Solar Physics (2023) 298:82
https://doi.org/10.1007/s11207-023-02171-0

RESEARCH

Automatic Burst Detection in Solar Radio Spectrograms Using Deep Learning: deARCE Method

Javier Bussons Gordo[1,2] · Mario Fernández Ruiz[3] · Manuel Prieto Mateo[3] · Jorge Alvarado Díaz[4] · Francisco Chávez de la O[5] · J. Ignacio Hidalgo[4] · Christian Monstein[6]

Received: 22 December 2022 / Accepted: 19 May 2023

Abstract
We present in detail an automatic radio-burst detection system, based on the AlexNet convolutional neural network, for use with any kind of solar spectrogram. A full methodology for model training, performance evaluation, and feedback to the model generator has been developed with special emphasis on i) robustness tests against stochastic and overfitting effects, ii) specific metrics adapted to the unbalanced nature of the solar-burst scenario, iii) tunable parameters for probability-threshold optimization, and iv) burst-coincidence cross match among *e-Callisto* stations and with external observatories (NOAA-SWPC). The resulting neural network configuration has been designed to accept data from observatories other than *e-Callisto*, either ground- or spacecraft-based. Typical False Negative and False Positive Scores in single-observatory mode are, respectively, in the $10-16\%$ and $6-8\%$ ranges, which improve further in cross-match mode. This mode includes new services (deARCE, Xmatch) allowing the end-user to check at a glance if a solar radio burst has taken place with a high level of confidence.

Keywords e-Callisto · Solar radio burst · Spectrogram · Deep learning

1. Introduction

Solar radio bursts (SRBs) are a valuable tool for fundamental solar physics and space weather. These transient enhancements of solar radio emission contain relevant informa-

J. Bussons Gordo
javier.bussons@uah.es

[1] FISPAC Group, Physics Department, Universidad de Murcia, 30071 Murcia, Spain

[2] Physics and Mathematics Department, Universidad de Alcalá, 28805 Alcalá de Henares, Spain

[3] Space Research Group, Universidad de Alcalá, 28805 Alcalá de Henares, Spain

[4] Adaptive and Bioinspired Systems Research Group, Universidad Complutense de Madrid, 28040 Madrid, Spain

[5] Universidad de Extremadura, 06800 Mérida, Spain

[6] Istituto ricerche solari Aldo e Cele Daccò (IRSOL), Università della Svizzera italiana, Locarno, Switzerland

Published online: 19 June 2023

tion about the physical processes behind them, including magnetic reconnection, electric-current induction or electron acceleration in the solar corona, storage, explosive release, and transport of energy and particles across the interplanetary medium toward their final interaction with the Earth's magnetosphere. Statistical studies of observational SRB parameters and their correlation with solar flares or coronal mass ejections (CME) may establish the link between physical processes occurring in the solar atmosphere, at various heights in the corona, and out into the interplanetary medium (Mahender et al., 2020). When studied on long, yearly scales, they may also shed light on the mysteries of the solar activity cycle. Moreover, SRBs act as early signatures of interplanetary plasma disturbances such as CMEs and solar energetic particle (SEP) bursts, which take longer to reach the Earth and may cause damage to radio communications, satellite data, aircraft and spacecraft navigation, and astronaut health. Thus round-the-clock monitoring of solar radio emission becomes a tool for space-weather forecasting (Klein, Salas Matamoros, and Zucca, 2018 and references therein, Hou et al., 2020; Ma et al., 2022).

Moreover, solar studies in the radio domain are advantageous because they can be made with rather simple instruments – antennas of different types – which can operate under cloudy skies with a high duty cycle and, being protected by the Earth's magnetosphere and atmosphere, can issue real-time space-weather hazard alerts.

e-Callisto (e-callisto.org, Benz et al., 2009), the international network of *Compound Astronomical Low-frequency Low-cost Instruments for Spectroscopy and Transportable Observatories*, can play a key role in such 24-hour monitoring scheme. Its main strengths are: 24-hour coverage of the Sun provided by nearly one hundred stations deployed worldwide thanks to their low cost; geo-redundancy, allowing for event cross matching among a number of stations observing the Sun at any given time; and low-frequency radio coverage to trace the acceleration of electrons in the solar corona and study the fine-scale structure of different types of bursts (Zucca et al., 2012; Ndacyayisenga et al., 2021). Its native dynamic range, 45 – 870 MHz, can be extended down to 20 MHz with enough sensitivity to detect important events, such as reverse-drift and J-type bursts in the 20 – 85 MHz range not covered by other types of instruments (Klein et al., 2022). Geo-redundancy and worldwide coverage make it an excellent tool for cross-match studies between ground-based and spacecraft-borne detectors.

At present, burst identification in *e-Callisto* data is carried out through daily visual inspection of thousands of spectrograms (typically 40 per instrument per day, times 70 instruments) by an expert on duty, who then produces an event list. Monthly *e-Callisto* event reports since 2020 are available on the World Wide Web (e-callisto.org). This tedious task calls for an automatic system, which is the object of this article.

Deep learning emerges as a suitable tool for such automatization as it can adopt a computer-vision approach with images – spectrograms in our case – as input signals for a cascade of convolutional layers capable of finding their best abstract representation in terms of class-discriminating power.

Deep learning deals with high-level abstraction by using computational architectures that transform input signals to extract non-explicit features, which may provide the answer to a complex problem. It can work with signals of all kinds: music, video, text, images, etc. using algorithms based on deep-belief networks, deep neural networks (multimodal or otherwise), convolutional neural networks, etc. For more information, see Moujahid (2016), Ongsulee (2017), and Zhang et al. (2020).

Several works can be found in the literature dealing with automatic radio-burst identification in dynamic spectra, using either statistical approaches (Lobzin et al., 2009, 2010; Singh et al., 2019; Afandi et al., 2020) or different kinds of artificial intelligence (Chen et al., 2016;

Scully et al., 2021; Guo et al., 2022). In this work, dynamic spectrograms representing solar radio observations are passed as images to a convolutional neural network (CNN) charged with the task of automatically classifying them as bursts or not-bursts with an efficiency close to or better than that of a human.

In Section 2, the origin (observatories, instruments), location, format, and characteristics of our input data are described, and two historical periods are defined depending on whether expert burst reports based on human inspection exist (2020 – present) or not (2012 – 2019).

In Section 3, we present in detail the three phases of our automatic burst identification method: preprocessing of input spectrograms, training of the deep-learning classifier (model), and application to the target database. We pay special attention to careful generation of the training database, performance evaluation mechanisms, and proper feedback of such evaluation to the model before it can be applied to a target database. Emphasis is made in the definition of new metrics adapted to our practical goal: render the *e-Callisto* data more useful by providing the end user, in near-real time, with a list of runs containing all the relevant bursts with a minimum of false positives. This section ends with a description of the products supplied by the system: new reports in the case of reportless periods, updated reports otherwise, and in all cases, cross-match plots of events coincident in several stations.

In Section 4, results are shown for the performance evaluation and post-feedback re-evaluation of selected models, including both single- and multi-observatory (hybrid) models. A sample of bursts that had passed unnoticed to the human eye is given as a supplement.

In Section 5, we provide some conclusions and future directions.

2. Data: Solar Radio Spectrograms and Burst Reports

Dynamic spectra, heretofore spectrograms (Figure 1), from *e-Callisto* stations on the Earth dayside are continuously uploaded to a central server located at the FachHochschule Nord-Westschweiz – at the time of writing, a second server is being installed in Casa del Doncel, Sigüenza, Spain – and made available on the *e-Callisto* web site as FITS-format files (e-Callisto/Data/data.htm). Aside from an ASCII header with general information such as observatory, instrument, and run time, the file features a 3D binary table with the spectrogram itself: the signal intensity in analogue-to-digital-converter units for each given time and frequency. Canonical data files are matrices of dimension 3600 × 200 with 3600 time stamps and 200 frequencies. Standard observing runs span 15 minutes with a temporal step of 0.25 seconds (four "pixels" per second); during this time, the CALLISTO spectrometer is capable of making a full sweep of 200 different frequencies across the whole dynamic range. Dynamic ranges of *e-Callisto* instruments vary within the 20 – 850 MHz interval.

Figure 2 shows the ranking of the best performing *e-Callisto* observatories (stations) in 2021 in terms of the number of visually detected bursts. For the development of our burst identification method, the top-four observatories were selected: ASSA (Astronomical Society of South Australia), Glasgow (University of Glasgow, UK), Humain (Royal Observatory of Belgium), and Landschlacht (Switzerland). As many *e-Callisto* observatories host more than one instrument, an instrument ID number (focus code) is appended to the observatory name: ASSA_02, Landschlacht_01 (changed to 62 since 04 October 2021), and Glasgow_59 belong to the same frequency group (20 – 85 MHz), whereas Humain_59 is used for studies in the 45 – 435 MHz range.

Cross-match tests at European geo-longitudes include data from stations at Birr Castle (Ireland), Graz (Austria), and Heiterswil (Switzerland). With a view to future extension

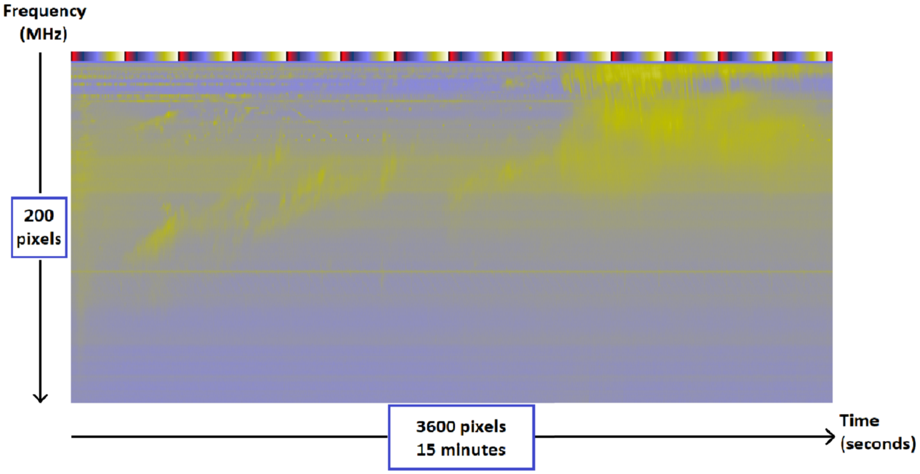


Figure 1 *e-Callisto* spectrogram (signal intensity for each time and frequency) showing a solar radio burst.

