DEEPGRAVILENS: A MULTI-MODAL ARCHITECTURE FOR CLASSIFYING GRAVITATIONAL LENSING DATA

Nicolò Oreste Pinciroli Vago
Department of Electronics, Information and Bioengineering
Politecnico di Milano
Via Giuseppe Ponzio, 34
nicolooreste.pinciroli@polimi.it

Piero Fraternali
Department of Electronics, Information and Bioengineering
Politecnico di Milano
Via Giuseppe Ponzio, 34
piero.fraternali@polimi.it

ABSTRACT

Gravitational lensing is the relativistic effect generated by massive bodies, which bend the space-time surrounding them. It is a deeply investigated topic in astrophysics and allows validating theoretical relativistic results and studying faint astrophysical objects that would not be visible otherwise. In recent years Machine Learning methods have been applied to support the analysis of the gravitational lensing phenomena by detecting lensing effects in data sets consisting of images associated with brightness variation time series. However, the state-of-art approaches either consider only images and neglect time-series data or achieve relatively low accuracy on the most difficult data sets. This paper introduces DeepGraviLens, a novel multi-modal network that classifies spatio-temporal data belonging to one non-lensed system type and three lensed system types. It surpasses the current state of the art accuracy results by $\approx 3\%$ to $\approx 11\%$, depending on the considered data set. Such an improvement will enable the acceleration of the analysis of lensed objects in upcoming astrophysical surveys, which will exploit the petabytes of data collected, e.g., from the Vera C. Rubin Observatory.

Keywords Multi-modal Deep Learning · Fusion · Gravitational Lensing · Time series

1 Introduction

In astrophysics, a gravitational lens is a matter distribution (e.g., a black hole) able to bend the trajectory of transiting light, similar to an optical lens. Such apparent distortion is caused by the curvature of the geometry of space-time around the massive body acting as a lens, a phenomenon that forces the light to travel along the geodesics (i.e., the shortest paths in the curved space-time). Strong and weak gravitational lensing focus on the effects produced by particularly massive bodies (e.g., galaxies and black holes), while microlensing addresses the consequences produced by lighter entities (e.g., stars). This research proposes an approach to automatically classify strong gravitational lenses with respect to the lensed objects and to their evolution through time.

Automatically finding and classifying gravitational lenses is a major challenge in astrophysics. As [103, 91, 39, 44] show, gravitational lensing systems can be complex, ubiquitous and hard to detect without computer-aided data processing. The volumes of data gathered by contemporary instruments make manual inspection unfeasible. As an example, the Vera C. Rubin Observatory is expected to collect petabytes of data [108].

Moreover, strong lensing is involved in major astrophysical problems: studying massive bodies that are too faint to be analyzed with current instrumentation; characterizing the geometry, content and kinematics of the universe; and investigating mass distribution in the galaxy formation process [103]. Discovery is only the first step, yet a fundamental

one, in the study of gravitational lenses. Finding evidence of strong gravitational lensing enables the validation and the advancement of existing astrophysical theories, such as the theory of general relativity [91], and supports specialized studies aimed at modeling the effects of gravitational lensing on specific entities, such as wormholes [91], Simpson-Visser black holes [39], and Einstein-Gauss-Bonnet black holes [44].

The gravitational lenses discovery task takes as input spatiotemporal observations consisting of images and time series and associates each observation with one class (e.g., "Lens", "No lens", "Lensed galaxy"...). Images are obtained from specific regions of the electromagnetic field (e.g., visible and infrared [17], ultra-violet [82], and green, red, and near-infrared [64]), depending on the specific experiment. Time series are also collected in specific electromagnetic field regions. They typically describe brightness variation through time (e.g. [64, 105]), and their sampling frequency depends on the technological constraints of the acquisition instrument. In general, they can be multivariate time series [73, 72]. Observations can be either real (i.e., collected by actual instruments) or simulated (i.e., generated by a software system that replicates the characteristics of real instruments).

Several gravitational lenses discovery approaches and tools have been introduced in the past. Originally, observations were analyzed without the aid of computers [118]. Even after the advent of computer science, observations were initially processed without automated classification systems [33, 49, 104]. More recently, Machine Learning (ML) methods have been exploited. The works [18, 100] use Convolutional Neural Network (CNNs) to classify gravitational lensing images, [62] exploits a Bayesian approach to categorize image data, and [64] applies a multi-modal approach to classify spatio-temporal data in four simulated data sets generated by the `deeplenstronomy` simulator [66]. In particular, [64] classifies gravitational lensing data by applying a CNN to the image and a Long Short-Term Memory (LSTM) network to the brightness time series and then fusing the outputs of the two branches, achieving a test accuracy ranging from 48.7% to 78.5%.

State-of-the-art lens detection systems, however, still present several limitations. Some of them (e.g., [18, 100, 62]) rely on images only and neglect time-domain data and thus cannot detect transient phenomena such as supernovae explosions, which are of great importance for estimating the rate of expansion of the Universe [65]. The work [64] considers spatio-temporal data, but the proposed DeepZipper multi-modal (image + time series) multi-class ("no lens", "lensed galaxy", "lensed type-Ia supernova", "lensed core-collapse supernova") classification architecture shows relatively low accuracy on the most challenging simulated data sets. Moreover, the simulated data set, as presented in [64], contains ≈ 4000 samples in each test set, obtained after an eight-fold data augmentation. The unique samples, before augmentation, then, amount to 500, making the number of test samples of some sub-classes (e.g., "SN-Ia") low (≈ 14 samples) and yielding high uncertainty on the test set accuracy results.

The authors of [64] have recently proposed the DeepZipper II architecture [65], which exploits a multi-modal (image + time series) binary ("lensed supernova" vs "other") classification architecture similar to that of [64], achieving an accuracy of 93% over a mix or real and simulated data. The work [48], applies to image time series (i.e., sequences of images), but the classifier works only on the observations where a supernovae is known to be present to infer if lensing has occurred or not.

Multi-modal classification architectures have been exploited in many fields other than astrophysics (e.g., remote sensing and medicine) [84, 28, 27]. Only a few approaches consider the combination of a single image and one or more time series [23, 43, 26], and most approaches are similar to the architecture proposed in [64, 65]. Other modalities have been also considered (e.g., videos and texts), but such inputs differ from those relevant to astrophysical observations and thus such architectures do not carry over the gravitational lensing discovery task. Section 2 briefly surveys them.

Finally, the evaluation of an automatic system for gravitational lens classification poses specific challenges due to the very nature of the task. In real astrophysical observations, gravitational lenses, especially lensed supernovae, are extremely rare and only a few discoveries have already been validated by the scientific community. The extreme scarcity of ground truth data (i.e., verified discoveries) challenges both training and testing of classification algorithms and motivates the use of simulators for creating synthetic data sets. Such data sets can be used for training, validating and testing a classifier in the usual way. However, when it comes to real data, evaluation can only be done a posteriori by submitting the candidate lensing phenomena to the expert judgement for verification.

This paper presents DeepGraviLens, a novel architecture for the classification of strong gravitational lensing multi-modal data. The considered classes concern both transient and non-transient phenomena, and this research shows the superiority of DeepGraviLens over other spatio-temporal networks not only at finding gravitational lenses, but also at finding gravitationally-lensed supernovae, rare objects of particular interest to the astrophysical community. The contributions can be summarized as follows:

- We introduce the architecture of DeepGraviLens, which takes in input spatio-temporal data of real or simulated astrophysical observations and produces in output a multi-class single-label classification of each spatio-temporal sample. DeepGraviLens exploits three complementary sub-networks trained independently

and combines their outputs by means of a SVM final stage. The three sub-networks apply different and complementary ways of combining image and time-series data, taking advantage of both the local and the global features of the input data.

- We evaluate the designed architecture on four simulated data sets formed by $\approx$ 20000 unique examples, split into a train set with $\approx$ 14000 samples (70% of the data set), a validation set with $\approx$ 3000 samples (15% of the data set), and a test set with $\approx$ 3000 samples (15% of the data set). We compare the predictions of DeepGraviLens to the results obtained by the DeepZipper network [64] and by a version of DeepZipper II [65] extended from 2 to 4 classes. DeepGraviLens yields accuracy improvements ranging from $\approx 10\%$ to $\approx 36\%$ with respect to the best version of DeepZipper on each test set and significantly reduces the confusion between similar classes, one of the major issues of gravitational lenses classification.
- We have also compared DeepGraviLens with STNet [23], a spatio-temporal multi-modal neural network recently proposed in remote sensing applications, with an improvement in accuracy ranging from $\approx 3\%$ to 11%.
- Finally, we demonstrate that DeepGraviLens is able to detect the presence of gravitational lenses, and specifically gravitationally-lensed supernovae, in real Dark Energy Survey (DES) data [20].

The obtained improvements in the classification of lensing phenomena will enable a faster and more accurate characterization of future real observations, such as those of the Vera C. Rubin Observatory, and will open the way to the discovery of lensed supernovae, which are among the hardest bodies to detect due to their rarity, scattered spatial distribution and relatively short observable life [40, 68, 20, 31, 112].

The rest of this paper is organized as follows: Section 2 surveys the related work; Section 3 describes the data set and the architecture of DeepGraviLens; Section 4 describes the adopted evaluation protocol and presents quantitative and qualitative results; finally, Section 5 draws the conclusions and outlines our future work.

2 Related Work

This section surveys the previous research in the fields of automated gravitational lensing analysis and multi-modal Deep Learning, which are the foundations of this work.

2.1 Automated Gravitational Lensing Analysis

Classifying gravitational lensing phenomena is a challenging task and the subject of many studies. This section concentrates on data-driven techniques, as opposed to the analytical methods that focus on the design of mathematical models capable of explaining the observed data. It considers the specific case of lensed supernovae, as representatives of transient phenomena, as they are particularly interesting for the astrophysics community. Some of the most recent and promising approaches are listed in Table 2.1.

In gravitational lens search, finding lensed supernovae (LSNe) is challenging, as they are rare and fast transient phenomena. The main challenges connected with rarity have been thoroughly analyzed in [88]. A common problem across several lens-finding approaches is the lack of large data sets comprising a sufficient number of real gravitational lens observations. The work [76], then, proposes a training set with mock lenses and real non-lensed data, which is a widespread strategy. Several works [65, 8, 48] also test their trained models on real data and propose some candidate gravitational lenses.

The second major challenge is considering the transient nature of supernovae. The explosion of a supernova leads to a peak in its brightness, which first increases and can then decline at a slower rate in a few months [37]. The benefits of considering brightness time series in the LSNe case have been illustrated by [64, 65], and [48] uses image time series to consider brightness variability. [64] justifies the extraction of brightness time series from image time series noticing that the differences between images in a series are negligible in 17 representative sub-classes of lensed and non-lensed astrophysical objects. For this reason, their input is formed by a representative image and a normalized brightness time series. The work [48], instead, uses image time series for finding lensed supernovae and shows promising results on simulated data. However, it considers only two classes: non-lensed supernovae and lensed supernovae, while [64] considers also other astrophysical objects, both lensed and non-lensed, making the input used by [48] a particular case of theirs.

The work described in [62] applies a Bayesian approach to classify high-resolution images of non-transient phenomena to reproduce the categorization performed by human experts. However, high-resolution images are not always available and the human classification ("Definitely not a lens", "Possibly a lens", "Probably a lens", "Definitely a lens") is intrinsically imprecise and prone to bias depending on the human classifier.

Table 1: This table summarizes the main approaches for finding gravitational lenses using data-driven techniques. In the "Metric" column, "*" indicates that the metric was computed on real data. The "Real data" column indicates whether the algorithm was tested also on real data, the "Trans." column indicates whether transient phenomena are considered, "LSNe class" indicates whether the "LSNe" class is present in the data set, and "Class. type" is the classification type, which can be either binary (B) or multi-class (M)

Paper	Year	Algorithm	Survey	Metric	Metric result	Real data	LSNe discoveries	Trans.	LSNe class	Class. type
[65]	2022	Multi-modal NN	DES	Accuracy	0.930	Y	Y	Y	Y	B
[48]	2022	Spatiotemporal NN	YSE	Accuracy	0.950	Y	Y	Y	Y	B
			LSST	Accuracy	0.990	N	None			
[87]	2022	CNN committee	CFIS	Precision*	0.014	Y	N/A	N	N	B
[64]	2022	Multi-modal NN	DES (DES-wide)	Accuracy	0.487	N	None	Y	Y	M
			DES (DES-deep)	Accuracy	0.573	N	None			
			DES (DESI-DOT)	Accuracy	0.735	N	None			
			LSST	Accuracy	0.785	N	None			
[96]	2021	Tree-based	Gaia	Found lenses*	14	Y	N/A	Y	N	B
[8]	2020	CNN	Pan-STARRS 3π	Accuracy	0.942	Y	N/A	N	N	B
[11]	2020	Rule-based	HSC	FPR	2.30%	Y	N/A	Y	N	B
[10]	2020	HSC	Rule-based	Found lenses*	6	Y	N/A	Y	N	B
[51]	2020	CNN	KiDS	FPR	<0.4%	Y	N/A	N	N	B
[12]	2020	ConvAE and BGM	Euclid	Accuracy	0.773	N	N/A	N	N	B
[76]	2019	CNN	KiDS	Recall*	0.750	Y	N/A	N	N	B
[19]	2019	Tree-based	Gaia	AUC*	0.997	Y	N/A	N	N	B
[78]	2019	CNN	KiDS	Precision*	0.025	Y	N/A	N	N	B
[46]	2019	Tree-based	KiDS	Precision*	0.013	Y	N/A	N	N	M
[74]	2018	CNN	Various	Accuracy	0.982	N	N/A	N	N	B
[89]	2018	CNN committee	GGSLC	AUC	0.988	N	N/A	N	N	B
[34]	2017	SVM	Euclid	AUC	0.89	N	N/A	N	N	B
			KiDS	AUC	0.95	Y				
[77]	2017	CNN	KiDS	Precision*	0.029	Y	N/A	N	N	B
[62]	2009	Bayes	HSC	Completeness	0.900	Y	N/A	N	N	M

An alternative to Bayesian methods [34] relies on domain-specific features and separates lensed and non-lensed systems using an SVM, whose output is assessed by human experts. The classifier obtains, in the best case, an AUC of 0.95 on simulated data, but the presence of manually-defined features makes this approach less general than deep learning methods. In particular, it exploits specific hard-coded characteristics of lenses, such as the prevalence of a specific color, which can hardly generalize to multi-label classification tasks or to scenarios where transient phenomena are relevant.

