of a Transient FITS file, the non-transient FITS header allocates the information of Table 4, which includes the CRTS_ID, the RA/Dec coordinates, the number of images in the sequence, and the image source from which it was extracted.

The first HDU in each FITS non-transient file allocates the information in Table 5: the HDU_Extension for each RoI, the date of observation, the MJD, the Field_ID, and the cutout.

Table 3. Identifiers stored first HDU: Transient files*

Key	Description	Type
HDU_Ext	HDU extension of the RoI (From 2 to `N_Images+1`)	int
Set_Number	Stands for the sequence (or set number)	str
Date	Date of observation (YYMMMDD)	str
MJD	Modified Julian Date	float
Field_ID	Field identifier	str
Obs_In_Seq	Refers to the observation's number in the sequence	str
Cutout	Cutout matrix location. Each cutout covers an area of about 5×5 arcminutes	str

*Basic information in the first HDU about the image sequence in each transient FITS file.

Table 4. FITS header of the non-transient objects

Header Dict	Description	Type
CRTS_ID	Catalina Real-time Transient Survey ID	str
RA_(J2000)	Right Ascension (degrees)	float
Dec_(J2000)	Declination (degrees)	float
N_Images	Total number of images for CRTS ID	int
Img_Ref	Image of reference where the non-transient object was identified	str

Table 5. Identifiers stored first HDU: Non-transient files*

Key	Description	Type
HDU_Ext	HDU extension of the RoI (From 2 to `N_Images+1`)	int
Date	Date of observation (YYMMMDD)	str
MJD	Modified Julian Date	float
Field_ID	Field identifier	str
Cutout	The cutout matrix location Each cutout covers an area of about 5×5 arcminutes	str

*Basic information in the first HDU of the Non-Transient objects about the image sequence in each FITS file.

3.3. Python-based tools

In the folder `data` of the transients repository, there is a Jupyter notebook to manipulate the data. The `Read_dataset` Jupyter notebook shows the mechanism for reading the FITS files for transient and non-transient objects. In the folder `paper` in the same repository, we provide the `Explore_data set` Jupyter notebook, this shows how to compute some statistics from Deep-TAO to obtain the Figures 2, 3, 4 and 5, assuming that the `data/NON` folder from the non-transient's repository is located in the `data` folder of the transient's repository. This notebook also creates a plain text file in the `paper` folder called `statistic.csv`. This file has 16,307 rows, one by object in Deep-TAO, and four columns with the class name `class` (BZ,AGN,CV,OTHER,SN, or NON), the number of images by sequence `nimages_seq`, the median of the signal/noise measure `signal_noise_median`, and the median of the signal `signal_median`.

4. Linking Deep-TAO images to MANTRA lightcurves

In Neira et al. (2020), the authors presented MANTRA, an annotated machine-learning reference light curve dataset also built from the CRTS. MANTRA contains 4,869 transients and 71,207 non-transients as a plain text file to facilitate a standardized quantitative comparison of astronomical transient event recognition algorithms. The classes included in MANTRA are supernovae, cataclysmic variables, active galactic nuclei, high proper motion stars, blazars, and flares. The data set is publicly available and easy to access [3].

In the `mantra` folder of the Deep-TAO transients repository[4], we provide the `Connection_MANTRA` Jupyter notebook to link the image sequence from Deep-TAO to the lightcurve from MANTRA. This connection is established through the unique CRTS ID. For non-transients, this connection between images and light curves cannot be

[3] https://github.com/MachineLearningUniandes/MANTRA
[4] https://github.com/MachineLearningUniandes/TAO_transients

Table 6. F-measure for the binary task*

Set	Data	Model	Transient	Non-Transient	F1 ($\mu \pm \sigma$)
Validation	Images	TAO-Net	74.46	95.06	84.76 ± 10.30

*F-measure for each class in the validation set for the binary task. The last column reports the average F-measure.