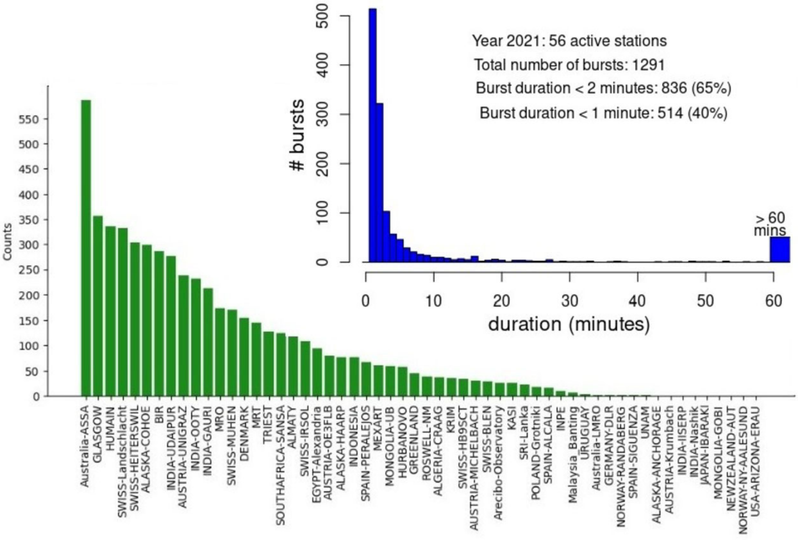


Figure 2 (**a**) Ranking of the best performing *e-Callisto* stations in 2021 in terms of the number of visually detected bursts. (**b**) Burst duration.

of our work to the period without reports (2012 – 2019) for certain tests, we have incorporated stations in Africa (Rwanda_59, University of Rwanda) and America (Roswell_58, New Mexico, USA), which at the time ranked high in burst detection. Frequency ranges and geographical coordinates for all the stations used, ordered by longitude, are listed in Table 1.

A wealth of solar radio observations has been, and is being, collected by *e-Callisto* observatories operating as a worldwide network since 2012. In this article, the 2021 *e-Callisto* FITS database is used in the training and application of the burst classifier. For our purposes, two different periods must be considered depending on whether burst reports based

Table 1 *e-Callisto* instruments used in this article ordered by geographical longitude: coordinates and selected frequency range.

Instrument	Longitude [°]	Latitude [°]	Frequency [MHz]
ASSA_02 (Australia)	139.64	−34.66	25 – 80
Rwanda_59	30.07	−1.95	45 – 81
Unigraz_01 (Austria)	15.49	47.07	47 – 80
Landschlacht_01/62 (Switzerland)	9.24	47.63	18 – 84
Heiterswil_59 (Switzerland)	9.13	47.30	47 – 80
Humain_59 (Belgium)	5.26	50.19	45 – 435
Glasgow_59 (United Kingdom)	−4.30	55.90	45 – 80
Birr_01 (Ireland)	−7.92	53.09	20 – 87
Roswell_58 (USA)	−104.52	33.44	20 – 90

on expert visual inspection are routinely issued, as they may be used as the ground truth for burst occurrence. Thus we can distinguish between the 2012 – 2019 reportless period and the 2020 – present documented period. According to its shape and extent in frequency and time, an SRB is assigned one of five major spectral types (I to V). Throughout this article, we extensively use the daily burst reports issued by *e-Callisto* with date, time interval, burst type, and stations involved in the detection (soleil.i4ds.ch/solarradio/data/BurstLists/2010-yyyy_Monstein).

In some cases, we also use the daily Solar Event Lists issued by the US National Oceanic and Atmospheric Administration's Space Weather Prediction Center (NOAA-SWPC), which include beginning-maximum-end event times, reporting observatory, a quality tag, event type, heliographic location, frequency, and additional information. Participating observatories include ground-based stations in Australia (Culgoora, Learmonth), the USA (New Mexico, Hawaii, Puerto Rico, Massachusetts), and Italy (San Vito) and spacecraft-borne instruments (GOES). Event types include radio bursts and storms as well as optical or X-ray flares. For more information, visit www.swpc.noaa.gov.

3. deARCE Method

The automatic burst identification method presented here (deARCE, pronounced /de ˈa r θ e/, deep Automatic Radioburst Compilation Engine) begins with the download of the input spectrograms from the web server via a Python script, a process which takes approximately half a minute per observatory per day on an average laptop. A day's worth of data from a given *e-Callisto* instrument is around 10 Mb on average – a bit more on long Summer days and less on Winter days.

The method is divided into three stages: preprocessing of input spectrograms, training of a burst/not-burst classifier (model) using a training dataset, and classification of the target dataset (burst detection).

3.1. Preprocessing

Each input 15-minute FITS spectrogram is chopped into fifteen 1-minute png frames via a preprocessing stage (Figure 3a), which includes background subtraction and frequency-range selection. Given that we end up working with png files, we will often refer to spectrograms as images and to time or frequency matrix elements as pixels.

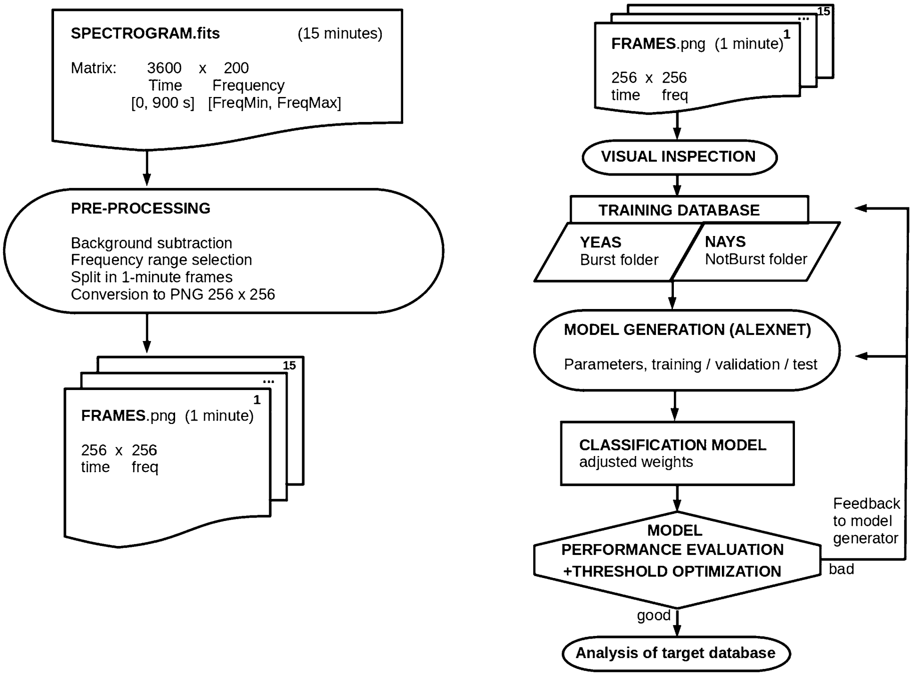


Figure 3 (**a**) Preprocessing of input spectrograms. (**b**) Steps in the training process: model generation and evaluation.

Several background subtraction schemes have been tried out, such as subtraction of a constant value from all pixels (either minimum, mean, or median intensity; also no subtraction at all) or subtraction of the average spectrum over time, i.e. the mean of 3600 spectra. After careful evaluation, we have settled on the latter because it eliminates noise sources that are constant along the run time and show up as horizontal bands of constant frequency in the spectrograms.

A frequency range of study can be selected to avoid observatory, or even instrument-dependent, noise-swarmed frequencies (for instance, the bad channel at the top of Figure 1). Table 1 shows the ranges selected for this work.

To better fit the images to the neural network input size (256 × 256 pixels) and to minimize resolution losses, the spectrograms are split into one-minute images, heretofore *frames* of $3600/15 = 240$ pixels in time and 200 pixels in frequency, before being resized to 256 × 256 (Figure 4). Apart from fulfilling the network requirement, this operation renders burst signals more visible, as they become the main feature of the frame they belong to. Most bursts have short durations (65% of them last less than two minutes, Figure 2b), thus fitting well in one or two one-minute frames; longer bursts may span several frames, but for our purposes, it will suffice to recognize them in any one of the frames.

The output of the preprocessing stage is thus a set of 256 × 256 png frame files, which are then transferred to the site where the CNN-based classification system resides. For reference, we provide representative numbers for the transfer of one day's worth of data from one observatory by an ftp-client such as FileZilla or WinSCP: 40 data runs, 15 png frames per run, 20 kb per frame, i.e. 12 Mb, transferred in 15 seconds.

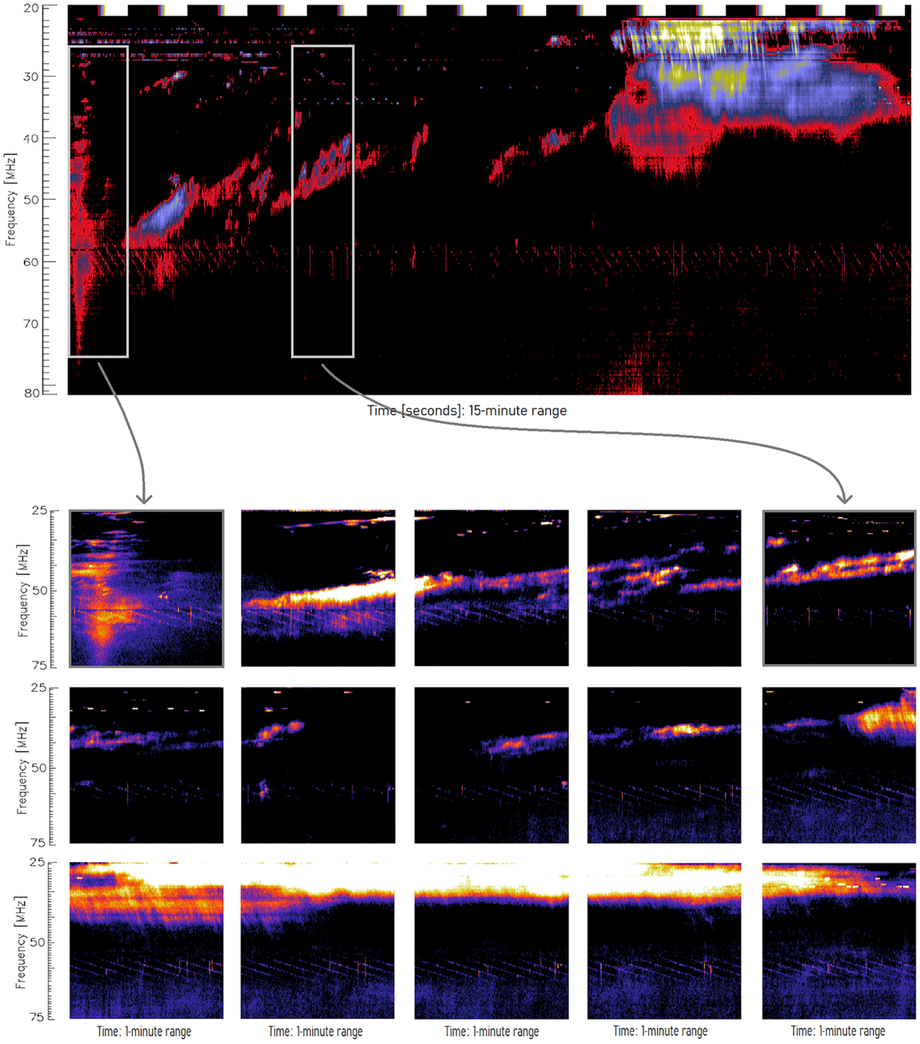


Figure 4 Cropping of the *e-Callisto* spectrogram shown in Figure 1 into 15 one-minute frames. The *first* and *fifth frames* are marked to guide the eye.