Deep learning-based methods rely mostly on Convolutional Neural Networks (CNNs), as in the binary classifiers illustrated in [18], [76] and in [80], which do not consider time-domain information nor support the fine-grain classification of lensed systems. The work [8] exploits a CNN architecture and tests it also on real data, reporting good results on a binary classification problem that does not focus on LSNe. The authors observe that in their experiments CNN performances relied "heavily on the design of lens simulations and on the choice of negative examples for training, but little on the network architecture". The works [64, 65] argue instead that architecture design can lead to great improvements in the results, reporting that multi-modal architectures outperform single-modality CNNs on transient phenomena data. The work [74] describes a CNN-based algorithm trained and tested only on simulated data, which achieves an accuracy of 98% and finds the position of the gravitational lens in the input image. However, the classifier is binary and does not consider LSNe. An interesting approach has been proposed in [89], which focuses on the binary classification of simulated data and proposes a committee of networks, yielding an improvement with respect to individual networks.

As an alternative to supervised methods, [12] defines an unsupervised method for binary classification, which first uses an autoencoder to denoise the image (reducing its resolution), then applies a second autoencoder to extract features from the denoised image, and finally exploits a Bayesian Gaussian Mixture (BGM) to cluster the extracted features. This approach, however, requires human intervention for associating labels to clusters corresponding to the lensed objects.

Several works focused on finding other gravitationally lensed transient phenomena, such as quasars. [64] shows that, compared to supernovae, the brightness of quasars changes in a timescale of several years, because they are not explosive phenomena. For this reason, many studies targeting lensed quasars do not use time series information. [46] exploits the image magnitudes in different bands, which is an ad-hoc method that would need adaptation to be applied to the LSNe search. [11] also focuses on finding lensed quasars, but aims at finding quadruply-lensed quasars using an essentially rule-based pipeline. While this method can be effective for the specific application, it should also be modified to tackle more generic and complex cases.

Differently from binary approaches, DeepZipper [64] casts the problem as a multi-class single-label classification task for data sets consisting of images associated with time series of brightness variation. To analyze both images and time-series data, the authors propose a multi-modal network, formed by a CNN and an LSTM, whose outputs are then fused. The resulting system is applied to four simulated data sets corresponding to different astronomical surveys (DES-wide, LSST-wide, DES-deep, and DESI-DOT). This approach, although relatively simple, achieves relatively good results on all four data sets, with accuracies ranging from 48.7% to 78.5%. DeepZipper II [65], an evolution of DeepZipper, introduces minor changes to the network, casts the problem as a binary classification task ("LSNe" vs "other") instead of a multi-class one, and performs testing on a new data set partially based on real data. It reaches an accuracy of 93% on DES data and a false positive rate of 0.02%. Three new candidate lensed supernovae found in the DES survey are offered to the astrophysical scientific community for confirmation.

DeepGraviLens, similarly to DeepZipper, casts the problem as a multi-class single-label classification task, on the same types of classes and data sets. Compared to previous approaches, it employs more effective unimodal networks and more advanced fusion techniques, which improve the effectiveness in dealing with shared information between the two modalities.

2.2 Multi-modal Deep Learning and Fusion

Several phenomena in the most varied disciplines are characterized by heterogeneous data that give complementary information about the subject under investigation. Multi-modal DL has proved its effectiveness in those domains that require the integrated analysis of multiple data types (e.g., images, videos, and time series). The survey [84] overviews the advances and the trends in multi-modal DL until 2017 and documents usage in such areas as medicine [5, 9, 52], human-computer interaction [4] and autonomous driving [29, 58]. The recent survey [98] discusses several applications combining image and text [54, 55], video and text [56, 83], and text and audio [22, 92]. Some applications rely on physiological signals for behavioral studies, such as face recognition [15, 50, 42]. In the medical field, [93] overviews the use of AI in oncology and shows the benefits of multi-modal DL. The work [113] diagnoses cervical dysplasia with the integrated analysis of images and numerical data. [30] employs multi-modal DL for classifying malware using textual data from different sources. [107] exploits images and texts to detect hate speech in memes. [85] uses multiple robotic sensors (e.g., cameras, tactile and force sensors) for object manipulation.

From the architecture viewpoint, the processing of heterogeneous inputs can be performed by analyzing the individual data types separately and then fusing the outcome of the different branches to produce an output (late fusion), by stacking the inputs, which are processed together (early fusion), or by introducing fusion at a middle stage (intermediate fusion) [94, 84]. The survey [28] overviews DL methods for multi-modal data fusion in general whereas [94] focuses on biomedical data fusion. The work [98] broadens the comparison beyond DL and contrasts alternative methods employed in multi-modal classification tasks, including SVMs [38], RNNs [42, 109], CNNs [35, 69] and even GANs [110]. The combination of a single image and a time series has been considered by a few works, mainly in the remote sensing [32] and medicine [3] fields. It is apparently similar to the problem of classifying data formed by a video and a time series [86, 7, 81]. However, the combination of a single image and a time series, differently from the case of videos, does not require addressing the time-dependent synchronization, connection and interaction between modalities [53]. Another similar case is the joint analysis of image and text. However, text processing poses different challenges and adopts different methods with respect to numeric signals [24]. Another correlated problem is classifying image time series (i.e., sequences of images), as done in several remote sensing applications (e.g., [97, 75, 21]). This task, addressed also by [48] for gravitational lensing data, is best applied when images in the time series vary noticeably. In gravitational lensing data applications such as the one addressed in this paper, instead, the images in the series have small variations. In such a scenario, the use of time series is preferred to the us of image sequences and can be regarded as the extraction of the relevant features from the image sequence [64, 65].

2.2.1 Image and time series analysis

The data considered by DeepGraviLens are formed by a single image (the average of the real or simulated observations) and a time series (representing the brightness variation through the observation). Table 2 summarizes the most representative works based on the combination of a single image and a time series.

Table 2: This table summarizes representative approaches based on the combination of an image and a time series

Paper	Year	Field of application
[47]	2022	Remote sensing
[41]	2022	Remote sensing
[23]	2022	Remote sensing
[60]	2022	Medicine
[64]	2022	Astrophysics
[67]	2021	Medicine
[43]	2021	Medicine
[26]	2020	Remote sensing
[70]	2018	Music genre classification
[106]	2018	Medicine

In the medicine field, [106] focuses on the classification of Parkinson's disease severity. It proposes an architecture based on convolutional neural networks to analyze both time series and image data so that the network focuses on local features [116]. The authors show the advantage of a multi-modal approach with respect to the unimodal ones. The work [71] considers images, time series, and audio, and proposes a multi-modal approach to classify emotions. The fusion process relies on the computation of RMSE on continuous values predicted by the network, and assigns weights to different modalities based on the errors associated with them. Different from [106], this method relies on the comparison with GT continuous values (namely, arousal and valence) to determine the weights used during fusion. For this reason, this approach is not extendable to case studies which lack the GT usable for quantifying the prediction errors. The work [67] focuses on diagnosing two heart-related syndromes and proposes the use of ECGs and chest X-rays given in input to a multi-modal network exploiting CNNs for both images and time series. The works in [60, 43] are two similar approaches that employ multi-modal networks for COVID-19 prediction. Both consider audio signals (for cough, speech, and breathing) and CT scans of the patient's lung. Two networks (one for audio signals, and the other for images) are trained independently. Then, the outcomes are combined with tree-based approaches. In particular, [60] shows that using a decision tree for fusion is more beneficial than using MaxVoting. These approaches are different from the one proposed in [64] because the fusion parameters are learned during a joint training of the two sub-networks.

Representative works in the field of remote sensing have focused on crop yield prediction [41], air pollution prediction [47], crop classification [26], and urban informal settlements classification [23]. The work [41] proposes an approach for predicting crop yield in Ecuador considering spatio-temporal data. It combines a CNN, for image data, an LSTM, for time series, and a FCNN for late fusion, similarly to the architectures of [64, 65]. The work [47] also focuses on a prediction problem and combines a CNN and an LSTM-based subnetwork. Late fusion is performed by finding the optimal weights associated with each output feature obtained from the unimodal networks. The work [26] addresses the problem of crop classification, uses a CNN for the input image and compares different networks (LSTM, CNN, BiLSTM) for the temporal data, showing that the use of a CNN achieves slightly better performances. The final classification step is performed by fusing the unimodal networks decisions using SVM. The work [23] aims at classifying urban informal settlements and proposes a transformer-based approach for fusion.

The combination of images and time series has proven to be beneficial also in the field of music genre classification [70]. This work considers the audio signal and the album cover to classify music. The proposed network uses two CNNs to analyze both modalities, similarly to [67], and fuses their feature vectors using a FCNN.

The work in [117] proposes a decision-level fusion approach that leverages the uncertainty associated with each modality, employing a Softplus activation function to quantify uncertainty. This method aims to enhance the credibility of the model's output by considering the uncertainty of each modality, thereby improving the accuracy of the overall results. It has been proposed for generic input modalities, so it can be adapted to the combination of images and time series.

DeepGraviLens introduces a novel approach for the classification of images and time series. The proposed architecture exploits three multi-modal networks whose results are assembled using SVM. The three multi-modal networks consider the data in different ways: LoNet exploits intermediate fusion and emphasizes the local features of the image; GloNet

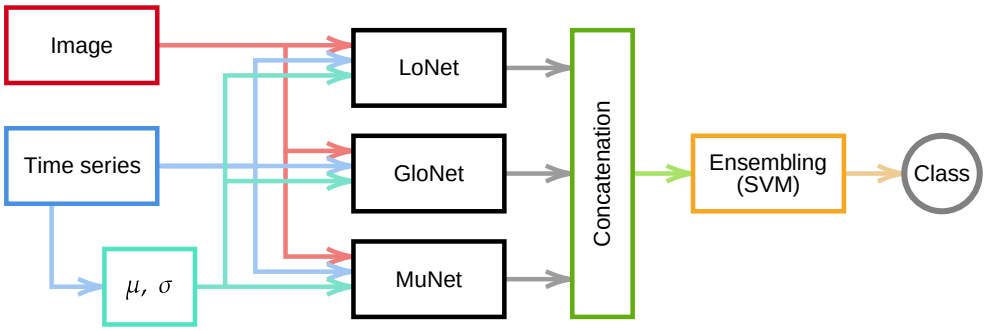

Figure 1: The DeepGraviLens pipeline comprises four steps: (1) the inputs are fed into three independent networks (LoNet, GloNet, and MuNet); (2) the outputs of the three networks are concatenated; (3) the ReFuse network receives the concatenated outputs and (4) outputs a predicted class

applies early fusion and accentuates the global features of both types of input; MuNet employs intermediate fusion but extracts both local and global features from the image.

3 Data Sets and Methods

3.1 Data sets

An input to the lensed object classification task consists of four images and four brightness variation time series, which together represent an astrophysical observation. One image and one time series are provided for each band of the *griz* photometric system, widely used in CCD cameras [90]. In this system, the g band is centered on green, the r band is centred on red, the i band is the near-infrared one, and the z band is the infrared one.

Each input is labeled with one of four classes: "No Lens" (no lensed system), "Lens" (Galaxy-Galaxy lensing), "LSNIa" (the lensed object is a Type-Ia supernova), and "LSNCC" (the lensed object is a core-collapse supernova). Section 4.2 shows various examples of input samples and of their classification by DeepGraviLens.

Four distinct data sets (DESI-DOT, LSST-wide, DES-wide, and DES-deep) are built via simulation and are used for training and evaluating DeepGraviLens. The details of their construction are similar to the ones presented in [66, 6, 64]. Each data set simulates a current or next-generation cosmic survey and is characterized by different specifications of the images and of the associated time series.

The DESI-DOT data set simulates the observations made by the Dark Energy Camera (DECam) [25] and mirrors the real observing conditions of the DES wide-field survey reported in [1]. The exposure time, a simulation parameter that affects the image quality (higher is better), was set to 60 seconds. The LSST-wide data set simulates the LSST survey images acquired using the LSSTCam camera [95]. The simulation parameters were estimated from the conditions of the first year of the survey and the exposure time was set to 30 seconds [61]. The DES-wide data set emulates the images from the DECam and uses the real observing conditions from the DES wide-field survey, but the exposure time is 90 seconds. The DES-deep data set also reproduces the images from DECam but its characteristics are simulated according to the DES SN program [2] with the exposure time set to 200 seconds.

Due to the use of the four-bands *griz* photometric system, each image has 4 layers. The image size is $45 \times 45 \times 4$ pixels for all the four data sets. The length of the time series depends on the technical limitations of the simulated instruments. DESI-DOT, LSST-wide, and DES-deep time series contain 14 samples for each band, while DES-wide contains 7 samples for each band.

For each data set, 17 astrophysical systems were defined and grouped into the four classes "No Lens", "Lens", "LSNIa", and "LSNCC" as proposed in [64]. The examples of the four classes were generated randomly: each class covers $\approx 25\%$ of each data set and the distribution of the 17 subsystems is the same in all the data sets. Each data set comprises $\approx 20,000$ elements, split into the train set ($\approx 70\%$), the validation set ($\approx 15\%$), and the test set ($\approx 15\%$).

3.2 Extraction of statistical quantities

Two statistical quantities (mean μ and standard deviation σ) are extracted from the brightness time series and used as inputs. Such derived data have a physical meaning. For example, an empty sky is expected to have approximately the

same mean value for the four bands and a high standard deviation (because the fluctuations are random). A non-lensed star is expected to be characterized by a low standard deviation, as the means are approximately constant. Even when they manifest a transient behavior (e.g., the explosion of a supernova), the brightness variation is attenuated by the distance. Lensed bodies instead are expected to have a higher standard deviation, because when they display a transient behavior their brightness is amplified by the lens. The contribution of such derived inputs is quantified in the ablation study described in Section 4.

3.3 Overall architecture

Figure 1 illustrates the multi-stage multi-modal inference pipeline of DeepGraviLens. It is formed by three sub-networks (LoNet, GloNet, and MuNet), whose outputs are ensembled using SVM. LoNet and MuNet, in turn, rely on unimodal sub-networks focusing on local or global features in the images and time series. Table 3 summarizes the characteristics of the three networks. GloNet exploits the combination of the image and time-series data, which are merged using early fusion. This approach emphasizes the global features of the multi-modal inputs. LoNet focuses on the local features of the distinct data types: the image and the time series pass through two separate sub-networks and then intermediate fusion is applied. Finally, MuNet extracts both local and global features from the image, using an FC sub-network and a CNN in parallel, and then applies intermediate fusion. The next sections present the three proposed multi-modal networks.