Table 7. F-measure for the transient classification*

Set	Data	Model	BZ	AGN	CV	OTHER	SN	F1 ($\mu \pm \sigma$)
Validation	Light curves	RF	19.74	42.67	53.60	56.06	55.36	45.49 ± 13.75
Validation	Images	CNN	25.17	49.77	59.48	64.04	63.39	52.37 ± 14.53

*The last column reports the average F-measure of the 5 transient categories.

established between Deep-TAO and MANTRA because both have different non-transient objects.

Figure 6 shows an example of an AGN. Using the MJD information, it is possible to connect points in the light curve to images in the sequence. In the light curve of Figure 6, the red crosses correspond to the images plotted below in this figure. Due to the constraints in the RoI construction (Section 2), not all points in the MANTRA lightcurve have a corresponding image in Deep-TAO. Another reason is that Deep-TAO includes only three-year intervals of observations.

5. Example of a Deep-TAO application

Here, we show some examples of Deep-TAO applications using a Convolutional Neural Network (CNN) to gauge its performance on three basic classification tasks:

1. binary classification between Transients and Non-Transients.
2. fine-grained classification into five transient classes (Blazar, AGN, Cataclysmic Variables, Supernovae, and Other)
3. fine-grained classification into five transient classes and Non-Transients as a sixth class.

We evaluated all tasks using metrics that were robust to class imbalances. For each class, we report the maximum F-measure (F1) from the Precision-Recall (PR) curve that we constructed by setting different thresholds on the output probabilities of each class. The global performance is the F1 average across individual classes, with an uncertainty computed as the standard deviation. In all experiments, we used 70% of Deep-TAO data for training, 25% for validation, and 5% for testing.

The CNN used here is based on the previous work by Gómez et al. (2020). They used TAO-Net, a neural network composed of two modules. First, a CNN based on the DenseNet architecture is used to extract a feature representation, and then a Recurrent Neural Network (RNN) that uses these representations to solve the classification task. Here, we only use the first part, a CNN based on a Densely Connected Convolutional Network (DenseNet) (Huang et al., 2017) with $L = 70$ layers and a growth rate $k = 32$.

We model temporal information by selecting images from the complete sequences. We considered images at three different dates in sequential order, such that they reflect differences in brightness for transient classes. We included the observation date in the three-year period when the transient object had the maximum brightness and one observation before and after that date. For the Non-Transient class, we considered the first, middle, and last dates of the sequence of ordered images. At each date, we took the first available observation and then merged the temporal information by sampling images from the complete sequences at three different dates in sequential order. This selection reflects the evolution of temporal information, evidencing the differences in brightness for transient classes.

Table 6 summarizes the results of the binary classification tasks. As expected, it was considerably easier to classify a sequence as non-transient (F1 of 95.06) than as transient (F1 of 74.46).

For the five-class transient classification task, we performed an experiment that consisted of the traditional approach for transient classification using the light curves from the CRTS. We computed the discriminatory features from the light curves to train a Random Forest (RF) classifier. All details on feature extraction and the RF classifier can be found in Neira et al. (2020). These results are equal to those of Gómez et al. (2020) because we share the same dataset and algorithm parameters.

Table 7 lists the F-scores of the transient classification tasks. The results show that classification with images using a CNN is a better option that makes a classification with light curves using a RF algorithm. With RF on the light curves, the best classification was for the OTHER class with 56.06, followed by the SN class with 55.36. The worst is the BZ class with 19.74, and the average F1-score is 45.49. The CNN on images is better with an average F1-score of 52.37, where the best classification is for OTHER with 64.04,

Table 8. F-measure for the multi-class detection*

Set	Data	Model	BZ	AGN	CV	OTHER	SN	Non-T	F1 ($\mu \pm \sigma$)
Validation	Images	CNN	21.82	37.45	54.76	40.22	46.59	95.29	49.36 ± 22.84

*The last column reports the average F-measure of the 6 classes.

followed by SN with 63.39, and the worst classification is for BZ with 25.17.