Other preprocessing schemes, which may help in the detection, classification, and analysis of solar radio bursts, are currently under investigation.

3.2. Training of the Deep Learning Classifier

Figure 3b summarizes our training scheme: first, a two-class training database is manually created with selected examples (tagged images) of what is to be considered as Burst (Yeas or Positive-class folder) and Not-Burst (Nays or Negative-class folder); this database, suitably broken down in training, validation, and test subsets, is presented to a convolutional neural network, which learns to classify the frames and generates a classification model; finally, before applying such a model to the analysis of untagged images (target database), we evaluate its performance and incorporate the lessons learned from incorrect predictions into improved versions of the model in a feedback loop, which is repeated until performance is considered satisfactory. Needless to say, throughout the article, no training data are ever used as target data.

This process is currently carried out in a DIGITS environment (NVIDIA Deep Learning GPU Training System, developer.nvidia.com/digits, version 6.1.1) with Caffe version 0.17.3 as machine learning framework. TensorFlow has also been tried but with poorer results. DIGITS is a user-friendly web-interface software specialized in training convolutional neural networks for computer vision problems. It allows the user to create the classes to be discriminated, optionally use pretrained CNNs, tune the CNN to a given problem, and test it. It can be installed in a Linux environment or, in its latest versions, in Docker containers.

3.2.1. Training Database Generation

Building an adequate training database is one of the keys to success (see Section 3.2.6 for emphasis on proper feedback being injected into the training database). As the well-known "garbage in, garbage out" saying goes, the selection, and eventually certain balance, of the images in each class folder is decisive.

With this in mind, we initially created single-observatory models, i.e. instrument-dependent models built from images taken in only one observatory to avoid distortions introduced by different local backgrounds or instrument-induced noise. The Yeas folder contains images of reported bursts, and the Nays folder contains frames outside reported burst intervals. The availability of both types of images is such that the Yea/Nay ratio is of order 10% (hundreds of Yeas, thousands of Nays), i.e. a clear though not aggressive imbalance often found in real-world problems. Several works in the literature (He and Garcia, 2009 and references therein) prove the ability of this kind of neural networks to work with unbalanced classes, but to be on the safe side, the results presented in this article have been thoroughly tested for overfitting (ruled out, see Figure 5), and a metric specifically adapted to unbalanced cases [G_{mean}] is used. Nevertheless, a few data augmentation schemes have been tried: those based on transformation of existing burst images do not work well in our nontime-reversible and translational-invariant scenario. Schemes based on magnetohydrodynamic models of known burst types are left for the future.

3.2.2. Model Generation

Three different neural networks available in DIGITS (GoogLeNet, LeNet, and AlexNet) were tested before settling on the latter as the best adapted to our problem and image resolution requirements. Different parameter configurations of AlexNet were tried out, which converged on the following configuration: 100 training epochs with validation interval of 10 epochs; NVCaffe as blob format; Stochastic Gradient Descent (SGD) solver type with 0.01 Base Learning Rate (parameter that determines the step size at each iteration); and a step-down policy for decay with 33% step size and 0.1 γ-factor (10% learning rate reduction at the 33% and 66% training stages). SGD has proved to perform better than others such as ADAM when working with AlexNet (Manataki, Vafidis, and Sarris, 2021).

As shown in Figure 5, accuracy and loss values improve rapidly during the first epochs and then reach a plateau, so it is definitely not necessary to go beyond 100 epochs – in fact, 30 may be enough (Table 3).

Except for specific robustness tests (Section 3.2.3), we use a Training/Validation/Test percentage breakdown of 75/25/0, i.e. we do not allocate any space for tagged testing as we

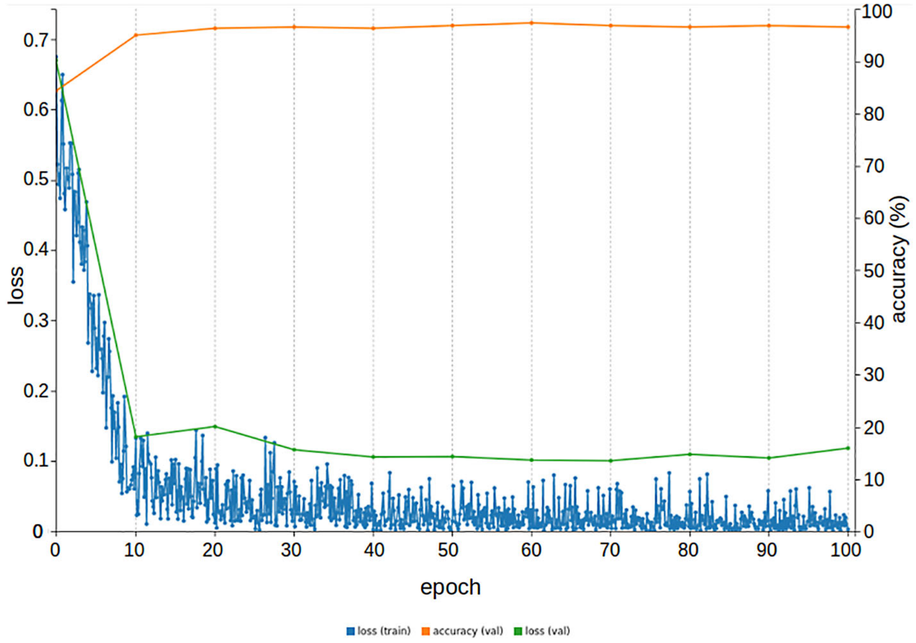


Figure 5 Validation accuracy (*orange*), validation loss (*green*), and train loss (*blue*) for an example model (Humain). At the validation stage, all our models reach accuracies above 98%, thus ruling out overfitting, with losses in the 4% to 12% range.

Table 2 Models used in this work. Abbreviations: A=ASSA, L=Landschlacht, H=Humain, G=Glasgow, B=Birr; HYB=hybrid. Indices 1, 2 (as in G1, G2) refer to subsequent versions of a model when feedback is incorporated. Training dataset: number of days worth of data [in brackets] used for a given month of the year 2021. The ratio Y/N of positive(Yeas)/negative(Nays) samples is shown in the last column.

Model	Instrument	Frequency [MHz]	Training dataset month [#days]	Burst [Yeas]	NotBurst [Nays]	Y/N [%]
G1	GLASGOW_59	45 – 81	May [2]	198	1031	16
G2	GLASGOW_59	45 – 81	May [29]	284	2317	11
A1	ASSA_02	25 – 80	May [10], June [1]	75	927	7
A2	ASSA_02	25 – 80	May [31]	308	2782	10
L1	Landschlacht_01	18 – 84	May [2]	318	3672	8
L2	Landschlacht_01	18 – 84	May [30]	334	3865	8
H	HUMAIN_59	45 – 435	May [4]	267	1176	19
HYB	A+L+G	18 – 84	May [31]	924	8964	9

prefer to do the testing directly on untagged frames. This breakdown has been found to work well empirically.

The output model, mathematically represented by the best set of adjusted convolutional weights, is automatically saved as a .caffemodel file (> 200 Mb). The model is now ready to be executed (Section 3.3). Table 2 shows the characteristics of the models built for this article, whereas Figure 5 shows an example of accuracy and loss curves. At the validation stage, all models reach accuracies above 98%, thus ruling out severe overfitting effects. The losses lie in the 4% to 12% range.

Once the single-observatory models were satisfactorily tested, we dared develop a hybrid model by combining training sets with images from three different observatories: ASSA, Landschlacht, and Glasgow (HYB: A+L+G in Table 2).

Model output predictions are given in terms of the probability P of the event being a burst. An event will be classified as Burst if $P \geqslant P_{\rm thr}$, where the *probability threshold* $P_{\rm thr}$ is by default set to $T_{50} = 50\%$ but can be optimized ($T_{\rm opt}$, Section 3.2.5).

Model performance can be evaluated using images either from the "Test" subsets or from "Documented" data, i.e. data for which an expert's event report exists. Either way, we can check whether the predictions made by the model are correct or not. In the two following subsections, both performance evaluations are presented.

3.2.3. Model Performance Evaluation Using "Test" Data

We have used the one-minute frames of our training dataset, already tagged as Real Positives or Negatives, to measure the robustness of our model (Positive and Negative "predictions" in machine learning jargon) against changes in the following parameters: distribution of images among the Training/Validation/Test subsets, random seeds, and number of training epochs.

A k-fold cross-validation procedure (Anguita et al., 2012) has been conducted for model selection and error estimation to obtain a reliable and rigorous estimation of the misclassification probability. A value of $k = 6$ has been used by dividing the training dataset into 6 subsets, which are then shuffled to form combinations each with 4/1/1 Training/Validation/Test subsets (i.e. 66%, 17%, 17%, respectively), thus resulting in 30 different models. Performance results are shown in Table 3 in terms of False Negative Rate (*FNR*, ratio of False Negatives to Real Positives), False Positive Rate (*FPR*, ratio of False Positives to Real Negatives), and their geometric mean:

$$G_{\rm mean} = \sqrt{(1 - FNR)\ (1 - FPR)}\ ,$$

as a measure of the combined goal of minimizing both the false positives and the false negatives: low false negative rates, ranging from 11% for the least-developed model to 7% for the most-developed one, are achieved while keeping the false positives at very low rates (2.5 – 0.9%). For unbalanced datasets like ours, the geometric mean of the recalls $(1 - FNR)$ and $(1 - FPR)$ (often called recall and specificity, respectively) is a more suitable, more conservative performance indicator than the recalls themselves or even than the balanced accuracy (arithmetic mean). The resulting $G_{\rm mean} = 93 – 96\%$ indicates that the system has, in spite of its unbalanced nature, high discriminating power.