Table 3: The three sub-networks pursue different goals: GloNet emphasizes global features and applies early fusion; LoNet accentuates local features and employs intermediate fusion; MuNet extracts both global and local image features

Fusion type	**Feature extraction**	
	Global	**Local**
Early	GloNet	
Intermediate	MuNet	
		LoNet

3.4 LoNet, a network focusing on local features

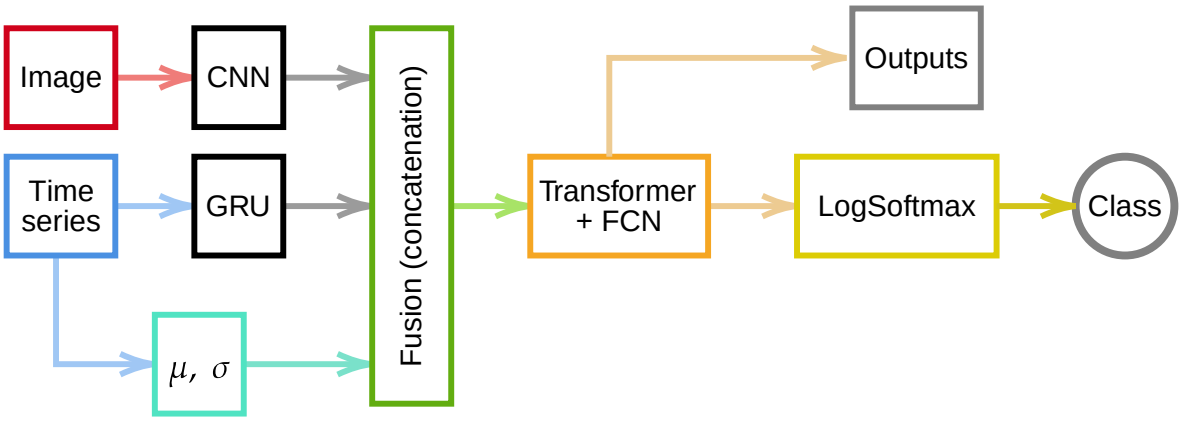

Figure 2: **LoNet** architecture. The time series is processed by the GRU module and the image by a CNN. The two outputs together with the statistics are fused and fed as input to a final transformer module

Figure 2 shows the architecture of the LoNet sub-network and Table 4 summarizes its features. It comprises two branches, one for the image (processed through a CNN) and one for the time series (processed by a GRU). This structure is similar to the one of ZipperNet [64] but replaces the LSTM [36] module with a GRU module with a smaller hidden

unit size [13] and batch normalization. The benefits of GRU over LSTM have been shown in several applications [99, 14, 16, 45, 114]. In the considered data sets, the short length of the time series makes GRU advantageous over LSTM because the former has fewer training parameters and thus better generalization abilities.

The use of CNN for extracting features from images privileges the focus on contiguous pixels (i.e., small regions of the image), as shown in several studies [57, 79, 63].

Two feature vectors from the CNN and the GRU, the means and the standard deviations of the time series are concatenated and fed in input to a Transformer, similarly to [23].

Table 4: Summary of the LoNet neural network architecture showing its layers, output shape, and number of parameters

	Output Shape	**Parameters #**
GRU (features)	[128, 64]	4,068
CNN (features)	[128, 64]	2,337,284
Transformer1d: 1-3	[128, 136]	–
– TransformerEncoder: 2-15	[128, 136]	–
—- ModuleList: 3-2	–	2,537,248
Sequential: 1-4	[128, 32]	–
– Linear: 2-16	[128, 64]	8,768
– ReLU: 2-17	[128, 64]	–
– BatchNorm1d: 2-18	[128, 64]	128
– Dropout: 2-19	[128, 64]	–
– Linear: 2-20	[128, 32]	2,080
– ReLU: 2-21	[128, 32]	–
– BatchNorm1d: 2-22	[128, 32]	64
Sequential: 1-5	[128, 4]	–
– Linear: 2-23	[128, 8]	264
– ReLU: 2-24	[128, 8]	–
– BatchNorm1d: 2-25	[128, 8]	16
– Dropout: 2-26	[128, 8]	–
– Linear: 2-27	[128, 4]	36
Total params: 4,889,956		
Trainable params: 4,889,956		
Non-trainable params: 0		
Total mult-adds (G): 3.40		

3.5 GloNet, a network focusing on global features

Figure 3 shows the architecture of the GloNet sub-network and Table 5 summarizes its features. GloNet, differently from LoNet, applies early fusion and relies on a Fully Connected sub-network applied to the flattened inputs. This approach is complementary to the one of LoNet: it combines the original time series and the original image up-front, rather than merging the features derived from their pre-processing by the GRU and CNN modules. Table 5 also shows that the number of parameters is higher than in LoNet. Having more parameters allows learning from more complex patterns, which compensates for the absence of convolutional layers.

Table 5: Summary of the GloNet neural network architecture showing its layers, output shape, and number of parameters. In this case, a time series of 14 steps is considered

	Output Shape	**Parameters #**
Sequential: 1-1	[128, 32]	–
– Linear: 2-1	[128, 4096]	33,443,840
– ReLU: 2-2	[128, 4096]	–
– BatchNorm1d: 2-3	[128, 4096]	8,192
– Dropout: 2-4	[128, 4096]	–
– Linear: 2-5	[128, 2048]	8,390,656
– ReLU: 2-6	[128, 2048]	–
– BatchNorm1d: 2-7	[128, 2048]	4,096
– Dropout: 2-8	[128, 2048]	–
– Linear: 2-9	[128, 1024]	2,098,176
– ReLU: 2-10	[128, 1024]	–
– BatchNorm1d: 2-11	[128, 1024]	2,048
– Dropout: 2-12	[128, 1024]	–
– Linear: 2-13	[128, 512]	524,800
– ReLU: 2-14	[128, 512]	–
– BatchNorm1d: 2-15	[128, 512]	1,024
– Dropout: 2-16	[128, 512]	–
– Linear: 2-17	[128, 256]	131,328
– ReLU: 2-18	[128, 256]	–
– BatchNorm1d: 2-19	[128, 256]	512
– Linear: 2-20	[128, 128]	32,896
– ReLU: 2-21	[128, 128]	–
– BatchNorm1d: 2-22	[128, 128]	256
– Dropout: 2-23	[128, 128]	–
– Linear: 2-24	[128, 64]	8,256
– ReLU: 2-25	[128, 64]	–
– BatchNorm1d: 2-26	[128, 64]	128
– Dropout: 2-27	[128, 64]	–
– Linear: 2-28	[128, 32]	2,080
Sequential: 1-2	[128, 4]	–
– Linear: 2-29	[128, 8]	264
– ReLU: 2-30	[128, 8]	–
– Linear: 2-31	[128, 4]	36
Total params: 44,648,588		
Trainable params: 44,648,588		
Non-trainable params: 0		
Total mult-adds (G): 5.72		

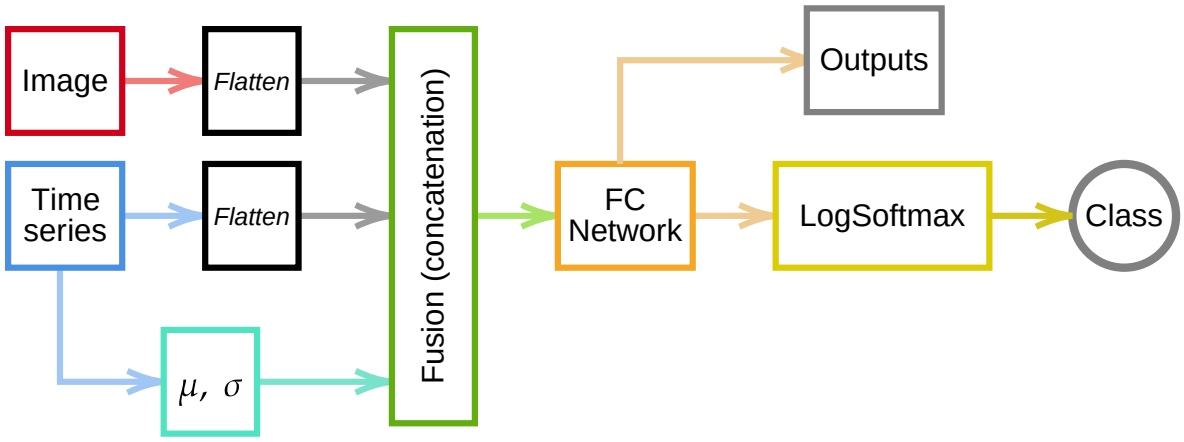

Figure 3: **GloNet** architecture. The input data are (1) flattened, (2) concatenated, and (3) fed to a FC module

3.6 MuNet, a network focusing on local and global features

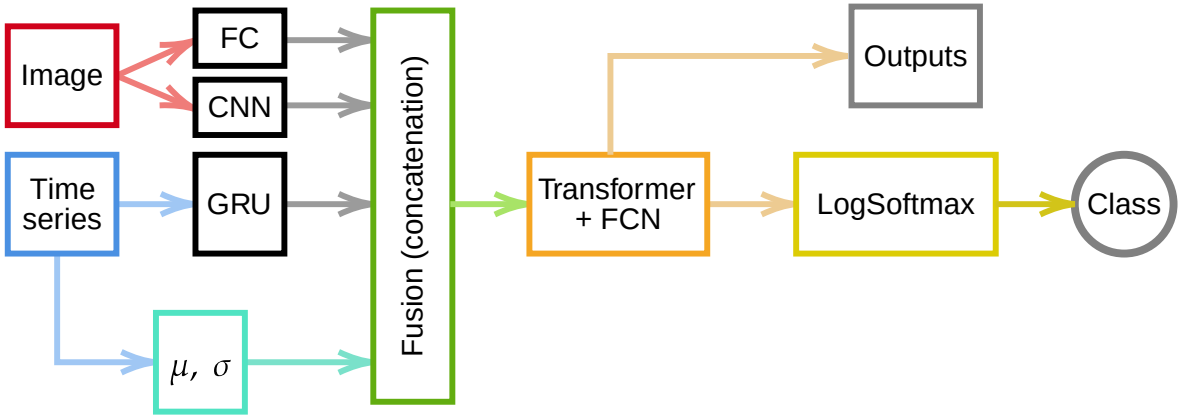

Figure 4: **MuNet** architecture. While LoNet processes the image using only a CNN, MuNet employs both a CNN and a FC component

Figure 4 shows the architecture of the MuNet sub-network and Table 6 summarizes its features. It processes the image using two parallel branches: a CNN and an FC sub-network. The time series is processed in the same way as in LoNet. Compared to LoNet, MuNet adds the FC module applied to the image, to extract local and global features simultaneously. The latter may provide a relevant contribution due to the small size of the images. To avoid overfitting, the number of parameters in the FC sub-network is smaller than in GloNet. In total, the number of parameters is similar to the one of LoNet.

3.7 Ensembling

The three multi-modal networks introduced in this study extract distinct information from the data, emphasizing local features, global features, or a combination of both. To fully leverage the complementary information provided by these networks, ensemble methods can be employed. Table 7 details the ensemble methods used in this study and their associated experimental parameters. For each parameter combination of every method, accuracy is computed on both the train and validation sets. The best parameter combination is then selected based on the highest validation set result, and the accuracy is finally computed on the test set. Moreover, an ablation study is conducted to assess the performance of the best ensemble method when using only two out of the three networks.

Table 6: Summary of the MuNet neural network architecture showing its layers, output shape, and number of parameters

	Output Shape	**Parameters #**
FC: 1-1	[128, 32]	2,337,284
GRU: 1-2	[128, 64]	4,068
CNN: 1-3	[128, 64]	2,118,804
Linear: 1-4	[128, 32]	2,080
Sequential: 1-5	[128, 32]	–
– Linear: 2-23	[128, 64]	8,768
– ReLU: 2-24	[128, 64]	–
– BatchNorm1d: 2-25	[128, 64]	128
– Dropout: 2-26	[128, 64]	–
– Linear: 2-27	[128, 32]	2,080
– ReLU: 2-28	[128, 32]	–
– BatchNorm1d: 2-29	[128, 32]	64
Sequential: 1-6	[128, 4]	–
– Linear: 2-30	[128, 16]	528
– ReLU: 2-31	[128, 16]	–
– BatchNorm1d: 2-32	[128, 16]	32
– Dropout: 2-33	[128, 16]	–
– Linear: 2-34	[128, 8]	136
– ReLU: 2-35	[128, 8]	–
– BatchNorm1d: 2-36	[128, 8]	16
– Dropout: 2-37	[128, 8]	–
– Linear: 2-38	[128, 4]	36
Total params: 4,474,024		
Trainable params: 4,474,024		
Non-trainable params: 0		
Total mult-adds (G): 3.39		

Table 7: Experimental parameters of ensemble methods for aggregating decisions of LoNet, GloNet, and MuNet

Methods	**Parametrization**
AdaBoost	estimators: 200
Random Forest	estimators: [10, 50, 100, 200, 500, 1000, 2000]
Extra Trees	estimators: 100
Fuzzy ranking [59]	None
Average	None
MLP	hidden layer sizes: 100 activation: 'relu' solver: 'adam' alpha: 0.0001
KNN	neighbours: [2, 4, 6, 8, 16, 32]
FCNN	Early stop
Max	None
SVM	C: [10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}] kernel: ['poly', 'linear', 'rbf', 'sigmoid']

3.8 Training

The training process of DeepGraviLens is divided into two stages. In the first step, LoNet, GloNet, and MuNet are trained separately, using the same inputs. The second stage consists of training the SVM, which exploits as inputs the values obtained before the application of the final activation function of the LoNet, GloNet, and MuNet sub-networks. LoNet, GloNet, and MuNet are trained for a maximum of 500 epochs, and the Early Stopping patience is set to 20 epochs. In both stages, the best model is the one with the highest validation accuracy.

4 Evaluation

This section reports the quantitative and qualitative evaluation of DeepGraviLens on the data sets introduced in 3.1.

For each accuracy result, a confidence interval amounting to 1 standard deviation is calculated to take the limited size of the test set into account. C.R. represents the radius of the confidence interval [111]:

$$C.R. = \sqrt{\frac{a \cdot (1-a)}{n}} \tag{1}$$

where a is the mean accuracy (scaled to $[0, 1]$) on the test set and n is the number of samples in the test set.

4.1 Quantitative results

This section presents the outcome of the performance analysis of DeepGraviLens on the four data sets described in Section 3.1. For assessing the improvement induced by the proposed architecture, the approach of [64] is used as a baseline, since it is the only research which used a data set with the same classes as ours. Accuracy is used as the performance metrics because the data set is balanced. In addition, results were compared with other two multi-modal networks using the time and image modalities, presented in Table 8, and with seven unimodal networks, presented in Table 9. Both DeepZipper II [65] and STNet [23] have been adapted to use four classes rather than the original two.

Ablation experiments with respect to the sub-networks preceding the final ensembling stage are also performed to verify their contribution.