Finally, in Table 8 we present the F-scores of the multi-class classification problem, which includes the five transient classes and the non-transient class using only the CNN method with images. Compared with the previous task, the overall performance was worse for every transient class, indicating that this task is more difficult when non-transient objects are included. The F-measure shows that the best classification is for the non-transient class, with a score of 95.29. The best transient class classified correctly was the CV, with a score of 54.76, followed by the SN class, with a score of 46.59.

6. Conclusions

There is increasing interest in automated methods for detecting transient sources. Some of these methods are based on Deep Learning techniques that require the use of large, realistic datasets for training. Publicly available and easily accessible datasets can trigger the development of new deep learning applications for transient detection.

In this study, we present such a dataset. We named it Deep-TAO, for deep-learning transient astronomical objects. This is the first public and easily accessible dataset based on real images that can be used to train and improve Deep Learning algorithms for transient classification. The dataset is a compilation of images extracted and transformed from the Catalina Real-Time Transient Survey (CRTS). Deep-TAO includes 3,807 transient and 12,500 non-transient objects with a total of 1,249,079 real astronomical images. Deep-TAO is publicly available at https://github.com/MachineLearningUniandes/.

We demonstrated the utility of Deep-TAO using a set of deep learning experiments and comparisons against a machine learning algorithm. We explored the transient versus non-transient task, the fine-grained multi-classification task between five transient classes, and finally a fine-grained multi-classification task with six classes, five transient classes, and non-transient as another class.

In the three tasks we used the same architecture, a Densely Connected Convolutional Network with $L = 70$ layers and a growth rate $k = 32$ motivated by the more complex architecture proposed by Gómez et al. (2020). In the fine-grained multi-classification task between five transient classes, we compared a classification based on a CNN with images and the classification of light curves with a random forest with 200 trees based on the work by Neira et al. (2020). The results showed that CNN consistently performed better.

Deep-TAO is public with files in the FITS format to facilitate its usability in different projects. The realism of Deep-TAO provides an additional motivation to train new learning-based models to be used by next-generation experiments in time-domain astronomy, and hopefully, it will also motivate the creation of more datasets with a similar structure: realistic, fully labeled, open, and easy to access.

The authors thank the Office of the Vice Rector for Research at the Universidad de los Andes for supporting this project by the grant SPATIO TEMPORAL TRANSIENT OBJECT /P17.246622.004/01. JFSP and JEFR acknowledge the support of INV-2021-126-2256 and INV-2022-137-2394 projects of the Universidad de Los Andes, Facultad de Ciencias. We also thank contributors and collaborators of the open-source packages fundamental to our work: NumPy (Van Der Walt et al., 2011), the Jupyter notebook (Kluyver et al., 2016), matplotlib (Hunter, 2007) and pandas (McKinney et al., 2010). The CRTS and CSDR2 are supported by the U.S. National Science Foundation under NSF grants AST-1313422, AST-1413600, and AST-1518308. The CSS survey is funded by the National Aeronautics and Space Administration under Grant No. NNG05GF22G was issued through the Science Mission Directorate Near-Earth Objects Observations Program.

Data Availability

The Deep-TAO data set is publicly available at https://github.com/MachineLearningUniandes/ into two different repositories, one for transients objects (https://github.com/MachineLearningUniandes/TAO_transients) and other for non-transients (https://github.com/MachineLearningUniandes/TAO_non-transients).

References

Allam, T., & McEwen, J. D. 2024, RAS Techniques and Instruments, 3, 209, doi: 10.1093/rasti/rzad046

Bellm, E. C., Kulkarni, S. R., Graham, M. J., et al. 2019, PASP, 131, 018002, doi: 10.1088/1538-3873/aaecbe

Cabrera-Vives, G., Reyes, I., Förster, F., Estévez, P. A., & Maureira, J.-C. 2017, ApJ, 836, 97, doi: 10.3847/1538-4357/836/1/97

Carrasco-Davis, R., Cabrera-Vives, G., Förster, F., et al. 2019, PASP, 131, doi: 10.1088/1538-3873/aaef12