The k-fold tests conducted show that stochastic effects derived from shuffling and the use of random seeds are kept under control, with standard deviations of 2.3 – 5.6% in *FNR*, less than 1% in *FPR*, and 1 – 3% in $G_{\rm mean}$. Moreover, the effect of random seeds alone has been measured, via ten repetitions of the hybrid (A+L+G) model with different seeds, to be less than 2% in *FNR* and less than 1% in *FPR* and $G_{\rm mean}$.

Finally, we compare the output of the hybrid model after 30 and 100 validation epochs using five repetitions in each case (bottom rows in Table 3): by Epoch 30, most of the learning has already been achieved. The toll paid to go from 95% to 96% in $G_{\rm mean}$ is to more than triple the computing time. In terms of energy consumption and savings, this is a key point to be considered.

Table 3 Robustness tests: shuffling effects (k-fold) for the least and most developed models (Glasgow and Hybrid, respectively); effect of random seeding; training epochs choice.

Test	Model	FNR [%]	FPR [%]	G_{mean} [%]
k-fold	Glasgow-G1	11.3 ± 5.6	2.5 ± 1.0	93.0 ± 3.1
	Hybrid	7.0 ± 2.3	0.9 ± 0.3	96.0 ± 1.2
random seeds	Hybrid	7.0 ± 1.4	0.7 ± 0.1	96.1 ± 0.8
100 epochs	Hybrid	7 ± 1	1 ± 1	96 ± 1
30 epochs	Hybrid	9 ± 2	1 ± 1	95 ± 1

In conclusion, our models are very efficient with tagged one-minute images ($G_{\mathrm{mean}} = 93-96\%$) and robust against stochastic effects induced by shuffling of the training, validation and test datasets, and by the use of random seeds.

3.2.4. Model Performance Evaluation Using "Documented" Data

The metrics chosen for a performance evaluation based on data for which an expert's event report exists are defined here:

The ground truth for what is or is not a solar radio burst is based on the event lists issued by the *e-Callisto* network, which provide beginning and end times for each detected burst (*burst interval*).

For *one-minute images* with frame times inside reported burst intervals, a positive model prediction will be considered a True Positive (TP), whereas a negative model prediction will be considered a False Negative (FN); conversely, for one-minute images with times outside reported burst intervals, a negative model prediction will be considered a True Negative (TN), whereas a positive prediction will be counted as a False Positive (FP).

At this point the reader is reminded that in this work the term prediction is used to denote the CNN classifier output and has no bearing with any forecasting capabilities.

From these basic metrics, computed on a minute-by-minute basis, we can define other metrics from different perspectives:

From a *burst interval* perspective, a burst interval is assigned a Burst True Positive (BTP) when at least one of the one-minute frames inside the interval (±1 minute) is a TP; otherwise, a Burst False Negative (BFN) is assigned. Demanding that all one-minute frames in a reported burst interval be positive would be completely unrealistic, especially on long intervals, often defined during visual inspection by loosely grouping together several solar events.

Regarding 15-minute *data runs*, all runs involved in a BTP are considered True Positive Runs [TP_{run}], whereas a run will be considered a False Positive Run [FP_{run}] if there are no bursts in that 15-minute period and at least one of the one-minute frames in the run was classified as FP.

Now we define the metrics to be used in the evaluation of a model's performance:

the *False Negative Burst Score* [FN_{bs}], which accounts for the fraction of reported *bursts* not detected by the model,

$$FN_{\mathrm{bs}} = \frac{BFN}{BTP + BFN} \,, \tag{1}$$

and the *False Positive Run Score [FP_{rs}]* or fraction of false positive *runs*,

$$FP_{\mathrm{rs}} = \frac{FP_{\mathrm{run}}}{R}\,, \tag{2}$$

where R is the total number of runs analyzed.

The user is spared browsing a large amount of runs and is, instead, given a reduced fraction, $(FP_{\mathrm{run}} + TP_{\mathrm{run}})/R$, with the first term [$FP_{\mathrm{rs}}$] representing the wrongly selected part.

If the fraction of bursts missed by the model [FN_{bs}] is low enough and the number of false alarms has been sufficiently reduced (low FP_{rs}), then the model is ready to be applied (Burst Detection, Section 3.3); otherwise, the model needs further improvement (Feedback to Model Generator, Section 3.2.6).

3.2.5. Threshold Optimization

As mentioned in Section 3.2.2, probability threshold values P_{thr} other than 50% (called T_{50}) may be used in the Positive/Negative output decision – the higher the threshold, the lower the total amount of Positives (green TP + FP line in Figure 6). This figure illustrates the search for the threshold value T_{opt} that optimizes our performance scores using an array of thresholds T in steps of 2.5%. As optimization parameter, we choose the distance d to the ideal ($FP_{\mathrm{rs}} = 0$, $FN_{\mathrm{bs}} = 0$) case:

$$d = \sqrt{{FP_{\mathrm{rs}}}^2 + {FN_{\mathrm{bs}}}^2},$$

which is shown as a black line. In the ASSA case, for instance, minimum d is found at $T_{\mathrm{opt}} = 52.5\%$, but the whole range $40 \leq T_{\mathrm{opt}} \leq 57.5$ is within 1% of d_{min}. This is why, in Table 6, ±1% ranges are added in brackets. In contrast, for the hybrid model applied on Birr data, the d function grows steadily on both sides of $T_{\mathrm{opt}} = 57.5\%$. The separation between the green (TP + FP) and the blue (FP) lines is a measure of the number of True Positives, which is higher in the ASSA observatory (left plot), consistent with Figure 2.

However, other optimization schemes or parameters might be envisaged depending on the user's goals: some users may prefer to detect only the most intense events with a minimum of false alarms; others may focus on detecting as many bursts as possible at the expense of having many FPs. For instance, we might want to make sure that the number of missed bursts does not exceed 15%: a horizontal dotted line at 15% has been added to the plot to guide the eye, and the highest threshold fulfilling this requirement, T_{fn15}, has been appended to Table 6.

For easy model intercomparison, ROC-like curves (Receiver Operating Characteristics) in FP_{rs}–FN_{bs} space are shown in Figure 7. For each model, the probability thresholds P_{thr} within the $30-90\%$ range are applied (red dots from right to left, with circle and asterisk marks for the 50% and optimal thresholds, respectively). The location and extent of the optimal range of thresholds is represented by the dots closest to the dashed arc of radius d_{min}, e.g. the ten rightmost dots in the case of the single ASSA model, which correspond to the $40-62.5\%$ range around $T_{\mathrm{opt}} = 52.5\%$. Examples of significant advantage in using T_{opt} over the default $T_{50} = 50\%$ are the single Glasgow model or the hybrid model applied to either ASSA or Birr target data – notice the large distance between red circle and dashed line.

A summary of threshold optimization results can be found in Table 6 (right half), the detailed discussion of which is deferred to Section 4.2.

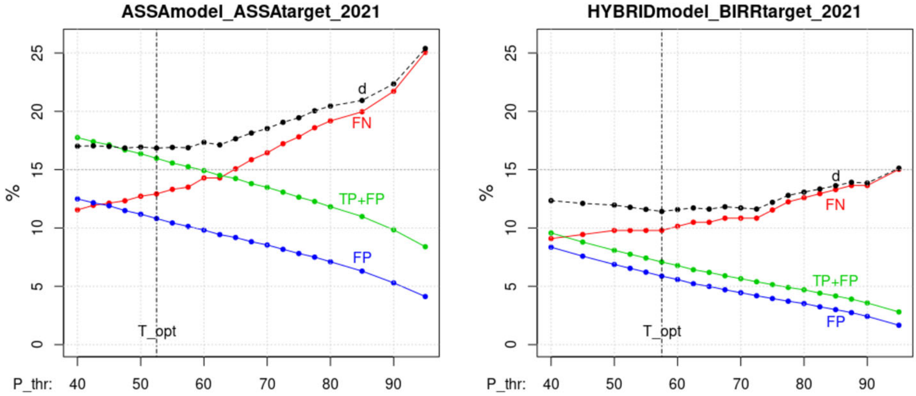


Figure 6 Threshold optimization examples for a single and a hybrid model: scores of False Negatives [FN], Positives [sum of true and false: $TP + FP$], and False Positives [FP] as a function of probability threshold [P_{thr}]. T_{opt} is the optimal threshold that minimizes the distance [d] to the ideal case with neither false positives nor false negatives.

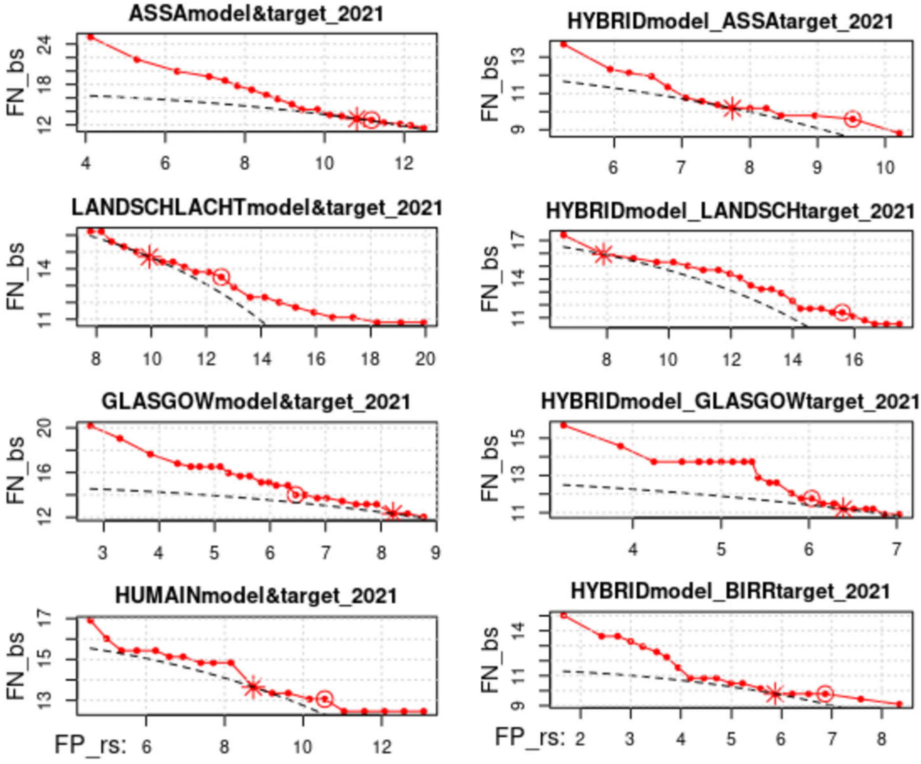


Figure 7 ROC-like curves in FP_{rs}–FN_{bs} space. For each model, probability thresholds P_{thr} within the 30 – 90% range are applied (*red dots* from right to left, with *circle* and *asterisk* marks for the 50% and optimal thresholds, respectively). A *dashed arc* of radius d_{min} is shown.