4.1.1 Prediction performance

Table 8: **Accuracy** – Comparison of the accuracy of DeepGraviLens and of the best result obtained using state of the art multi-modal methods. An improvement of $\approx 10\%$ to $\approx 36\%$ is achieved with respect to DeepZipper [64], the only work using a data set with the same classes as DeepGraviLens. When compared to the best result obtained by reproducing state of the art approaches, the improvement ranges between $\approx 3\%$ and $\approx 11\%$

	DESI-DOT	**DES-deep**	**DES-wide**	**LSST-wide**
DeepZipper [64]	77.1	58.6	51.7	74.3
DeepZipper II [65]	78.9	57.4	49.8	70.7
STNet [23]	85.1	58.4	82.5	84.3
EvidentialLoNet (Ours)	81.6	65.4	79.9	84.5
EvidentialMuNet (Ours)	81.1	65.6	79.5	84.2
LoNet (Ours)	87.0	67.5	85.8	87.2
GloNet (Ours)	77.2	62.3	76.8	76.8
MuNet (Ours)	87.9	67.9	86.5	88.5
DeepGraviLens (Ours)	**88.7**	**69.6**	**87.7**	**88.8**
Improvement	3.6	11.0	5.2	4.5

Table 9: **Comparison of the unimodal networks and DeepGraviLens** – The table shows the performance of different unimodal networks on image and time modalities, used in Deep Zipper, STNet, and DeepGraviLens. The best unimodal results are highlighted in bold, and the proposed network's performance is underlined

Modality	**Unimodal network**	**Multi-modal networks**	**DESI-DOT**	**DES-deep**	**DES-wide**	**LSST-wide**	**Average**
Image	ResMixer	STNet	**81.4**	65.1	**82.5**	**82.1**	**77.8**
	CNN (Ours)	LoNet, MuNet	78.4	**65.9**	79.7	81.4	76.4
	CNN (DZ)	DeepZipper	74.3	61.9	70.0	74.3	70.1
	FCNN (Ours)	MuNet	67.7	57.7	62.3	59.3	61.8
Time	GRU (DZ)	DeepZipper	**70.9**	28.5	**39.1**	**67.0**	**51.4**
	GRU (Ours)	LoNet, MuNet	69.2	**32.4**	38.9	61.0	50.4
	PDNet	STNet	63.8	28.5	32.9	60.9	46.5
Multi-modal	**DeepGraviLens** (Ours)		**88.7**	**69.6**	**87.7**	**88.8**	**83.7**

Table 8 presents the accuracy results on the four considered test sets. The test set accuracy is similar for the DESI-DOT, LSST-wide and DES-deep data sets and decreases for the more complex DES-wide data set. In all cases, the accuracy shows an improvement with respect to both the DeepZipper baseline and the best method in the state of the art. Such improvement is observed not only in the case of DeepGraviLens, but also for LoNet and GloNet, making them viable alternatives to state-of-the-art approaches. Moreover, the performance of GloNet, a simple network, are similar to the ones of DeepZipper and DeepZipper II.

In addition to LoNet and MuNet, the networks EvidentialLoNet and EvidentialMuNet were also implemented and tested. These networks exploit the evidence-based late fusion approach proposed in [117], which dynamically weights the contribution of each modality based on the degree of uncertainty associated with its predictions. Our experiments show that the proposed intermediate fusion approach outperforms the evidence-based fusion approach, with an average improvement of $\approx 4.5\%$.

Figure 5 illustrates the confusion matrices for the four data sets. For the DES-deep data set, the greatest confusion is observed between "LSNIa" and "LSNCC". A similar, yet more accentuated pattern, was found in [64] too.

For the DES-wide data set, the confusions between classes are similar, different from [64], in which the greatest confusion is between "LSNIa" and "LSNCC". This demonstrates that DeepGraviLens is more effective at discerning

between different gravitationally-lensed transient phenomena, reducing the confusion with respect to the baseline [64] significantly.

For the DESI-DOT data set, the confusion between classes is lower than the one presented in [64]. The greatest confusion is between the "No Lens" and the "Lens" classes, which can be justified by the similarity of the brightness time series of some systems. An example is the "Galaxy + Star" system, in which a galaxy and a star appear close together but without the lensing effect, and the "Galaxy-Galaxy Lensing + Star" system, in which a galaxy stands in front of another galaxy producing the lensing effect and a star appears close to the lensed galaxy from the point of view of the observer.

For the LSST-wide data set the greatest confusion is between the "LSNIa" and the "LSNCC" classes as in DES-deep, similarly to the pattern observed in [64].

The reported results prove that DeepGraviLens can classify the samples of all the data sets accurately and with a significant performance improvement with respect to the compared methods. The results on DES-wide show a significant improvement, reducing the confusion between lensed supernovae classes. This data set is particularly challenging because lensed galaxies are fainter due to the simulated optical depth of the images, which depends on the technical characteristics of the simulated instrumentation. Moreover, the time series are shorter than in the other data sets and thus contain less information.

4.1.2 Ablation studies

Table 10: Comparison of 10 ensemble methods accuracies. The underlined results are the best ones for each data set. The values in bold are the ones comprised in the 1σ confidence interval of the best results. The best performances are obtained using SVM on DESI-DOT, DES-deep, and DES-wide, while Max is the best on LSST-wide

Ensemble method	**DESI-DOT**	**DES-deep**	**DES-wide**	**LSST-wide**	**Average**
AdaBoost	87.2 ± 0.6	66.7 ± 0.9	86.1 ± 0.6	87.9 ± 0.6	82.0
Random Forest	$\mathbf{88.3 \pm 0.6}$	68.6 ± 0.8	87.0 ± 0.6	$\mathbf{88.8 \pm 0.6}$	83.2
Extra Trees	$\mathbf{88.3 \pm 0.6}$	68.7 ± 0.8	$\mathbf{87.1 \pm 0.6}$	$\mathbf{88.8 \pm 0.6}$	83.2
Fuzzy ranking [59]	$\mathbf{88.3 \pm 0.6}$	$\mathbf{68.8 \pm 0.8}$	$\mathbf{87.2 \pm 0.6}$	$\mathbf{88.6 \pm 0.6}$	83.2
Average	$\mathbf{88.1 \pm 0.6}$	$\mathbf{68.8 \pm 0.8}$	$\mathbf{87.3 \pm 0.6}$	$\mathbf{88.7 \pm 0.6}$	83.2
MLP	88.0 ± 0.6	$\mathbf{68.9 \pm 0.8}$	$\mathbf{87.6 \pm 0.6}$	$\mathbf{88.5 \pm 0.6}$	83.3
KNN	$\mathbf{88.3 \pm 0.6}$	68.7 ± 0.8	$\mathbf{87.1 \pm 0.6}$	$\mathbf{88.9 \pm 0.6}$	83.3
FCNN	$\mathbf{88.4 \pm 0.6}$	$\mathbf{69.2 \pm 0.8}$	$\mathbf{87.6 \pm 0.6}$	88.4 ± 0.6	83.4
Max	$\mathbf{88.6 \pm 0.6}$	68.7 ± 0.8	$\mathbf{87.3 \pm 0.6}$	$\underline{\mathbf{89.1 \pm 0.6}}$	83.4
SVM	$\underline{\mathbf{88.7 \pm 0.6}}$	$\underline{\mathbf{69.6 \pm 0.8}}$	$\underline{\mathbf{87.7 \pm 0.6}}$	$\mathbf{88.8 \pm 0.6}$	83.7
Improvement w.r.t. MuNet	0.8	1.7	1.2	0.3	1.0

Table 10 compares SVM with other ensemble methods. The use of SVM brings an average 1% improvement over the best multi-modal network (MuNet) and surpasses the performances of other ensemble methods in three data sets out of four. Considering the LSST-wide data set, Max performs better than SVM, but the SVM result is inside Max's confidence interval. Moreover, Max's accuracy on DES-deep is outside the SVM confidence interval. Considering the analyzed ensemble methods, only SVM, Fuzzy Ranking [59] and Average are inside the confidence interval of the best ensemble approach for all the data sets. However, both Fuzzy Ranking and Average have an accuracy significantly inferior to that of SVM.

Table 11 presents the results of the ablation experiments with respect to the multi-modal sub-networks. The presence of the three sub-networks guarantees the highest accuracy, with the results obtained ensembling one or two networks being often outside the confidence interval of the result obtained by ensembling three networks. In particular, combining three networks yields an improvement ranging from $+0.3\%$ to $+12.0\%$ with respect to single networks, and a change ranging from 0.0% to $+1.7\%$ with respect to the combination of two networks.

In DESI-DOT, the contribution of GloNet is dominated by that of the other two sub-networks and thus eliminating GloNet does not affect accuracy. This can be explained by the use of early fusion in GloNet which does not preserve the information of the image, which is immediately fused with the time series.

Table 11: **Ablation studies on SVM ensemble** – When a single network is considered, accuracy refers to the results obtained by applying it without any additional decision-level algorithm. The underlined results are the best mean accuracy results for every data set, and results in bold are contained within the confidence intervals of the best results. All the values are expressed in %

Data set	LoNet	GloNet	MuNet	Accuracy $\pm$ 1σ
DESI-DOT	✓			87.0 $\pm$ 0.6
		✓		77.2 $\pm$ 0.8
			✓	87.9 $\pm$ 0.6
	✓	✓		87.0 $\pm$ 0.6
	✓		✓	**88.7 $\pm$ 0.6**
		✓	✓	87.9 $\pm$ 0.6
	✓	✓	✓	**88.7 $\pm$ 0.6**
DES-deep	✓			67.5 $\pm$ 0.9
		✓		62.3 $\pm$ 0.9
			✓	67.9 $\pm$ 0.9
	✓	✓		68.4 $\pm$ 0.8
	✓		✓	68.7 $\pm$ 0.8
		✓	✓	68.7 $\pm$ 0.8
	✓	✓	✓	**69.6 $\pm$ 0.8**
DES-wide	✓			85.8 $\pm$ 0.6
		✓		76.8 $\pm$ 0.8
			✓	86.5 $\pm$ 0.6
	✓	✓		**87.2 $\pm$ 0.6**
	✓		✓	**87.3 $\pm$ 0.6**
		✓	✓	86.9 $\pm$ 0.6
	✓	✓	✓	**87.7 $\pm$ 0.6**
LSST-wide	✓			87.2 $\pm$ 0.6
		✓		76.8 $\pm$ 0.8
			✓	**88.5 $\pm$ 0.6**
	✓	✓		87.4 $\pm$ 0.6
	✓		✓	**88.5 $\pm$ 0.6**
		✓	✓	**88.5 $\pm$ 0.6**
	✓	✓	✓	**88.8 $\pm$ 0.6**

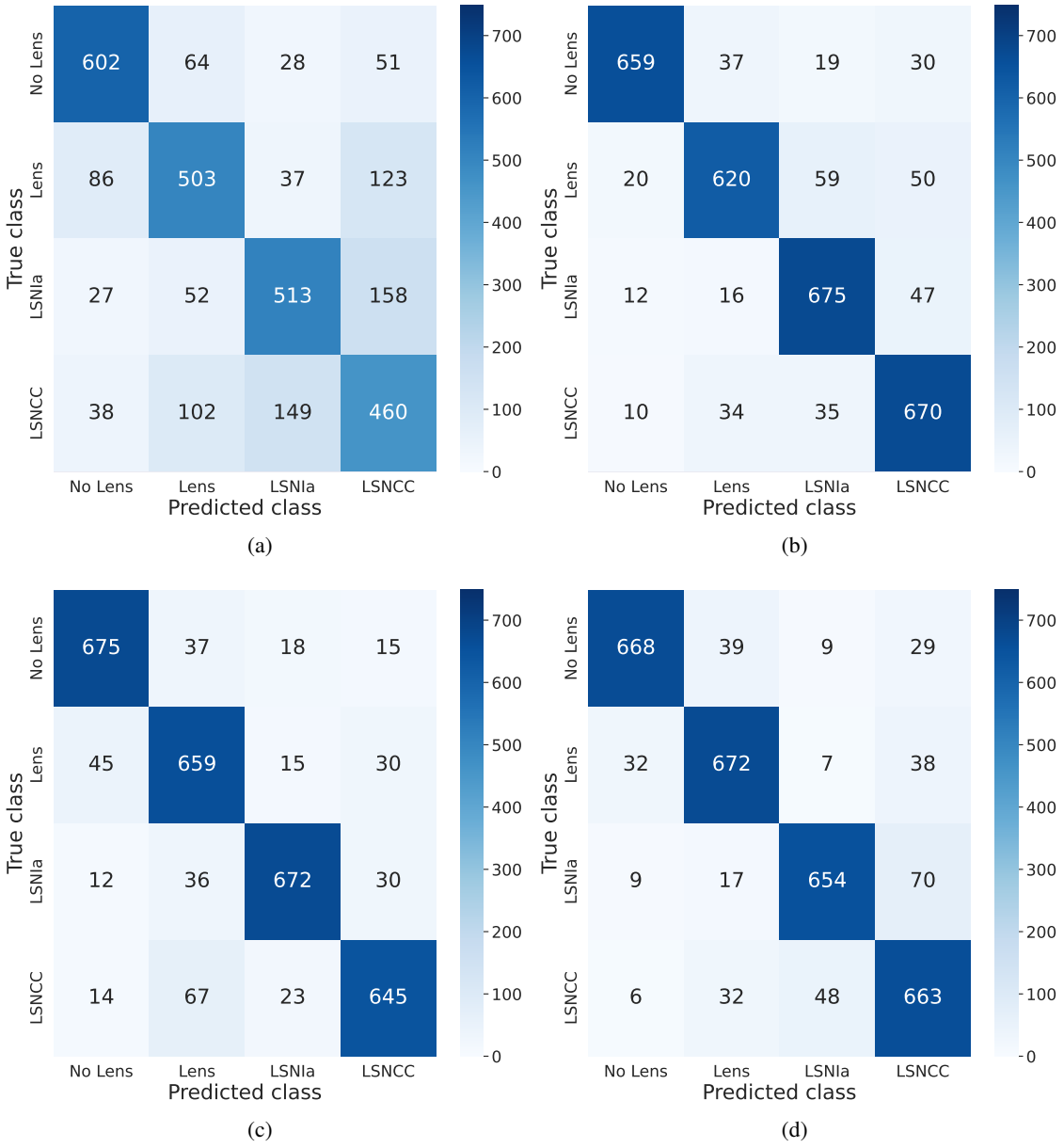

Figure 5: **Confusion matrices** of the (a) DES-deep, (b) DES-wide, (c) DESI-DOT, and (d) LSST-wide data sets. In general, the greatest confusion is observed between "Lens" and "No Lens", and in the case of the DES-wide data set, between "LSNCC" and "LSNIa", due to the low sampling rate

The introduction of the means μ and standard deviations σ of the time series yields an additional modest average improvement of 0.5% in accuracy consistently across the data sets. Compared to the predictions made using a random forest with inputs μ and σ, DeepGraviLens accuracy improves from 18% to 49%.