Chambers, K. C., Magnier, E. A., Metcalfe, N., et al. 2016, arXiv e-prints, arXiv:1612.05560, doi: 10.48550/arXiv.1612.05560
D'Isanto, A., Cavuoti, S., Brescia, M., et al. 2016, MNRAS, 457, 3119, doi: 10.1093/mnras/stw157
Drake, A. J., Djorgovski, S. G., Mahabal, A., et al. 2009, AJ, 696, 870, doi: 10.1088/0004-637X/696/1/870
Dyer, M. J., Steeghs, D., Galloway, D. K., et al. 2020, in Ground-based and Airborne Telescopes VIII, ed. H. K. Marshall, J. Spyromilio, & T. Usuda, Vol. 11445, International Society for Optics and Photonics (SPIE), 114457G, doi: 10.1117/12.2561008
Gieseke, F., Bloemen, S., van den Bogaard, C., et al. 2017, MNRAS, 472, 3101, doi: 10.1093/mnras/stx2161
Gómez, C., Neira, M., Hoyos, M. H., Arbeláez, P., & Forero-Romero, J. E. 2020, MNRAS, 499, 3130, doi: 10.1093/mnras/staa2973
Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. 2017, in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2261–2269, doi: 10.1109/CVPR.2017.243
Hunter, J. D. 2007, CSE, 9, 99, doi: 10.1109/MCSE.2007.55
Kaiser, N. 2004, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 5489, Ground-based Telescopes, ed. J. Oschmann, Jacobus M., 11–22, doi: 10.1117/12.552472
Killestein, T. L., Lyman, J., Steeghs, D., et al. 2021, MNRAS, 503, 4838, doi: 10.1093/mnras/stab633
Kluyver, T., Ragan-Kelley, B., Pérez, F., et al. 2016, in IOS Press, 87–90, doi: 10.3233/978-1-61499-649-1-87
Law, N. M., Kulkarni, S. R., Dekany, R. G., et al. 2009, PASP, 121, 1395, doi: 10.1086/648598
Mahabal, A. A., Djorgovski, S. G., Drake, A. J., et al. 2011, BASI, 39, 387. https://arxiv.org/abs/1111.0313
Martínez-Palomera, J., Förster, F., Protopapas, P., et al. 2018, AJ, 156, 186, doi: 10.3847/1538-3881/aadfd8
McKinney, W., et al. 2010, in Proceedings of the 9th Python in Science Conference, Vol. 445, Austin, TX, 51–56
Muthukrishna, D., Narayan, G., Mandel, K. S., Biswas, R., & Hložek, R. 2019, PASP, 131, 118002, doi: 10.1088/1538-3873/ab1609
Neira, M., Gómez, C., Suárez-Pérez, J. F., et al. 2020, ApJS, 250, 11, doi: 10.3847/1538-4365/aba267
Nidever, D. L., Dey, A., Fasbender, K., et al. 2021, AJ, 161, 192, doi: 10.3847/1538-3881/abd6e1
Sánchez-Sáez, P., Reyes, I., Valenzuela, C., et al. 2021, AJ, 161, 141, doi: 10.3847/1538-3881/abd5c1
Scalzo, R. A., Yuan, F., Childress, M. J., et al. 2017, PASA, 34, e030, doi: 10.1017/pasa.2017.24
Smartt, S. J., Valenti, S., Fraser, M., et al. 2015, A&A, 579, A40, doi: 10.1051/0004-6361/201425237
Van Der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, CSE, 13, 22, doi: 10.1109/MCSE.2011.37
Van Roestel, J., Duev, D. A., Mahabal, A. A., et al. 2021, AJ, 161, 267, doi: 10.3847/1538-3881/abe853
Wyrzykowski, Ł., Kostrzewa-Rutkowska, Z., Kozłowski, S., et al. 2014, AcA, 64, 197, doi: 10.48550/arXiv.1409.1095