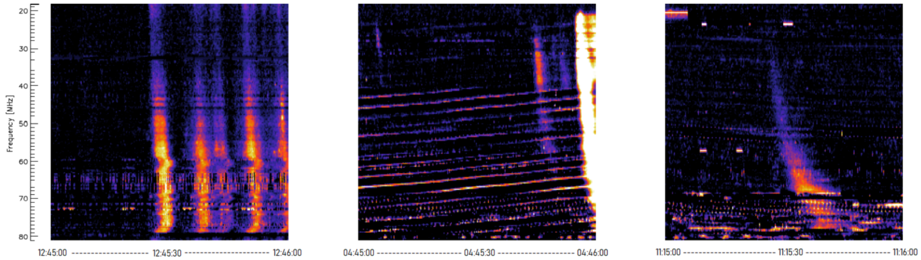


Figure 8 Discovery of unreported bursts [observatory, date, UT time]: GLASGOW 22 May 2021 12:45:00 UT (*left*); Landschlacht 23 May 2021 04:45:00 UT (*center*); Landschlacht 31 May 2021 11:15:00 UT (*right*).

3.2.6. Feedback to Model Generator

Feedback to the model generator may be applied by modifying either the CNN parameters or the training database (Figure 3b).

In the first case, parameters such as the number of training epochs have been optimized; the use of pretrained networks shortens the model-generation process but does not improve its efficiency. Another possibility is to try more efficient networks, which become readily available in this fast-evolving discipline.

In the second case, important lessons may be learned from careful study of common CNN misses: a close look at the false positives reveals families of pathological cases, which can be avoided by adding examples of such frames to the Nays folder; revision of false positives brings us the discovery of a good number of unreported bursts, which had gone unnoticed to the expert inspector. These could be called False False-Positives, of which Figure 8 shows a few examples.

A dynamic model generation scheme whereby a continuous flow of both burst and not-burst examples is added to the training database as new data come in can be implemented. This will be the subject of future work.

3.3. Burst Detection

Once an optimum classifier has been built, the goal is to apply it to a target database – data that have never gone through the CNN system before. As any other data, they must first be downloaded from the *e-Callisto* web server, preprocessed, and transferred to the CNN environment (DIGITS). Figure 9 summarizes the steps in the burst detection process.

Running the classification model yields a “prediction table” on a minute-by-minute basis, i.e. the probabilities $P(Y)$ and $P(N)$ that each one-minute frame belongs to the positive or negative class, respectively. As we mentioned earlier, this is the basic table from which other more elaborate products may be generated.

3.3.1. Performance Re-Evaluation (Documented Period)

In the case of target data that have been previously inspected by an expert, the ground truth is known, and therefore the classifying model may be re-evaluated. We use, as inputs for this process, the prediction table, the corresponding *e-Callisto* event report, and, for added confidence, the NOAA-SWPC reports.

The output is given in terms of the same metrics used during the original performance evaluation (Section 3.2.4): false negative burst score (FN_{bs}) and false positive run score

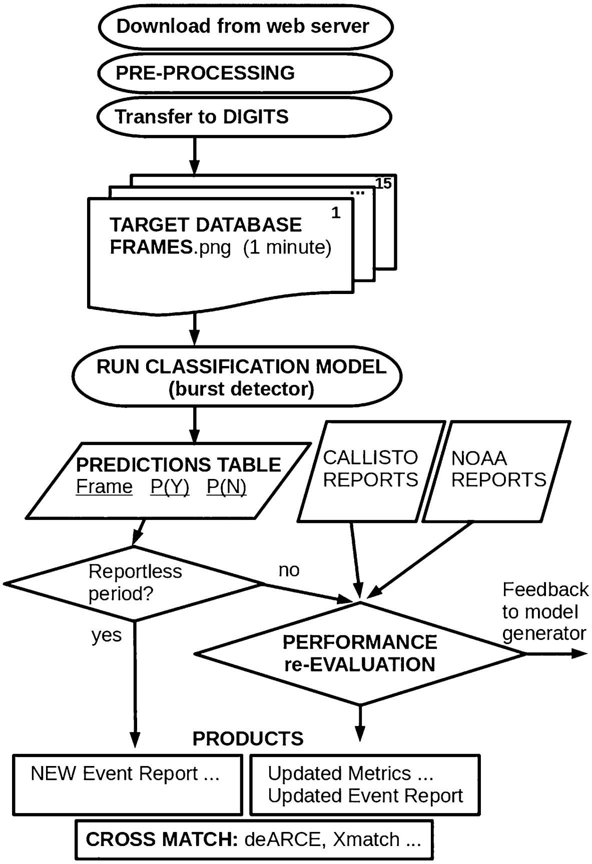


Figure 9 Application of the classification model to the target database for burst detection.

(FP_{rs}). Again, visual inspection of common misses – both false positives and false negatives – powers the feedback channel for further model improvement and, by the way, uncovers some unreported bursts.

3.3.2. Products for Reportless Periods

If, instead, the target data belong to a reportless period, no ground truth is available for performance evaluation, and the main product is the generation of previously unavailable (new) reports for the space-weather and scientific community. An example of these new reports, featuring cross-match among *e-Callisto* stations and/or with NOAA stations, is shown in Figure 11.

Cross-match products, regardless of the documented or reportless nature of the period, will be the subject of Section 4.5.

4. Results and Discussion

Having described in detail the method for automatic burst identification, we now summarize the most relevant results for the selected models.

Table 4 *Left*: confusion matrix for Glasgow data (11 days, November 2021). Pg, Ng: ground-truth Positives and Negatives; Pn, Nn: neural-net predictions. *Right*: derived scores. R: total number of runs.

	Pn	Nn	Score	Value
Pg	TP = 20	FN = 5	FN_{bs}	5/25 = 20%
Ng	FP = 63	TN = 429	FP_{rs}	63/517 = 12%
sum	83	434	R	517

Table 5 Initial performance evaluation of four models on target data from November 2021, to be compared with the performance re-evaluation shown in Table 6.

Model	Target	# days	# runs	$\%FN_{\text{bs}}$	$\%FP_{\text{rs}}$	$\%(FP+TP)_{\text{rs}}$
ASSA	ASSA	30	1303	30	11	16
Landschlacht	Landschlacht	30	1179	38	20	20
Glasgow	Glasgow	11	517	20	12	16
Humain	Humain	30	1380	30	7	8

4.1. Initial Performance Evaluations

As mentioned in Section 3.2.4, models are evaluated in terms of their ability to detect as many bursts as possible (low false negative score, FN_{bs}) without incurring in many false alarms (low FP_{rs}).

We started by building models of the best performing stations. Our first successful model was trained with data taken during two very active days in May 2021 at the Glasgow observatory, known to be a very reliable instrument at that time, and tested on eleven days worth of target data from November 2021. The resulting confusion matrix is shown in Table 4 along with the final scores $FN_{\text{bs}} = 20\%$ and $FP_{\text{rs}} = 12\%$, which were considered satisfactory for a first training round. The total percentage 16% of Positive runs, including True Positives, means that this model (see characteristics in Table 2) filters out 84% of all runs and that in the remaining set, we can find 80% of the bursts ($FN_{\text{bs}} = 20\%$).

Close inspection of the five missed bursts (false negatives) reveals that two of them are extremely weak – one is not observed in any other observatory around the world – and another two belong to long storms that indeed have been detected in adjacent runs, so in fact only one burst, i.e. 4% of the total sample, is a worrying miss. This has turned out to be the general trend in most other models.

A comparison of results of this initial performance evaluation for single-observatory models is given in Table 5. At this stage, the default probability threshold T_{50} is used.

4.2. Performance Re-Evaluations of Classifiers

We now turn to the results obtained for a full year's worth of data (2021) after feedback from inspection of false positives and false negatives has been incorporated to model training. A very significant reduction of the FN Burst Score is found for all single-observatory models (top half of Table 6) using the same T_{50} threshold as in Table 5: a remarkable 58% drop for ASSA and 30% for Glasgow, the FP Run Score not dropping for the former and falling 46% for the latter. Overall, the single model for Glasgow shows the best performance ($FN_{\text{bs}} = 14$, $FP_{\text{rs}} = 6.5$, resulting in $d = 15.4$) able to detect 86% of the reported bursts with false alarms in just 6.5% of the runs. However, threshold optimization offers room for further improvement. In this case, it brings the ROC d parameter down to a minimum of $d = 14.8$

Table 6 Summary of results for single and hybrid models (Mod) applied on a full year worth of data: 2021 (Target, number of Runs). Scores FN_{bs} and FP_{rs} are given for probability thresholds at 50% (T_{50}) and at T_{opt} along with the resulting distance d or d_{min}, respectively. In the seventh column, we show the highest threshold for which FN_{bs} is kept below 15%.

Mod	Target	Runs	FN_{bs} (T_{50})	FP_{rs} (T_{50})	d (T_{50})	T_{fn15}	T_{opt} (±1% range)	d_{min} (T_{opt})	FN_{bs} (T_{opt})	FP_{rs} (T_{opt})
A	ASSA	13,172	12.7	11.2	16.9	62.5	52.5 (40.0 – 57.5)	16.8	12.9	10.8
L	Lands.	16,213	13.5	12.6	18.4	65.0	65.0 (62.5 – 72.5)	17.7	14.7	9.9
G	Glasg.	13,917	14.0	6.5	15.4	55.0	30.0 (25.0 – 30.0)	14.8	12.3	8.2
H	Humain	16,607	13.1	10.6	16.8	67.5	60.0 (57.5 – 60.0)	16.2	13.7	8.7
HYB	ASSA	13,172	9.6	9.5	13.5	90.0	67.5 (65.0 – 75.0)	12.8	10.2	7.7
HYB	Lands.	16,213	11.4	15.6	19.3	82.5	95.0 (95.0)	17.8	15.9	7.9
HYB	Glasg.	13,917	11.8	6.0	13.2	90.0	42.5 (30.0 – 47.5)	12.9	11.2	6.4
HYB	Birr	20,868	9.8	6.9	12.0	90.0	57.5 (57.5)	11.4	9.8	5.9

at $T_{\mathrm{opt}} = 30\%$: a decrease of 1.7% in false negatives and an increase of 1.7% in false positives yield a more compensated model ($FN_{\mathrm{bs}} = 12.3$, $FP_{\mathrm{rs}} = 8.2$), which detects nearly 88% of bursts with false alarms in only 8% of the runs.