4.1.3 Execution time

DeepGraviLens has been trained using an NVIDIA GeForce GTX 1080 Ti for GloNet, MuNet and LoNet. On average, the network training requires less than 3 hours for a single data set. SVM training time is negligible with respect to the other networks.

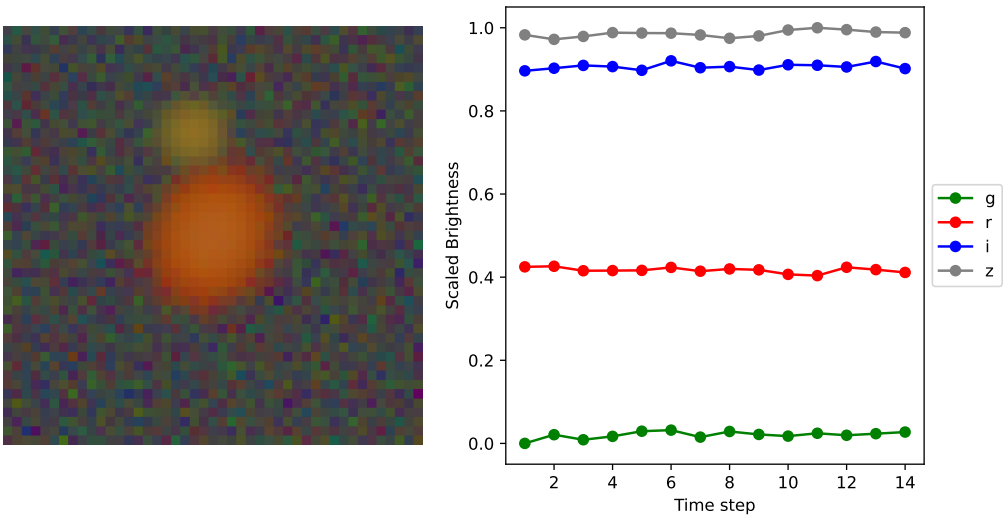

Figure 6: **A positive example on the LSST-wide data set** – This datum belongs to the "No Lens" class. The image shows two separate stars that have a spherical geometry, which suggests they are not lensed. Moreover, the curves on the right show no consistent brightness variation through time, which indicates the absence of transient phenomena

4.2 Qualitative results

Section 4.2.1 illustrates some representative examples of the results obtained by DeepGraviLens on the four test sets. All the images are obtained by adding the *griz* layers, as done in [64]. In the plots, the g band is displayed in green, the r band in red, the i band in blue and the z band in grey.

Section 4.2.2 shows how the application of DeepGraviLens to real data recognizes the presence of gravitational lensing phenomena, also confirming the three lensed supernovae candidate systems, a very rare occurrence, reported in [65].

4.2.1 Simulated data

Figure 6 presents a true positive example belonging to the "No Lens" class in the LSST-wide data set. It shows two stars close to each other, which exhibit a spherical symmetry, which suggests the absence of lensing. In addition, the brightness curves do not show consistent variations, which indicates the absence of transient phenomena.

Figure 7 presents a true positive example belonging to the "Lens" class, in the DESI-DOT data set. In this system, the lensing effect is manifested by the ring pattern on the central body. The flatness of the brightness curves indicates the absence of transient phenomena, as expected, because the system is formed by galaxies, which are not characterized by explosive events.

Figure 8 presents a true positive example belonging to the "LSNIa" class in the DESI-DOT data set. The peak in the time series indicates the presence of an exploding supernova and the image shows an elliptical shape, which signals the presence of lensing. The brightness in the g band is almost flat, which is distinctive of Type Ia supernovae. Type Ia and core-collapse supernovae release chemical elements during the explosion and produce photons at different wavelengths, which are detected by sensors in specific bands. During explosions, the emission of an element with a certain wavelength produces a temporary brightness peak in the corresponding band. Both types of supernovae release chemical elements whose detection can be observed in the g band, but Type-Ia supernovae emit less materials than core-collapse supernovae, which makes the latter exhibit a more pronounced peak in the g band. The absence of such a peak in Figure 8 justifies the "LSNIa" classification.

The same type of system is shown in Figure 9, from the DES-wide data set. In this case, the peaks are not detected because of the lower sampling rate, which misses rapid transient events. However, the network correctly classifies this example thanks to the information contained in the image.

Figure 10 presents a true positive example belonging to the "LSNCC" class, in the DESI-DOT data set. In this case, the presence of a supernova is indicated by the rapid variation in the brightness time series. Since also the g band exhibits a peak, the input is classified as a core-collapse supernova. The lensing effect is manifested in the image by the supernova (the green body), lensed by the galaxy in front of it. The green color confirms the presence of elements emitting photons in the g band and the body itself is visible because of the magnifying effect induced by the galaxy.

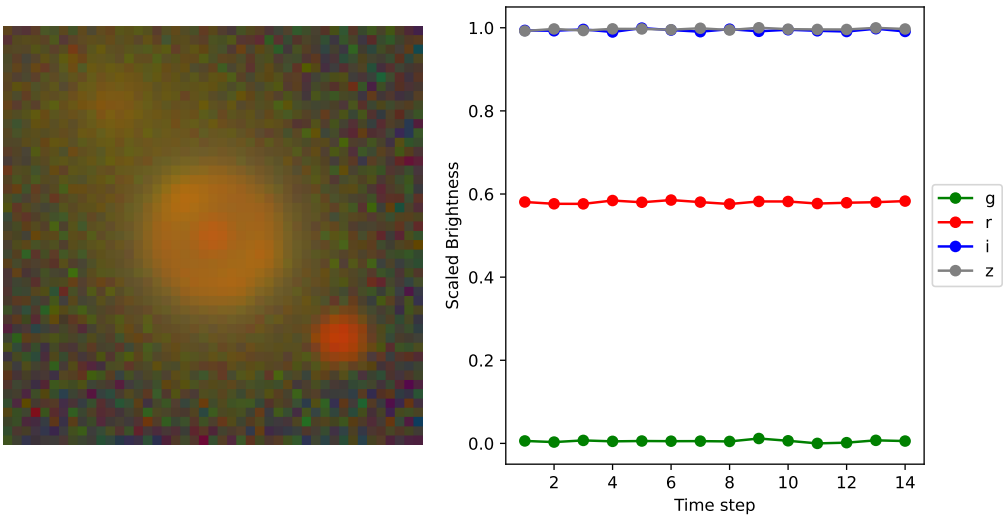

Figure 7: **A positive example on the DESI-DOT data set** – This datum belongs to the "Lens" class. The lensing effect is visible in the ring pattern around the central body. The flatness of the brightness time series, instead, indicates the absence of transient phenomena (e.g., explosions), which is expected because the involved entities are galaxies

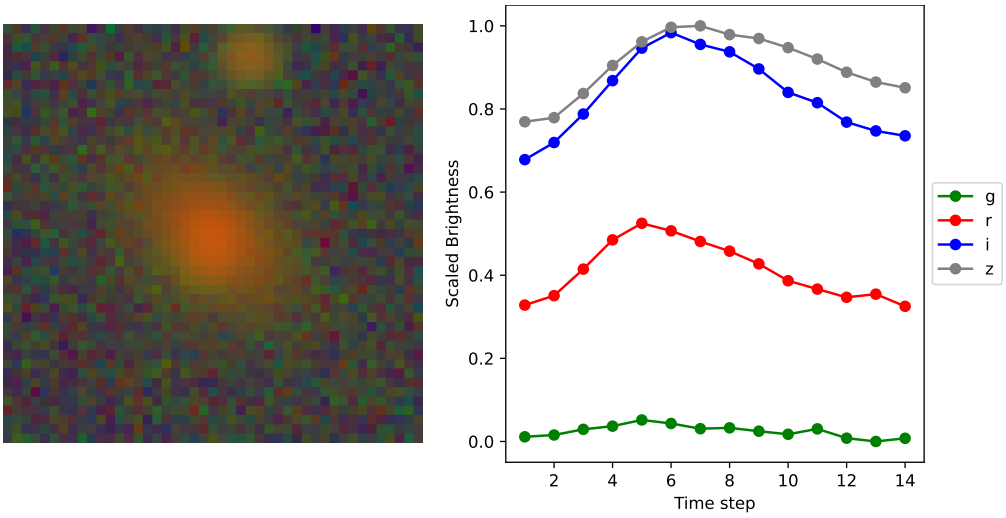

Figure 8: **A positive example on the DESI-DOT data set** – This datum belongs to the "LSNIa" class. The lensing effect is visible from the elliptical shape of the central body, while the presence of a supernova can be observed by the peaks in the brightness time series, which indicates the presence of explosive transient phenomena. The supernova type can be inferred from the flatness of the g band time series

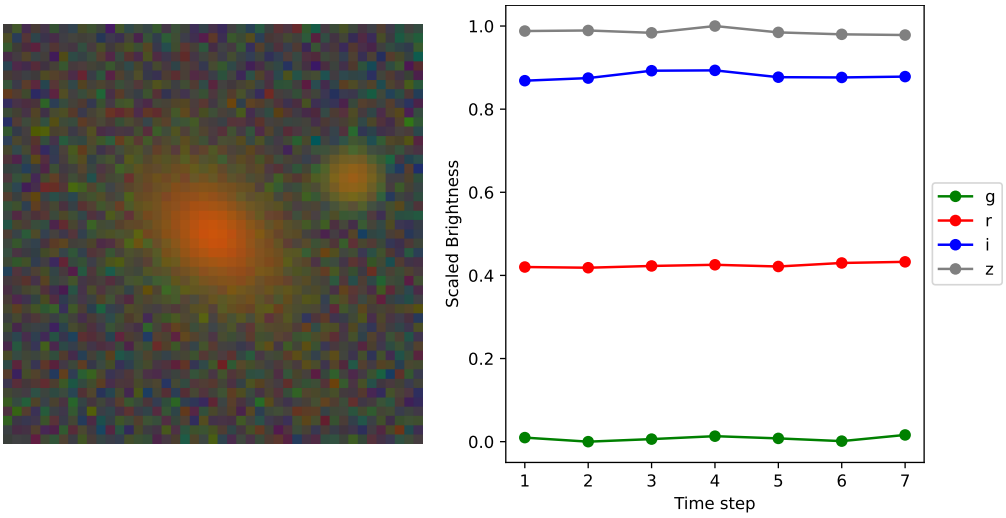

Figure 9: **A positive example on the DES-wide data set** – This datum belongs to the "LSNIa" class. The lensing effect is visible because of the elliptical shape of the central body. Even if the peaks that indicate the presence of transient phenomena are absent, the network is still able to correctly classify the datum

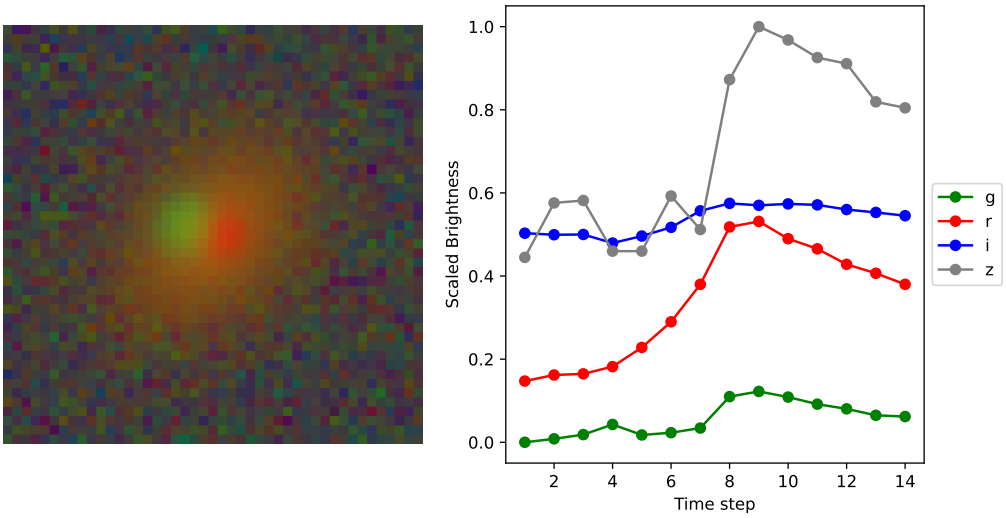

Figure 10: **A positive example on the DESI-DOT data set** – This datum belongs to the "LSNCC" class. In this case, the lensing effect is suggested both by the presence of varying time curves (indicating the presence of a supernova) and the green body lensed by the galaxy

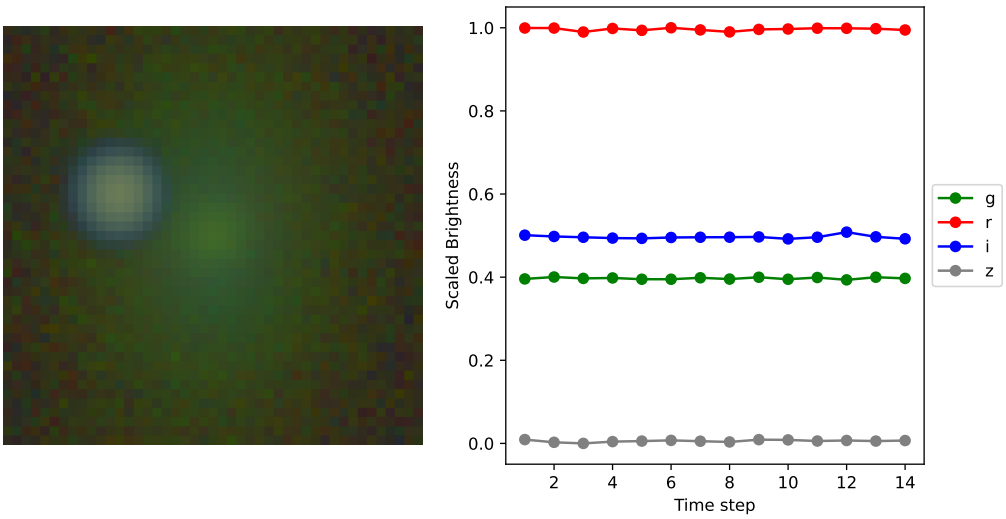

Figure 11: **A negative example on the LSST-wide data set** – This datum belongs to the "LSNCC" class, but has been classified as "Lens". The lensing effect is alluded by the halo surrounding the star, while the flat time series suggests the absence of a transient phenomenon, which induces the wrong classification

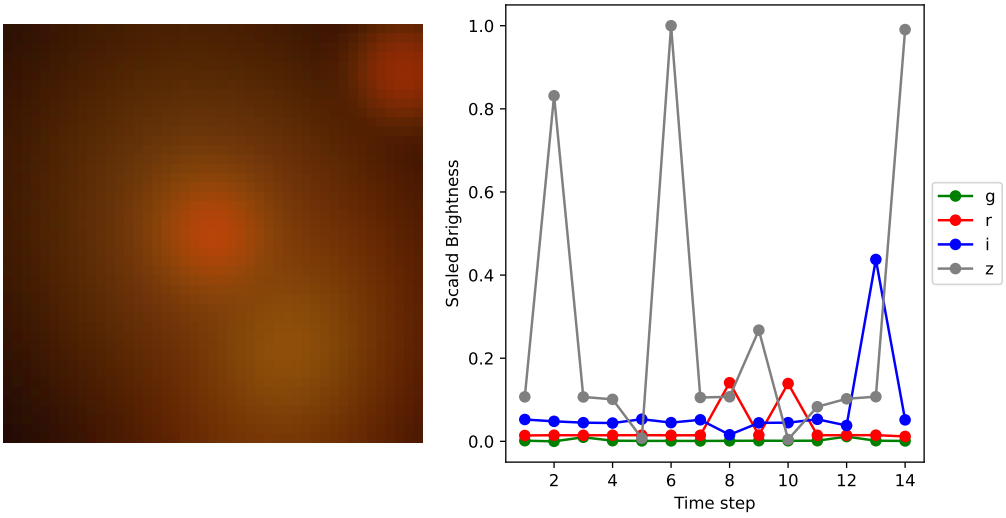

Figure 12: **A negative example on the DES-deep data set** – This datum belongs to the "Lens" class, but has been classified as "No Lens". The lensing effect is suggested by the halo surrounding the central body

Figure 11 presents a negative example in the LSST-wide data set. The datum belongs to the "LSNCC" class, but is classified as "Lens", which means that the model was not able to detect the presence of a supernova and interpreted the example as a lensed system without evident transient phenomena. The wrong classification is caused by the low-quality time series and the ambiguous image. The lensing effect is visible thanks to the faint halo surrounding the star in the background, but the time series (wrongly) suggest the absence of a transient phenomenon. The apparent lack of the transient phenomenon can be explained by considering that supernovae explosions can happen in a short time and the brightness variation may not be recorded by the camera. Soon after the explosion, the brightness returns to the original value, which explains the flatness of the curves.