Quite unexpectedly, the best results come from the application of the Hybrid model (Table 6, bottom half). For starters, the FN Burst Score remains below 15% for all thresholds up to at least 90% (except 65% for Landschlacht, see the T_{fn15} column). For the default T_{50} threshold, the Hybrid model performs better on all target datasets, and the same is true for the optimized T_{opt}. Compared to single-observatory results, 21% and 9% reductions of FN_{bs} are observed in ASSA and Glasgow, accompanied by 29% and 22% reductions in FP_{rs}, respectively. This remarkable improvement includes Birr data, which never took part in the training process – its results are even better than those obtained for other observatories, thus proving that a properly built hybrid model can be applied to many different observatories.

4.3. Products Generated for Documented Periods

In Figure 10, we can see an extract of an automatically generated report for an already documented *e-Callisto* period using predictions made by the Landschlacht model. The first two columns show the date and time interval of every detected burst – positive predictions tend to be very scattered. Besides, cross-validation with both NOAA's SWPC and *e-Callisto* burst reports is provided in the third and fourth columns, respectively: "Yes" is printed if there is at least a one-minute overlap between the Time column and the SWPC or *e-Callisto* report intervals; otherwise, a string of hyphens is printed. This information about coincident events in several stations belonging to two different observing networks is precious as it allows the scientific community to focus on the most probably interesting events.

A somewhat unexpected added value to our method is the discovery of *unreported bursts* (Section 3.2.6 and Figure 8). As an example, in 2021, 47 of them have been found in the ASSA database and 49 in Humain (i.e. 14% of the total number of bursts). They have all been visually double-checked by an expert team, and the corresponding official reports have been updated. In this article, they are given as a supplementary material:

e-CALLISTO_2021_UnreportedBurstsASSA.txt (47 bursts),

e-CALLISTO_2021_UnreportedBurstsHumain.txt (49 bursts).

```
# Landschlacht_2021_11

#Date     Time         In_NOAA                                   In_Callisto
#-----------------------------------------------------------------------------------------------------
...
20211108  10:26-10:26  ------                                    ------
20211108  11:49-11:49  Yes[11:49-11:49]                          Yes[11:49-11:49]
20211108  13:06-13:06  Yes[13:01-14:15]                          ------
20211108  13:08-13:08  Yes[13:01-14:15]                          Yes[13:08-13:08]
20211108  13:24-13:24  Yes[13:01-14:15,13:12-13:54]              Yes[13:15-14:04]
20211108  13:41-13:42  Yes[13:01-14:15,13:12-13:54]              Yes[13:15-14:04]
20211108  14:15-14:15  Yes[13:01-14:15]                          ------
20211108  14:26-14:26  ------                                    ------
20211109  07:43-07:43  ------                                    ------
20211109  12:32-12:32  ------                                    ------
20211109  15:17-15:17  Yes[13:07-21:14]                          ------
20211110  13:09-13:09  Yes[06:08-15:26,11:58-20:32,12:18-14:15]  ------
20211110  13:12-13:12  Yes[06:08-15:26,11:58-20:32,12:18-14:15]  ------
20211110  13:54-13:54  Yes[06:08-15:26,11:58-20:32,12:18-14:15]  ------
20211111  07:56-07:56  Yes[00:00-10:29,07:10-08:32]              Yes[03:15-10:20,06:14-09:45]
20211111  09:27-09:27  Yes[00:00-10:29]                          Yes[03:15-10:20,06:14-09:45,09:16-13:20]
20211111  10:02-10:02  Yes[00:00-10:29]                          Yes[03:15-10:20,09:16-13:20]
20211111  10:43-10:43  ------                                    Yes[09:16-13:20]
20211111  11:08-11:08  ------                                    Yes[09:16-13:20]
20211111  11:30-11:30  ------                                    Yes[09:16-13:20]
20211111  11:36-11:36  ------                                    Yes[09:16-13:20]
20211111  11:45-11:45  ------                                    Yes[09:16-13:20]
20211111  11:48-11:48  ------                                    Yes[09:16-13:20]
20211111  11:58-11:58  ------                                    Yes[09:16-13:20]
20211111  12:55-12:55  Yes[12:22-18:54]                          Yes[09:16-13:20]
20211111  13:39-13:39  Yes[13:36-13:39,12:22-18:54]              Yes[13:39-13:39]
20211111  14:24-14:24  Yes[12:22-18:54]                          ------
20211111  15:31-15:31  Yes[12:22-18:54]                          ------
20211112  09:04-09:04  Yes[07:19-10:41]                          ------
20211112  10:03-10:03  Yes[07:19-10:41]                          ------
20211112  13:57-13:57  Yes[13:49-16:18]                          ------
20211112  15:01-15:01  Yes[13:49-16:18]                          ------
20211113  07:56-07:56  ------                                    ------
...
```

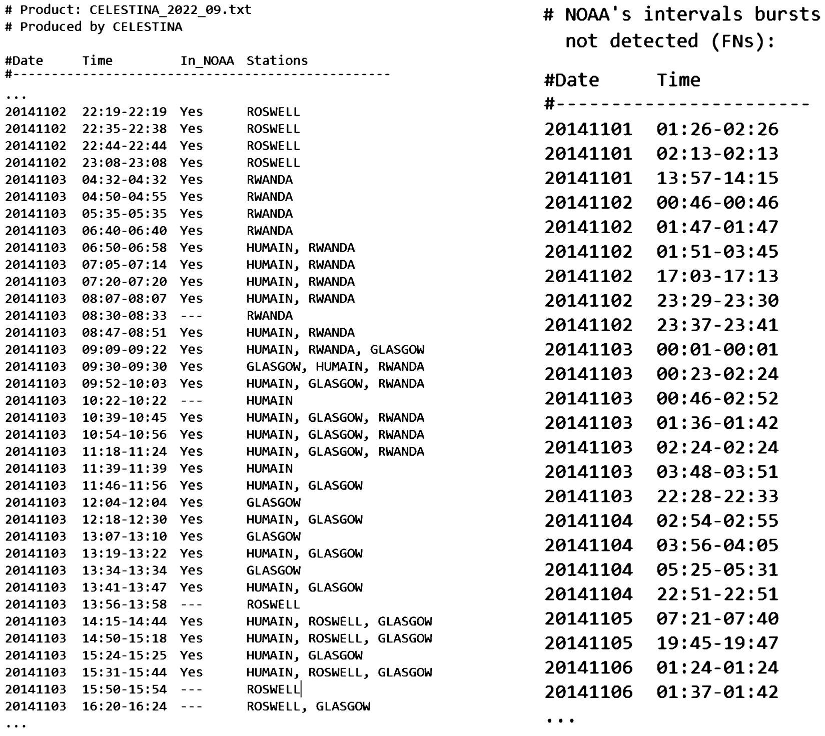
Figure 10 Extract of an automatically generated report for an already documented period (2021). Valuable information about cross-validation with NOAA-SWPC and other *e-Callisto* stations has been added.

Besides, the automatic identification output has led to the discovery and correction of typos in the official reports – the reader must bear in mind how tiring the daily visual inspection can be. We have also found cases of doubtful FPs, where the expert team cannot tell whether they correspond to real bursts or not.

In an ideal deep learning experiment, with perfect ground truth, there would not be any doubtful FPs or Unreported Bursts, but our real scenario is not perfect. Unlike the evaluation made on tagged Test data (Section 3.2.3), evaluation of model performance on real data will always be subject to unavoidable errors in the ground truth definition. This must be taken into account when comparing results with methods in other scenarios. Our conservative approach and the findings made during FN and FP revisions (very weak FNs, doubtful FPs, unreported bursts) suggest that the automatic method presented here performs even better than what the face-value figures show.

4.4. Products Generated for Reportless Periods

As mentioned in Section 3.3.2, we aim at producing reports for the years 2012 – 2019, for which data observations from *e-Callisto* stations are available but reports are not. Figure 11a is an extract of a report produced using predictions made by the Hybrid model on target data from five different stations in November 2014 – a reportless period for *e-Callisto* but not for NOAA's SWPC. It is similar to Figure 10, but in this case, burst intervals several minutes long are noticeably much more common. The reason is that in these reports, a



```
# Product: CELESTINA_2022_09.txt
# Produced by CELESTINA

#Date     Time        In_NOAA  Stations
#---------------------------------------------------
...
20141102  22:19-22:19  Yes     ROSWELL
20141102  22:35-22:38  Yes     ROSWELL
20141102  22:44-22:44  Yes     ROSWELL
20141102  23:08-23:08  Yes     ROSWELL
20141103  04:32-04:32  Yes     RWANDA
20141103  04:50-04:55  Yes     RWANDA
20141103  05:35-05:35  Yes     RWANDA
20141103  06:40-06:40  Yes     RWANDA
20141103  06:50-06:58  Yes     HUMAIN, RWANDA
20141103  07:05-07:14  Yes     HUMAIN, RWANDA
20141103  07:20-07:20  Yes     HUMAIN, RWANDA
20141103  08:07-08:07  Yes     HUMAIN, RWANDA
20141103  08:30-08:33  ---     RWANDA
20141103  08:47-08:51  Yes     HUMAIN, RWANDA
20141103  09:09-09:22  Yes     HUMAIN, RWANDA, GLASGOW
20141103  09:30-09:30  Yes     GLASGOW, HUMAIN, RWANDA
20141103  09:52-10:03  Yes     HUMAIN, GLASGOW, RWANDA
20141103  10:22-10:22  ---     HUMAIN
20141103  10:39-10:45  Yes     HUMAIN, GLASGOW, RWANDA
20141103  10:54-10:56  Yes     HUMAIN, GLASGOW, RWANDA
20141103  11:18-11:24  Yes     HUMAIN, GLASGOW, RWANDA
20141103  11:39-11:39  Yes     HUMAIN
20141103  11:46-11:56  Yes     HUMAIN, GLASGOW
20141103  12:04-12:04  Yes     GLASGOW
20141103  12:18-12:30  Yes     HUMAIN, GLASGOW
20141103  13:07-13:10  Yes     GLASGOW
20141103  13:19-13:22  Yes     HUMAIN, GLASGOW
20141103  13:34-13:34  Yes     GLASGOW
20141103  13:41-13:47  Yes     HUMAIN, GLASGOW
20141103  13:56-13:58  ---     ROSWELL
20141103  14:15-14:44  Yes     HUMAIN, ROSWELL, GLASGOW
20141103  14:50-15:18  Yes     HUMAIN, ROSWELL, GLASGOW
20141103  15:24-15:25  Yes     HUMAIN, GLASGOW
20141103  15:31-15:44  Yes     HUMAIN, ROSWELL, GLASGOW
20141103  15:50-15:54  ---     ROSWELL
20141103  16:20-16:24  ---     ROSWELL, GLASGOW
...
```

```
# NOAA's intervals bursts
  not detected (FNs):

#Date     Time
#-----------------------
20141101  01:26-02:26
20141101  02:13-02:13
20141101  13:57-14:15
20141102  00:46-00:46
20141102  01:47-01:47
20141102  01:51-03:45
20141102  17:03-17:13
20141102  23:29-23:30
20141102  23:37-23:41
20141103  00:01-00:01
20141103  00:23-02:24
20141103  00:46-02:52
20141103  01:36-01:42
20141103  02:24-02:24
20141103  03:48-03:51
20141103  22:28-22:33
20141104  02:54-02:55
20141104  03:56-04:05
20141104  05:25-05:31
20141104  22:51-22:51
20141105  07:21-07:40
20141105  19:45-19:47
20141106  01:24-01:24
20141106  01:37-01:42
...
```