Figure 12 presents a negative example from the DES-deep data set, belonging to the "Lens" class, but classified as "No Lens". The lensing effect is visible on the central body, which has a halo. However, because of the low image resolution, this effect is not as clear in most of the positive examples. In addition, the presence of multiple peaks is not frequently associated with the "Lens" class and induces the wrong classification.

As a final example, Figure 13 shows an ambiguous image in the DESI-DOT data set, incorrectly classified. The sample belongs to the "No Lens" class, but is classified as "Lens". The confusion is generated chiefly by the elliptical object,

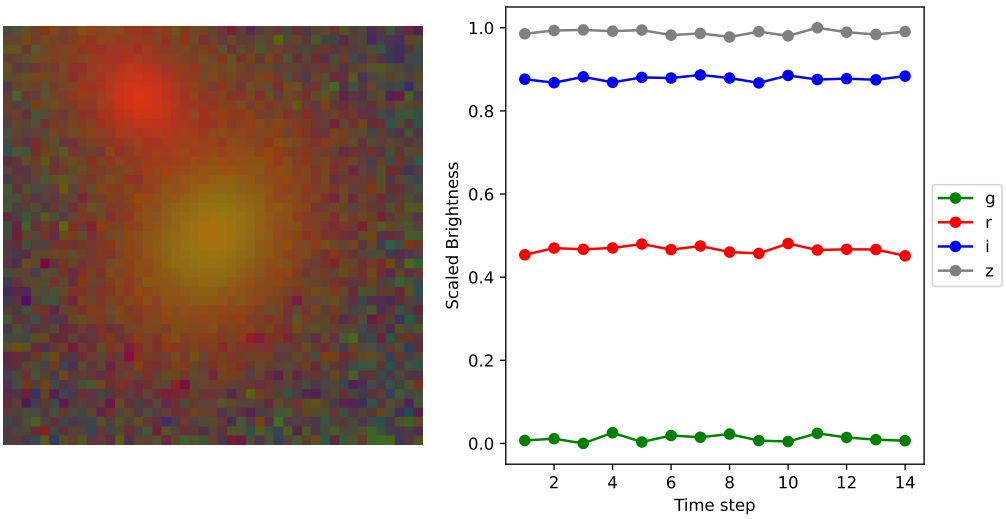

Figure 13: **A negative example on the DESI-DOT data set** – This datum belongs to the "No Lens" class, but it has been classified as belonging to the "Lens" class. The lensing effect is suggested by the elliptical shape, but such shape may suggest also the presence of a non-lensing elliptical galaxy. The flatness of the time series, in addition, does not allow to discern "Lens" and "No Lens" systems, as some "No Lens" systems also have flat time series

which is confused with a lensing effect, while it can represent, e.g., a non-lensed elliptical galaxy. The time series are flat, so they do not help discern "Lens" and "No Lens" systems, because some "No Lens" systems also have flat time series.

4.2.2 Real data

The authors of [65] analyze real data from the Dark Energy Survey over a five-year period (Y1-Y5) with the aim of detecting gravitationally-lensed supernovae. They identify three potential lensed supernova systems (identified as 691022126, 701263907, and 699919273), two of which were detected using only Y5 data, indicating that the supernovae likely exploded during that year. Our research tries to reproduce such results using public data provided by NoirLab[1], which currently only includes data up to Y4, using the network trained on the DES-deep data set.

DeepGraviLens successfully identified the lensed supernova with ID 691022126 and also detected the presence of a gravitational lens for the other two systems. To extract brightness time series, we followed a methodology similar to the one employed in [65], using 14 time steps with a 6-day interval between each step, resulting in a 78-day period. The corresponding image was obtained by averaging the images captured during this period. Each system has been observed for more than 78 days, and as such, multiple observations are associated with each system. Finally, images bigger than 45×45 pixels are resized to such dimension.

Table 12 presents a summary of our results on the real data. The number of observations associated with each system may differ slightly due to missing observations in the database. Our results confirm the findings of [65]. The systems in which a lensed supernova was discovered only in Y5 have a prevalence of "Lens" prediction.

The object with ID 691022126 is shown in Figure 14. It has been classified as "LSNCC" in 65% of the observations. The presence of a gravitational lens is signaled by the multiple objects visible in the image. Additionally, the peaks in the four bands indicate the presence of a supernova and the peak in the g band suggests it belongs to the "LSNCC" class, similarly to the case shown in Figure 10. Figure 15 shows the same system at a different time. Although the four objects are more clearly visible in the image, the time series appears more flat and does not exhibit the typical peaks of exploding supernovae. There are several possible explanations for this. One hypothesis is that the supernova has already exploded and the brightness change is no longer detectable. Another possibility is that real data are inherently more variable than simulated data and noise makes peaks difficult to detect.

Figure 16 presents the system identified with ID 699919273, which exhibits a clear gravitational lens. Additionally, this system contains multiple objects, which are likely to be lensed versions of the same astrophysical object. The authors of [65] classify this system as a gravitationally-lensed supernova, based on Y5 data (not publicly available). With the

[1] `https://datalab.noirlab.edu/` (As of March 2023)

Table 12: Summary of results on the considered real data, including system ID, coordinates, number of observations, predicted class, and the proportion of observations in which that class was observed. Here, RA indicates the right ascension, and DEC indicates the declination

System ID [65]	**Coordinates [deg]**		**# observations**	**Predicted class**	
	RA	**DEC**		**Class**	**Proportion**
691022126	53.898910	-28.912293	77	LSNCC	65%
701263907	40.969218	-0.619054	69	Lens	100%
699919273	10.155917	-44.437515	76	Lens	93%

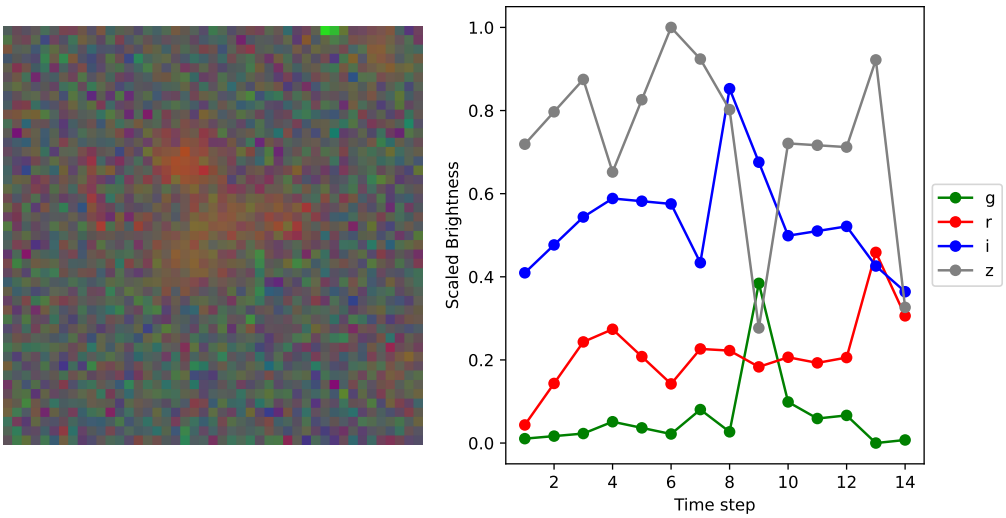

Figure 14: **The detection of a real gravitationally-lensed supernova** – This system is formed by four objects, whose boundaries are not well-defined. The time series shows the presence of peaks in the four bands. The presence of a peak in the g band suggests the presence of a LSNCC, as predicted by DeepGraviLens

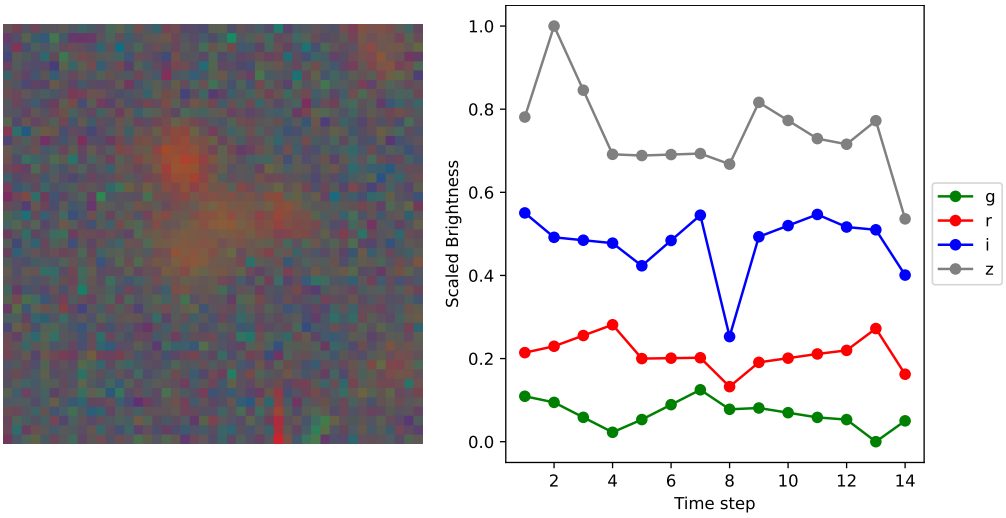

Figure 15: **The missed detection of a real gravitationally-lensed supernova** – The system presented in this figure is the same as the one in Figure 14, but the time series, for this time interval, does not show significant peaks, suggesting the absence of a transient phenomenon. The clearer separation between the four bodies in the image is not enough for suggesting the presence of a supernova

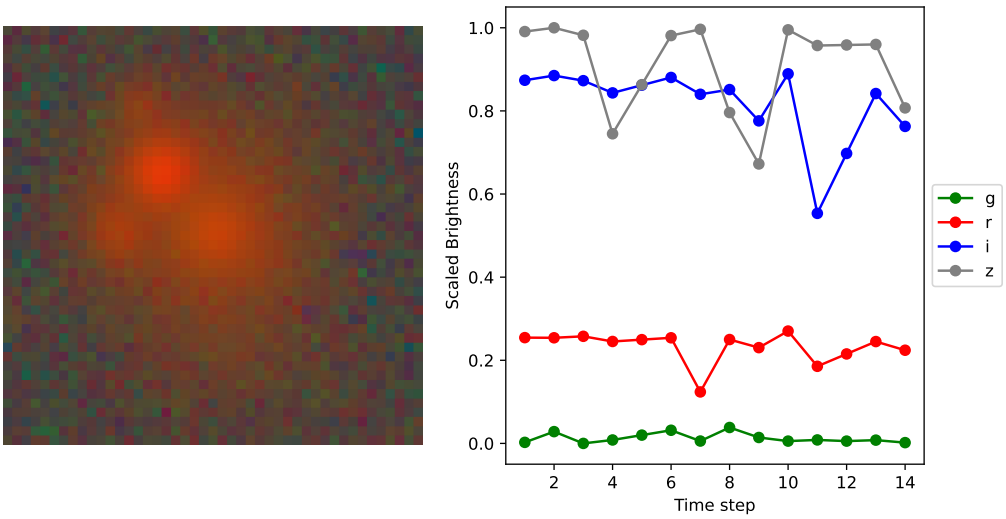

Figure 16: **A real gravitational lens** – The system presented in this figure has been classified as a gravitationally-lensed supernova by [65]. However, the detection was performed on the fifth year of the observation, which is not publicly available. At the time of the observation, the lens is already present, but the supernova explosion is not visible yet. The time series, indeed, are almost flat or noisy

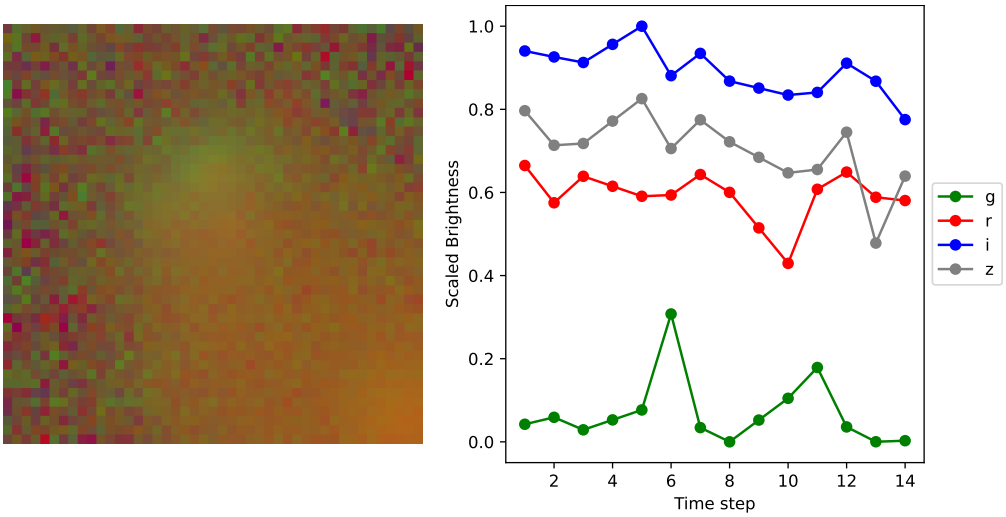

Figure 17: **A real gravitational lens** – The system presented in this figure has been indicated as a gravitationally-lensed supernova by [65]. However, the detection was performed on the fifth year of the observation, which is not publicly available. Before, the lens is already present, but the supernova explosion is not visible yet. The time series, indeed, are almost flat or noisy

available data up to Y4, the system is classified as a "Lens," which confirms the category assigned by [65] with the public data up to Y4.