Figure 11 (**a**) Extract of burst report generated for a period without reports (Nov. 2014). (**b**) Appendix with list of NOAA-SWPC bursts not detected by the model in any of the observatories used for the report.

coincident event is considered if the time difference is equal to or less than five minutes so as to account for loose grouping of solar bursts in some *e-Callisto* reports. Again, cross-validation information with NOAA's SWPC burst reports is provided in the third column, and "Yes" is written using the same criteria – intervals are omitted for aesthetic reasons, but they could also be included. In the last column, the names of the stations with at least one positive prediction belonging to the interval are written. Figure 11b is an appendix with the list of NOAA's SWPC burst intervals not detected by the model in any of the observatories used for the report – hence it could be considered a list of false negatives. The list is quite populous; nonetheless, most of the intervals correspond to times where the stations used – with the possible exception of Roswell – are in the night side.

4.5. Cross-Match Products: deARCE Xmatch

Cross-match between different observatories is a key tool to discriminate actual bursts from radio frequency interference (RFI), as RFI should be station-dependent (Prieto et al., 2020), whereas a solar burst may be observed by several observatories located at similar geographical longitudes.

For this purpose, we have developed a cross-match system named Xmatch. In Figure 12, it is applied to the five most active observatories at European longitudes during September 2021. It shows the number of coincidental positive predictions, or alarm level, on a minute basis (red circles, just one day). Black crosses indicate alarm levels according to reported

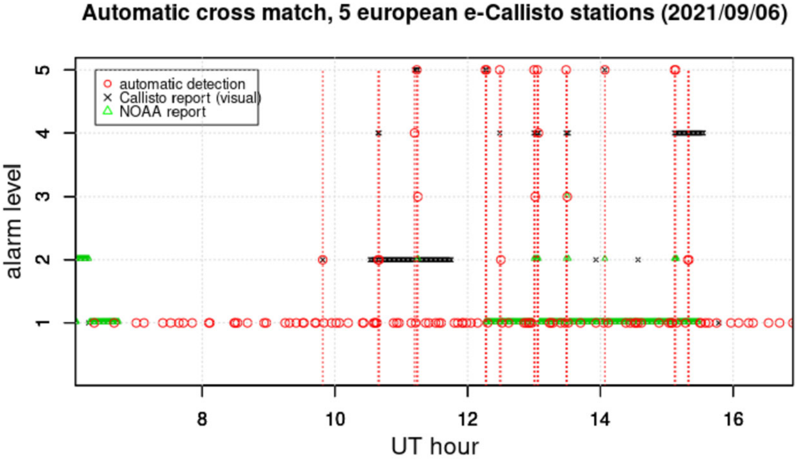


Figure 12 Example of cross-match among five European stations in automatic identification mode (Graz, Landschlacht, Heiterswil, Glasgow, Birr). Alarm-level definitions: for the *e-Callisto* report (visual inspection, *black crosses*), levels 1 – 5 correspond to having detected the burst respectively in 1, 2 – 3, 4 – 5, 6 – 9, 10 – 25 different stations; for the automatic system (*red circles*) and NOAA (*green triangles*), the alarm level is exactly equal to the number of positive stations.

visual inspection, with levels 1 to 5 corresponding, respectively, to detections in 1, 2 – 3, 4 – 5, 6 – 9, or ≥ 10 stations across the whole *e-Callisto* network. Green triangles indicate the number of NOAA stations reporting positive detection. Several "one-hit" events not correlated to reported bursts (i.e. red circles at Level 1 without green or black marks) can be found (around 08:00 or 16:00 UT) representing either unreported bursts or false positives – not so surprising in this test, where three stations (Graz, Heiterswil, and Birr) have not yet been included in any CNN training. It is remarkably clear that multihit events (red circles with dashed lines) correlate almost perfectly with reported bursts – the higher the coincidence number, the higher the confidence level. This service will be of great use to scientists carrying out cross-match studies between ground-based and satellite-borne detectors of the type shown by Gómez-Herrero et al. (2021).

For this reason, in October 2022, we have started issuing a daily plot, called Xmatch (celestina.web.uah.es/Xmatch) where all positive predictions from the previous day in (currently ten of) the most active observatories are represented and coincidental positives are easily visible (Figure 13).

Finally, we use the September 2021 test mentioned above (Figure 12, hybrid model) to quantify the performance of the cross-match tool by demanding coincident positives in a minimum number of observatories M.

We use the same metrics as in Section 3.2.4, adapted to the multistation scenario. Regarding FN_{bs}, a reported burst interval is considered as BTP if at least M observatories are positive in both prediction and report or if less than M observatories appear in the report but all of them are positive in the prediction as well; otherwise, it is counted as BFN. Regarding FP_{rs}, a run is considered an FP_{run} if it does not belong to any burst interval and there are at least M positive model predictions in at least one minute of the run; besides, R is defined as the total number of runs during which M or more observatories are active.

The results are shown in Table 7: demanding coincident positive predictions in at least $M = 2$ stations brings the False Positive Score below 4% – either 2.4% or 3.4% depending

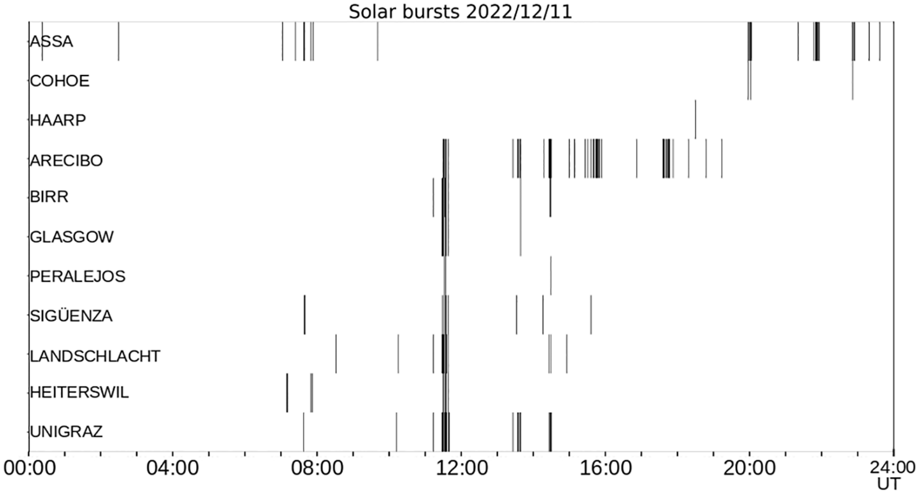


Figure 13 Example of cross-match plot posted daily at celestina.web.uah.es/Xmatch.

Table 7 Cross-match scores obtained for two different probability thresholds (T_{50}, T_{25}) when a minimum of M coincident positives is demanded. Model used: Hybrid. Target data: five European stations, September 2021. R: number of runs. $T_{\mathrm{fn}15}$: highest probability threshold with $FN_{\mathrm{bs}} < 15\%$.

M	R	$\%FN_{\mathrm{bs}}$ (T_{50})	$\%FP_{\mathrm{rs}}$ (T_{50})	$\%FN_{\mathrm{bs}}$ (T_{25})	$\%FP_{\mathrm{rs}}$ (T_{25})	$T_{\mathrm{fn}15}$
2	1785	11.9	2.4	8.9	3.4	60.0
3	1603	18.8	0.3	15.8	0.6	5.0

on the trade-off with False Negative Scores of 11.9% or 8.9%, respectively, and we are safely below $FN_{\mathrm{bs}} = 15\%$ for any threshold $P_{\mathrm{thr}} \leq 60$.

Increasing the minimum-coincidence demand to $M = 3$ (i.e. more than 60% of the observatories in this test) produces a virtually False-Positive-free sample (only 5 and 9 false alarms in 1603 runs, which a posteriori turned out to be Unreported Bursts) at the expense of raising the FN score above 15% (though still below 20%), which is not surprising for such a strong demand. FN_{bs} levels similar to those found for the $M = 2$, $P_{\mathrm{thr}} = 50\%$ case may be reached by setting a suitable combination of more stringent coincidence demand $M = 3$ and a more relaxed positive threshold $P_{\mathrm{thr}} = 1.55\%$. Both scores improve: $FN_{\mathrm{bs}} = 10.9\%$ and $FP_{\mathrm{rs}} = 1.7\%$.

5. Conclusions and Future Directions

In this article, we have presented an automatic radio-burst detection method (deARCE) based on artificial-intelligence techniques. A set of state-of-the-art deep neural networks have been assessed to select the one that best suits our problem. Through careful training with a wide range of observatories and vast amounts of data covering many years of operation of the *e-Callisto* array, a high-performance CNN configuration has been reached, which is stable, robust, and ready for use with past, present, and future data. Its high performance has been validated through internal cross-match among *e-Callisto* stations hundreds of kilometers apart and through external cross match with NOAA data.

Comparison with other results in automatic detection of SRBs or in ideal deep learning scenarios must be made with great care (see the end of Section 4.3). Afandi et al. (2020) use only 1491 spectrograms, all from a single day and in a perfect-tag scenario (perfect report, no chance of finding unreported bursts). Lobzin et al. (2010), Singh et al. (2019), and others do cover large periods of time but only in one observatory (Gauribidanur, Learmonth, etc.). In this article, False Negative and False Positive scores similar to those found in statistical method articles are obtained, but we cover full years worth of data simultaneously in many different observatories. The development of hybrid models, valid for a number of data sources, tunable probability threshold optimization and coverage of different frequency ranges (both the hundreds of MHz and the 20 – 100 MHz regimes are included in our study) are key advantages of our CNN-based method.

Perhaps our main contribution is a thorough analysis of the specific problem of automatic burst detection in solar radio spectra. The proposed solution, based on the AlexNet neural network within the DIGITS environment, is portable, flexible in terms of incorporation of new observatories or changes in their characteristics and has been designed to accept data from ground observatories other than *e-Callisto* and even space-borne observatories.

Future work should include: i) development of more sophisticated spectrogram denoising and preprocessing algorithms; ii) exploration of data augmentation techniques such as the generation of new images by means of probabilistic variations; iii) dynamic training system for constant improvement of the training dataset with new incoming data; iv) generation of hybrid models using ensemble techniques to combine more efficiently images from observatories with different weights; v) application of a genetic fuzzy algorithm during the cross-match stage for automatic compensation of the relative performance of each observatory.

The resulting products, beyond the automatic generation of daily reports for the 2012 – 2019 reportless gap and for the future, include a cross-match detection and plotting tool (Xmatch) freely available to the scientific community. Their key features are the simultaneous analysis of several data sources and the definition of a detection threshold. We believe that our tools will be very useful to the community, as they allow the end scientist to check at a glance if a solar radio burst has taken place with a high level of confidence. deARCE aims to be the first radio spectra analysis tool for space weather, which eventually may include not only burst detection and nowcasting, but also burst type classification and forecasting.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11207-023-02171-0.

Acknowledgments We thank the FachHochschule NordWestschweiz for hosting the *e-Callisto* Data Center and our equipment maintenance staff, Abel Moreno (Peralejos de las Truchas) and Patricia Rodrigo (Sigüenza), for their time and dedication. *In memoriam* Éric Nuss, resilient physicist and wise friend.

Author contributions J. Bussons Gordo and M. Fernández Ruiz devised the method, performed the analysis and drafted the manuscript with input from all authors. M. Prieto Mateo contributed to the design and implementation of the research, the analysis of results and writing of the manuscript. J. Alvarado Díaz, F. Chávez de la O, J.I. Hidalgo advised on the computational framework, worked out the technical details, and interpreted the deep-learning results. Ch. Monstein routinely performed expert visual inspection of the data and supervised the method and findings of this work.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

This research work has been funded by the Junta de Comunidades de Castilla–La Mancha and the European Union through the CELESTINA project (Castillian E-Callisto Leading Experimentation in Solar-Terrestrial Interaction with Novel Antennas, code SBPLY/19/180501/000237) in collaboration with Parque Científico y Tecnológico de Castilla–La Mancha and the universities of Alcalá and Murcia.

J. Bussons is supported by Fundación Séneca, Región de Murcia (20949/PI/18). J. Alvarado, F. Chávez, and J.I. Hidalgo are supported by PID2021-125549OB-I00 Ministerio de Ciencia e Innovación Spain and Y2018/NMT-4668 (Micro-Stress-MAP-CM) Comunidad de Madrid – FEDER.

Data Availability The datasets analyzed and generated during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests The authors declare no competing interests.

References

Afandi, N.Z.M., Sabri, N.H., Umar, R., Monstein, C.: 2020, Burst-finder: burst recognition for E-CALLISTO spectra. *Indian J. Phys.* **94**, 947. DOI.

Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., Ridella, S.: 2012, The 'K' in K-fold cross validation. In: *20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 441. i6doc.com/en/livre/?GCOI=28001100967420. Accessed on 21 December 2022.

Benz, A.O., Monstein, C., Meyer, H., Manoharan, P.K., Ramesh, R., Altyntsev, A., Lara, A., Paez, J., Cho, K.-S.: 2009, A World-Wide Net of Solar Radio Spectrometers: e-CALLISTO. *Earth Moon Planets* **104**, 277. DOI.

Chen, Z., Ma, L., Xu, L., Tan, C., Yan, Y.: 2016, Imaging and representation learning of solar radio spectrums for classification. *Multimed. Tools Appl.* **75**, 2859. DOI.

Gómez-Herrero, R., Pacheco, D., Kollhoff, A., Espinosa Lara, F., Freiherr von Forstner, J.L., Dresing, N., Lario, D., Balmaceda, L., Krupar, V., Malandraki, O.E., Aran, A., Bučík, R., Klassen, A., Klein, K.-L., Cernuda, I., Eldrum, S., Reid, H., Mitchell, J.G., Mason, G.M., Ho, G.C., Rodríguez-Pacheco, J., Wimmer-Schweingruber, R.F., Heber, B., Berger, L., Allen, R.C., Janitzek, N.P., Laurenza, M., De Marco, R., Wijsen, N., Kartavykh, Y.Y., Dröge, W., Horbury, T.S., Maksimovic, M., Owen, C.J., Vecchio, A., Bonnin, X., Kruparova, O., Pí ša, D., Souček, J., Louarn, P., Fedorov, A., O'Brien, H., Evans, V., Angelini, V., Zucca, P., Prieto, M., Sánchez-Prieto, S., Carrasco, A., Blanco, J.J., Parra, P., Rodríguez-Polo, O., Martín, C., Terasa, J.C., Boden, S., Kulkarni, S.R., Ravanbakhsh, A., Yedla, M., Xu, Z., Andrews, G.B., Schlemm, C.E., Seifert, H., Tyagi, K., Lees, W.J., Hayes, J.: 2021, First near-relativistic solar electron events observed by EPD onboard Solar Orbiter. *Astron. Astrophys.* **656**, L3. DOI. ADS.

Guo, J.C., Yan, F.B., Wan, G., Hu, X.J., Wang, S.: 2022, A deep learning method for the recognition of solar radio burst spectrum. *PeerJ Comput. Sci.* **1**, 36. DOI.

He, H., Garcia, E.A.: 2009, Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **21**, 1263. DOI.

Hou, Y.C., Zhang, Q.M., Feng, S.W., Du, Q.F., Gao, C.L., Zhao, Y.L., Miao, Q.: 2020, Identification and extraction of solar radio spikes based on deep learning. *Solar Phys.* **295**, 146. DOI. ADS.

Klein, K.-L., Salas Matamoros, C., Zucca, P.: 2018, Solar radio bursts as a tool for space weather forecasting. *C. R. Phys.* **19**, 36. DOI.

Klein, K.-L., Musset, S., Vilmer, N., Briand, C., Krucker, S., Francesco Battaglia, A., Dresing, N., Palmroos, C., Gary, D.E.: 2022, The relativistic solar particle event on 28 October 2021: evidence of particle acceleration within and escape from the solar corona. *Astron. Astrophys.* **663**, A173. DOI.

Lobzin, V.V., Cairns, I.H., Robinson, P.A., Steward, G., Patterson, G.: 2009, Automatic recognition of type III radio bursts: the Automated Radio Burst Identification System method and first observations. *Astrophys. J. Lett.* **7**, S04002. DOI.

Lobzin, V.V., Cairns, I.H., Robinson, P.A., Steward, G., Patterson, G.: 2010, Automatic recognition of coronal type II radio bursts: the Automated Radio Burst Identification System method and first observations. *Astrophys. J. Lett.* **710**, L58. DOI.

Ma, Q., Du, Q.F., Feng, S.W., Hou, Y.C., Ji, W.Z., Han, C.S.: 2022, Solar radio-burst forecast based on a convolutional neural network. *Solar Phys.* **297**, 130. DOI. ADS.

Mahender, A., Sasikumar Raja, K., Ramesh, R., Panditi, V., Monstein, C., Ganji, Y.: 2020, A statistical study of low-frequency solar radio type III bursts. *Solar Phys.* **295**, 153. DOI.

Manataki, M., Vafidis, A., Sarris, A.: 2021, Comparing Adam and SGD optimizers to train AlexNet for classifying GPR C-scans featuring ancient structures. In: *2021 11th International Workshop on Advanced Ground Penetrating Radar (IWAGPR)*, 1. DOI.

Moujahid, A.: 2016, *A Practical Introduction to Deep Learning with Caffe and Python*. adilmoujahid.com/posts/2016/06/introduction-deep-learning-python-caffe. Accessed on 12 December 2022.

Ndacyayisenga, T., Umuhire, A.C., Uwamahoro, J., Monstein, C.: 2021, Space weather study through analysis of solar radio bursts detected by a single-station CALLISTO spectrometer. *Ann. Geophys. (EGU)* **39**, 945. DOI.

Ongsulee, P.: 2017, Artificial intelligence, machine learning and deep learning. In: *Proc. 15th Internat. Conf. ICT Knowl. Eng.* DOI.

Prieto, M., Bussons, J., Rodríguez-Pacheco, J., Martínez, A., Sánchez, S., Russu, A., Monstein, C., Fernández, R.: 2020, Increase in interference levels in the 45 – 870 MHz band at the Spanish e-CALLISTO sites over the years 2012 and 2019. *Solar Phys.* **295**, 11. DOI.

Scully, J., Flynn, R., Carley, E., Gallagher, P., Daly, M.: 2021, Type III solar radio burst detection: a deep learning approach. In: *32nd Irish Signals Systems Conf. (ISSC)*, 1. DOI.

Singh, D., Sasikumar Raja, K., Subramanian, P., Ramesh, R., Monstein, C.: 2019, Automated detection of solar radio bursts using a statistical method. *Solar Phys.* **294**, 112. DOI.

Zhang, A., Lipton, Z.C., Li, M., Smola, A.J.: 2020, Dive into deep learning. d2l.ai. Accessed on 21 December 2022.

Zucca, P., Carley, E.P., McCauley, J., Gallagher, P.T., Monstein, C., McAteer, R.T.J.: 2012, Observations of low frequency solar radio bursts from the Rosse Solar-Terrestrial Observatory. *Solar Phys.* **280**, 591. DOI.