Figure 17 presents the more complex system with ID 701263907, in which the identification of individual objects is challenging due to their blurred boundaries. The presence of halos around the central bodies and in the bottom-right corner of the image suggests the existence of a gravitationally-lensed object. It is possible that the lens extends beyond the boundaries of the image, further complicating its identification. The absence of evident peaks in the time series data suggests the absence of transient phenomena. Specifically, the peaks observed in the g band do not correspond with significant peaks in other bands, indicating the absence of relevant transient effect. Similar to system 699919273, data up to Y4 hint at the presence of a lens, which DeepGraviLens correctly identifies.

5 Conclusions and Future Work

This work has introduced DeepGraviLens, a neural architecture for the classification of simulated and real gravitational lensing phenomena that processes multi-modal inputs by means of sub-networks focusing on complementary data aspects. DeepGraviLens surpasses the state-of-the-art accuracy results by $\approx 3\%$ to $\approx 11\%$ on four simulated data sets with different data quality. In particular, it attains a 4.5% performance increase on the LSST-wide data set, which simulates the acquisitions of the Vera C. Rubin Observatory whose operations are scheduled to start in 2023. The Vera C. Rubin Observatory is expected to detect hundreds to thousands of lensed supernovae systems, which represents a breakthrough with respect to the capacity of previous instruments. The enormous amount of data that will be acquired demands highly accurate and fast computer-aided classification tools, such as DeepGraviLens.

Future work will concentrate on the application of DeepGraviLens to real observations as soon as they become available. The envisioned research work will also pursue the objective of creating a scientist-friendly system that allows experts to import and manually classify data from real observations to create a non-simulated data set and compute relevant classification and object detection metrics for automated data analysis, following an approach similar to the one implemented in [101, 102, 115]. Finally, we plan to employ the multi-modal architecture designed for DeepGraviLens for the analysis of other (possibly non-astrophysical) data sets characterized by images and time series.

Acknowledgments

The authors thank Robert Morgan, author of [64], for having provided guidance in the use of the `deeplenstronomy` simulator and the generation and use of the data set, and professors Hans Georg Schaathun and Ben David Normann for proofreading the initial version of the article.

References

[1] ABBOTT, T. M. C., ABDALLA, F. B., ALLAM, S., AMARA, A., ANNIS, J., ASOREY, J., AVILA, S., BALLESTER, O., BANERJI, M., BARKHOUSE, W., BARUAH, L., BAUMER, M., BECHTOL, K., BECKER, M. R., BENOIT-LÉVY, A., BERNSTEIN, G. M., BERTIN, E., BLAZEK, J., BOCQUET, S., BROOKS, D., BROUT, D., BUCKLEY-GEER, E., BURKE, D. L., BUSTI, V., CAMPISANO, R., CARDIEL-SAS, L., ROSELL, A. C., KIND, M. C., CARRETERO, J., CASTANDER, F. J., CAWTHON, R., CHANG, C., CHEN, X., CONSELICE, C., COSTA, G., CROCCE, M., CUNHA, C. E., D'ANDREA, C. B., DA COSTA, L. N., DAS, R., DAUES, G., DAVIS, T. M., DAVIS, C., VICENTE, J. D., DEPOY, D. L., DEROSE, J., DESAI, S., DIEHL, H. T., DIETRICH, J. P., DODELSON, S., DOEL, P., DRLICA-WAGNER, A., EIFLER, T. F., ELLIOTT, A. E., EVRARD, A. E., FARAHI, A., NETO, A. F., FERNANDEZ, E., FINLEY, D. A., FLAUGHER, B., FOLEY, R. J., FOSALBA, P., FRIEDEL, D. N., FRIEMAN, J., GARCÍA-BELLIDO, J., GAZTANAGA, E., GERDES, D. W., GIANNANTONIO, T., GILL, M. S. S., GLAZEBROOK, K., GOLDSTEIN, D. A., GOWER, M., GRUEN, D., GRUENDL, R. A., GSCHWEND, J., GUPTA, R. R., GUTIERREZ, G., HAMILTON, S., HARTLEY, W. G., HINTON, S. R., HISLOP, J. M., HOLLOWOOD, D., HONSCHEID, K., HOYLE, B., HUTERER, D., JAIN, B., JAMES, D. J., JELTEMA, T., JOHNSON, M. W. G., JOHNSON, M. D., KACPRZAK, T., KENT, S., KHULLAR, G., KLEIN, M., KOVACS, A., KOZIOL, A. M. G., KRAUSE, E., KREMIN, A., KRON, R., KUEHN, K., KUHLMANN, S., KUROPATKIN, N., LAHAV, O., LASKER, J., LI, T. S., LI, R. T., LIDDLE, A. R., LIMA, M., LIN, H., LÓPEZ-REYES, P., MACCRANN, N., MAIA, M. A. G., MALONEY, J. D., MANERA, M., MARCH, M., MARRINER, J., MARSHALL, J. L., MARTINI, P., MCCLINTOCK, T., MCKAY, T., MCMAHON, R. G., MELCHIOR, P., MENANTEAU, F., MILLER, C. J., MIQUEL, R., MOHR, J. J., MORGANSON, E., MOULD, J., NEILSEN, E., NICHOL, R. C., NOGUEIRA, F., NORD, B., NUGENT, P., NUNES, L., OGANDO, R. L. C., OLD, L., PACE, A. B., PALMESE, A., PAZ-CHINCHÓN, F., PEIRIS, H. V., PERCIVAL, W. J., PETRAVICK, D., PLAZAS, A. A., POH, J., POND, C., PORREDON, A., PUJOL, A., REFREGIER, A., REIL, K., RICKER, P. M., ROLLINS, R. P., ROMER, A. K., ROODMAN, A., ROONEY, P., ROSS, A. J., RYKOFF, E. S., SAKO, M., SANCHEZ, M. L., SANCHEZ, E., SANTIAGO, B., SARO, A., SCARPINE, V., SCOLNIC, D., SERRANO, S., SEVILLA-NOARBE, I., SHELDON, E., SHIPP, N., SILVEIRA, M. L., SMITH, M., SMITH, R. C., SMITH, J. A., SOARES-SANTOS, M., SOBREIRA, F., SONG, J., STEBBINS, A., SUCHYTA, E., SULLIVAN, M., SWANSON, M. E. C., TARLE, G., THALER, J., THOMAS, D., THOMAS, R. C., TROXEL, M. A., TUCKER, D. L., VIKRAM, V., VIVAS, A. K., WALKER, A. R., WECHSLER, R. H., WELLER, J., WESTER, W., WOLF, R. C., WU, H., YANNY, B., ZENTENO, A., ZHANG, Y., ZUNTZ, J., JUNEAU, S., FITZPATRICK, M., NIKUTTA, R., NIDEVER, D., OLSEN, K., SCOTT, A., AND AND. The dark energy survey: Data release 1. *The Astrophysical Journal Supplement Series 239*, 2 (nov 2018), 18.

[2] ABBOTT, T. M. C., ALLAM, S., ANDERSEN, P., ANGUS, C., ASOREY, J., AVELINO, A., AVILA, S., BASSETT, B. A., BECHTOL, K., BERNSTEIN, G. M., BERTIN, E., BROOKS, D., BROUT, D., BROWN, P., BURKE, D. L., CALCINO, J., ROSELL, A. C., CAROLLO, D., KIND, M. C., CARRETERO, J., CASAS, R., CASTANDER, F. J., CAWTHON, R., CHALLIS, P., CHILDRESS, M., CLOCCHIATTI, A., CUNHA, C. E., D'ANDREA, C. B., DA COSTA, L. N., DAVIS, C., DAVIS, T. M., VICENTE, J. D., DEPOY, D. L., DESAI, S., DIEHL, H. T., DOEL, P., DRLICA-WAGNER, A., EIFLER, T. F., EVRARD, A. E., FERNANDEZ, E., FILIPPENKO, A. V., FINLEY, D. A., FLAUGHER, B., FOLEY, R. J., FOSALBA, P., FRIEMAN, J., GALBANY, L., GARCÍA-BELLIDO, J., GAZTANAGA, E., GIANNANTONIO, T., GLAZEBROOK, K., GOLDSTEIN, D. A., GONZÁLEZ-GAITÁN, S., GRUEN, D., GRUENDL, R. A., GSCHWEND, J., GUPTA, R. R., GUTIERREZ, G., HARTLEY, W. G., HINTON, S. R., HOLLOWOOD, D. L., HONSCHEID, K., HOORMANN, J. K., HOYLE, B., JAMES, D. J., JELTEMA, T., JOHNSON, M. W. G., JOHNSON, M. D., KASAI, E., KENT, S., KESSLER, R., KIM, A. G., KIRSHNER, R. P., KOVACS, E., KRAUSE, E., KRON, R., KUEHN, K., KUHLMANN, S., KUROPATKIN, N., LAHAV, O., LASKER, J., LEWIS, G. F., LI, T. S., LIDMAN, C., LIMA, M., LIN, H., MACAULAY, E., MAIA, M. A. G., MANDEL, K. S., MARCH, M., MARRINER, J., MARSHALL, J. L., MARTINI, P., MENANTEAU, F., MILLER, C. J., MIQUEL, R., MIRANDA, V., MOHR, J. J., MORGANSON, E., MUTHUKRISHNA, D., MÖLLER, A., NEILSEN, E., NICHOL, R. C., NORD, B., NUGENT, P., OGANDO, R. L. C., PALMESE, A., PAN, Y.-C., PLAZAS, A. A., PURSIAINEN, M., ROMER, A. K., ROODMAN, A., ROZO, E., RYKOFF, E. S., SAKO, M., SANCHEZ, E., SCARPINE, V., SCHINDLER, R., SCHUBNELL, M., SCOLNIC, D., SERRANO, S., SEVILLA-NOARBE, I., SHARP, R., SMITH, M., SOARES-SANTOS, M., SOBREIRA, F., SOMMER, N. E., SPINKA, H., SUCHYTA, E., SULLIVAN, M., SWANN, E., TARLE, G., THOMAS, D., THOMAS, R. C., TROXEL, M. A., TUCKER, B. E., UDDIN, S. A., WALKER, A. R., WESTER, W., WISEMAN, P., WOLF, R. C., YANNY, B., ZHANG, B., AND AND, Y. Z. First cosmology results using type Ia supernovae from the Dark Energy Survey: Constraints on cosmological parameters. *The Astrophysical Journal 872*, 2 (feb 2019), L30.

[3] ARIOZ, U., SMRKE, U., PLOHL, N., AND MLAKAR, I. Scoping review on the multimodal classification of depression and experimental study on existing multimodal models. *Diagnostics 12*, 11 (2022), 2683.

[4] AZAGRA, P., MOLLARD, Y., GOLEMO, F., MURILLO, A. C., LOPES, M., AND CIVERA, J. A multimodal human-robot interaction dataset. In *NIPS 2016, workshop Future of Interactive Learning Machines* (2016).

[5] BANOS, O., VILLALONGA, C., GARCIA, R., SAEZ, A., DAMAS, M., HOLGADO-TERRIZA, J. A., LEE, S., POMARES, H., AND ROJAS, I. Design, implementation and validation of a novel open framework for agile development of mobile health applications. *Biomedical engineering online 14*, 2 (2015), 1–20.

[6] BIRRER, S., SHAJIB, A. J., GILMAN, D., GALAN, A., AALBERS, J., MILLON, M., MORGAN, R., PAGANO, G., PARK, J. W., TEODORI, L., ET AL. lenstronomy ii: A gravitational lensing software ecosystem. *arXiv preprint arXiv:2106.05976* (2021).

[7] CAI, T., NI, H., YU, M., HUANG, X., WONG, K., VOLPI, J., WANG, J. Z., AND WONG, S. T. Deepstroke: An efficient stroke screening framework for emergency rooms with multimodal adversarial deep learning. *Medical Image Analysis 80* (2022), 102522.

[8] CAÑAMERAS, R., SCHULDT, S., SUYU, S., TAUBENBERGER, S., MEINHARDT, T., LEAL-TAIXÉ, L., LEMON, C., ROJAS, K., AND SAVARY, E. Holismokes-ii. identifying galaxy-scale strong gravitational lenses in pan-starrs using convolutional neural networks. *Astronomy & Astrophysics 644* (2020), A163.

[9] CAO, Y., STEFFEY, S., HE, J., XIAO, D., TAO, C., CHEN, P., AND MÜLLER, H. Medical image retrieval: a multimodal approach. *Cancer informatics 13* (2014), CIN–S14053.

[10] CHAN, J. H., SUYU, S. H., SONNENFELD, A., JAELANI, A. T., MORE, A., YONEHARA, A., KUBOTA, Y., COUPON, J., LEE, C.-H., OGURI, M., ET AL. Survey of gravitationally lensed objects in hsc imaging (sugohi)-iv. lensed quasar search in the hsc survey. *Astronomy & Astrophysics 636* (2020), A87.

[11] CHAO, D. C.-Y., CHAN, J. H.-H., SUYU, S. H., YASUDA, N., MORE, A., OGURI, M., MOROKUMA, T., AND JAELANI, A. T. Lensed quasar search via time variability with the hsc transient survey. *Astronomy & Astrophysics 640* (2020), A88.

[12] CHENG, T.-Y., LI, N., CONSELICE, C. J., ARAGÓN-SALAMANCA, A., DYE, S., AND METCALF, R. B. Identifying strong lenses with unsupervised machine learning using convolutional autoencoder. *Monthly Notices of the Royal Astronomical Society 494*, 3 (2020), 3750–3765.

[13] CHO, K., VAN MERRIËNBOER, B., GULCEHRE, C., BAHDANAU, D., BOUGARES, F., SCHWENK, H., AND BENGIO, Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[14] CHUNG, J., GULCEHRE, C., CHO, K., AND BENGIO, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[15] CIMTAY, Y., EKMEKCIOGLU, E., AND CAGLAR-OZHAN, S. Cross-subject multimodal emotion recognition based on hybrid fusion. *IEEE Access 8* (2020), 168865–168878.

[16] COLLINS, J., SOHL-DICKSTEIN, J., AND SUSSILLO, D. Capacity and trainability in recurrent neural networks. *arXiv preprint arXiv:1611.09913* (2016).

[17] DAHLE, H., KAISER, N., IRGENS, R. J., LILJE, P. B., AND MADDOX, S. J. Weak gravitational lensing by a sample of X-ray luminous clusters of galaxies. I. the data set. *The Astrophysical Journal Supplement Series 139*, 2 (2002), 313.

[18] DAVIES, A., SERJEANT, S., AND BROMLEY, J. M. Using convolutional neural networks to identify gravitational lenses in astronomical images. *Monthly Notices of the Royal Astronomical Society 487*, 4 (05 2019), 5263–5271.

[19] DELCHAMBRE, L., KRONE-MARTINS, A., WERTZ, O., DUCOURANT, C., GALLUCCIO, L., KLÜTER, J., MIGNARD, F., TEIXEIRA, R., DJORGOVSKI, S., STERN, D., ET AL. Gaia gral: Gaia dr2 gravitational lens systems-iii. a systematic blind search for new lensed systems. *Astronomy & Astrophysics 622* (2019), A165.

[20] DIEHL, H. T., NEILSEN, E., GRUENDL, R. A., ABBOTT, T. M. C., ALLAM, S., ALVAREZ, O., ANNIS, J., BALBINOT, E., BHARGAVA, S., BECHTOL, K., BERNSTEIN, G. M., BHATAWDEKAR, R., BOCQUET, S., BROUT, D., CAPASSO, R., CAWTHON, R., CHANG, C., COOK, E., CONSELICE, C. J., CRUZ, J., D'ANDREA, C., DA COSTA, L., DAS, R., DEPOY, D. L., DRLICA-WAGNER, A., ELLIOTT, A., EVERETT, S. W., FRIEMAN, J., NETO, A. F., FERTÉ, A., FRISWELL, I., FURNELL, K. E., GELMAN, L., GERDES, D. W., GILL, M. S. S., GOLDSTEIN, D. A., GRUEN, D., GULLEDGE, D. J., HAMILTON, S., HOLLOWOOD, D., HONSCHEID, K., JAMES, D. J., JOHNSON, M. D., JOHNSON, M. W. G., KENT, S., KESSLER, R. S., KHULLAR, G., KOVACS, E., KREMIN, A., KRON, R., KUROPATKIN, N., LASKER, J., LATHROP, A., LI, T. S., MANERA, M., MARCH, M., MARSHALL, J. L., MEDFORD, M., MENANTEAU, F., MOHAMMED, I., MONROY, M., MORAES, B., MORGANSON, E., MUIR, J., MURPHY, M., NORD, B., PACE, A. B., PALMESE, A., PARK, Y., PAZ-CHINCHÓN, F., PEREIRA, M. E. S., PETRAVICK, D., PLAZAS, A. A., POH, J., PROCHASKA, T., ROMER, A. K., REIL, K., ROODMAN, A., SAKO, M., SAUSEDA, M., SCOLNIC, D., SECCO, L. F., SEVILLA-NOARBE, I., SHIPP, N., SMITH, J. A., SOARES-SANTOS, M., SOERGEL, B., STEBBINS, A., STORY, K. T., STRINGER, K., TARSITANO, F., THOMAS, B., TUCKER, D. L., VIVAS, K., WALKER, A. R., WANG, M.-Y., WEAVERDYCK, C., WEAVERDYCK, N., WESTER, W., WETHERS, C. F., WILKENSON, R., WU, H.-Y., YANNY, B., ZENTENO, A., AND ZHANG, Y. Dark energy survey operations: years 4 and 5. In *Observatory Operations: Strategies, Processes, and Systems VII* (2018), A. B. Peck, R. L. Seaman, and C. R. Benn, Eds., vol. 10704, International Society for Optics and Photonics, SPIE, pp. 138 – 155.

[21] DO NASCIMENTO BENDINI, H., FONSECA, L. M. G., SCHWIEDER, M., KÖRTING, T. S., RUFIN, P., SANCHES, I. D. A., LEITAO, P. J., AND HOSTERT, P. Detailed agricultural land classification in the brazilian cerrado based on phenological information from dense satellite image time series. *International Journal of Applied Earth Observation and Geoinformation 82* (2019), 101872.

[22] ELIAS, I., ZEN, H., SHEN, J., ZHANG, Y., JIA, Y., WEISS, R. J., AND WU, Y. Parallel tacotron: Non-autoregressive and controllable tts. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2021), IEEE, pp. 5709–5713.

[23] FAN, R., LI, J., SONG, W., HAN, W., YAN, J., AND WANG, L. Urban informal settlements classification via a transformer-based spatial-temporal fusion network using multimodal remote sensing and time-series human activity data. *International Journal of Applied Earth Observation and Geoinformation 111* (2022), 102831.

[24] FEGHALI, J., JIMENEZ, A. E., SCHILLING, A. T., AND AZAD, T. D. Overview of algorithms for natural language processing and time series analyses. In *Machine Learning in Clinical Neuroscience: Foundations and Applications* (2022), Springer, pp. 221–242.

[25] FLAUGHER, B., DIEHL, H. T., HONSCHEID, K., ABBOTT, T. M. C., ALVAREZ, O., ANGSTADT, R., ANNIS, J. T., ANTONIK, M., BALLESTER, O., BEAUFORE, L., BERNSTEIN, G. M., BERNSTEIN, R. A., BIGELOW, B., BONATI, M., BOPRIE, D., BROOKS, D., BUCKLEY-GEER, E. J., CAMPA, J., CARDIEL-SAS, L., CASTANDER, F. J., CASTILLA, J., CEASE, H., CELA-RUIZ, J. M., CHAPPA, S., CHI, E., COOPER, C., DA COSTA, L. N., DEDE, E., DERYLO, G., DEPOY, D. L., DE VICENTE, J., DOEL, P., DRLICA-WAGNER, A., EITING, J., ELLIOTT, A. E., EMES, J., ESTRADA, J., NETO, A. F., FINLEY, D. A., FLORES, R., FRIEMAN, J., GERDES, D., GLADDERS, M. D., GREGORY, B., GUTIERREZ, G. R., HAO, J., HOLLAND, S. E., HOLM, S., HUFFMAN, D., JACKSON, C., JAMES, D. J., JONAS, M., KARCHER, A., KARLINER, I., KENT, S., KESSLER, R., KOZLOVSKY, M., KRON, R. G., KUBIK, D., KUEHN, K., KUHLMANN, S., KUK, K., LAHAV, O., LATHROP, A., LEE, J., LEVI, M. E., LEWIS, P., LI, T. S., MANDRICHENKO, I., MARSHALL, J. L., MARTINEZ, G., MERRITT, K. W., MIQUEL, R., MUÑOZ, F., NEILSEN, E. H., NICHOL, R. C., NORD, B., OGANDO, R., OLSEN, J., PALAIO, N., PATTON, K., PEOPLES, J., PLAZAS, A. A., RAUCH, J., REIL, K., RHEAULT, J.-P., ROE, N. A., ROGERS, H., ROODMAN, A., SANCHEZ, E., SCARPINE, V., SCHINDLER, R. H., SCHMIDT, R.,

SCHMITT, R., SCHUBNELL, M., SCHULTZ, K., SCHURTER, P., SCOTT, L., SERRANO, S., SHAW, T. M., SMITH, R. C., SOARES-SANTOS, M., STEFANIK, A., STUERMER, W., SUCHYTA, E., SYPNIEWSKI, A., TARLE, G., THALER, J., TIGHE, R., TRAN, C., TUCKER, D., WALKER, A. R., WANG, G., WATSON, M., WEAVERDYCK, C., WESTER, W., WOODS, R., AND AND, B. Y. THE DARK ENERGY CAMERA. *The Astronomical Journal 150*, 5 (oct 2015), 150.

[26] GADIRAJU, K. K., RAMACHANDRA, B., CHEN, Z., AND VATSAVAI, R. R. Multimodal deep learning based crop classification using multispectral and multitemporal satellite imagery. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2020), pp. 3234–3242.

[27] GAO, J., LI, P., CHEN, Z., AND ZHANG, J. A Survey on Deep Learning for Multimodal Data Fusion. *Neural Computation 32*, 5 (05 2020), 829–864.

[28] GAO, J., LI, P., CHEN, Z., AND ZHANG, J. A survey on deep learning for multimodal data fusion. *Neural Computation 32*, 5 (2020), 829–864.

[29] GEIGER, A., LENZ, P., STILLER, C., AND URTASUN, R. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research 32*, 11 (2013), 1231–1237.

[30] GIBERT, D., MATEU, C., AND PLANES, J. HYDRA: A multimodal deep learning framework for malware classification. *Computers & Security 95* (2020), 101873.

[31] GOLDSTEIN, D. A., AND NUGENT, P. E. How to find gravitationally lensed type Ia supernovae. *The Astrophysical Journal Letters 834*, 1 (2016), L5.

[32] GÓMEZ-CHOVA, L., TUIA, D., MOSER, G., AND CAMPS-VALLS, G. Multimodal classification of remote sensing images: A review and future directions. *Proceedings of the IEEE 103*, 9 (2015), 1560–1584.

[33] GORENSTEIN, M., SHAPIRO, I., COHEN, N., COREY, B., FALCO, E., MARCAIDE, J., ROGERS, A., WHITNEY, A., PORCAS, R., PRESTON, R., ET AL. Detection of a compact radio source near the center of a gravitational lens: quasar image or galactic core? *Science 219*, 4580 (1983), 54–56.

[34] HARTLEY, P., FLAMARY, R., JACKSON, N., TAGORE, A., AND METCALF, R. Support vector machine classification of strong gravitational lenses. *Monthly Notices of the Royal Astronomical Society 471*, 3 (2017), 3378–3397.

[35] HAZARIKA, D., PORIA, S., MIHALCEA, R., CAMBRIA, E., AND ZIMMERMANN, R. Icon: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (2018), pp. 2594–2604.

[36] HOCHREITER, S., AND SCHMIDHUBER, J. Long short-term memory. *Neural computation 9*, 8 (1997), 1735–1780.

[37] HOEFLICH, P., KHOKHLOV, A., WHEELER, J. C., PHILLIPS, M. M., SUNTZEFF, N. B., AND HAMUY, M. Maximum brightness and postmaximum decline of light curves of type supernovae ia: a comparison of theory and observations. *The Astrophysical Journal 472*, 2 (1996), L81.

[38] HUANG, Y., YANG, J., LIAO, P., AND PAN, J. Fusion of facial expressions and EEG for multimodal emotion recognition. *Computational intelligence and neuroscience 2017* (2017).

[39] ISLAM, S. U., KUMAR, J., AND GHOSH, S. G. Strong gravitational lensing by rotating Simpson-Visser black holes. *Journal of Cosmology and Astroparticle Physics 2021*, 10 (2021), 013.

[40] IVEZIĆ, Ž., KAHN, S. M., TYSON, J. A., ABEL, B., ACOSTA, E., ALLSMAN, R., ALONSO, D., ALSAYYAD, Y., ANDERSON, S. F., ANDREW, J., ET AL. LSST: from science drivers to reference design and anticipated data products. *The Astrophysical Journal 873*, 2 (2019), 111.

[41] JÁCOME-GALARZA, L.-R. Multimodal deep learning for crop yield prediction. In *Doctoral Symposium on Information and Communication Technologies: Second Doctoral Symposium, DSICT 2022, Manta, Ecuador, October 12–14, 2022, Proceedings* (2022), Springer, pp. 106–117.

[42] JAISWAL, M., ALDENEH, Z., AND MOWER PROVOST, E. Controlling for confounders in multimodal emotion classification via adversarial learning. In *2019 International Conference on Multimodal Interaction* (2019), pp. 174–184.

[43] JAYACHITRA, V., NIVETHA, S., NIVETHA, R., AND HARINI, R. A cognitive iot-based framework for effective diagnosis of covid-19 using multimodal data. *Biomedical Signal Processing and Control 70* (2021), 102960.

[44] JIN, X.-H., GAO, Y.-X., AND LIU, D.-J. Strong gravitational lensing of a 4-dimensional Einstein–Gauss–Bonnet black hole in homogeneous plasma. *International Journal of Modern Physics D 29*, 09 (2020), 2050065.

[45] JOZEFOWICZ, R., ZAREMBA, W., AND SUTSKEVER, I. An empirical exploration of recurrent network architectures. In *International conference on machine learning* (2015), PMLR, pp. 2342–2350.

[46] KHRAMTSOV, V., SERGEYEV, A., SPINIELLO, C., TORTORA, C., NAPOLITANO, N. R., AGNELLO, A., GETMAN, F., DE JONG, J. T., KUIJKEN, K., RADOVICH, M., ET AL. Kids-squad-ii. machine learning selection of bright extragalactic objects to search for new gravitationally lensed quasars. *Astronomy & Astrophysics 632* (2019), A56.

[47] KO, K.-K., AND JUNG, E.-S. Improving air pollution prediction system through multimodal deep learning model optimization. *Applied Sciences 12*, 20 (2022), 10405.

[48] KODI RAMANAH, D., ARENDSE, N., AND WOJTAK, R. AI-driven spatio-temporal engine for finding gravitationally lensed type Ia supernovae. *Monthly Notices of the Royal Astronomical Society 512*, 4 (03 2022), 5404–5417.

[49] LAWRENCE, C., SCHNEIDER, D., SCHMIDT, M., BENNETT, C., HEWITT, J., BURKE, B., TURNER, E., AND GUNN, J. Discovery of a new gravitational lens system. *Science 223*, 4631 (1984), 46–49.

[50] LI, M., XIE, L., LV, Z., LI, J., AND WANG, Z. Multistep deep system for multimodal emotion detection with invalid data in the internet of things. *IEEE Access 8* (2020), 187208–187221.

[51] LI, R., NAPOLITANO, N., TORTORA, C., SPINIELLO, C., KOOPMANS, L., HUANG, Z., ROY, N., VERNARDOS, G., CHATTERJEE, S., GIBLIN, B., ET AL. New high-quality strong lens candidates with deep learning in the kilo-degree survey. *The Astrophysical Journal 899*, 1 (2020), 30.

[52] LIANG, M., LI, Z., CHEN, T., AND ZENG, J. Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. *IEEE/ACM transactions on computational biology and bioinformatics 12*, 4 (2014), 928–937.

[53] LIANG, P. P., ZADEH, A., AND MORENCY, L.-P. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430* (2022).

[54] LIU, M., HU, H., LI, L., YU, Y., AND GUAN, W. Chinese image caption generation via visual attention and topic modeling. *IEEE transactions on cybernetics* (2020).

[55] LIU, M., LI, L., HU, H., GUAN, W., AND TIAN, J. Image caption generation with dual attention mechanism. *Information Processing & Management 57*, 2 (2020), 102178.

[56] LIU, S., REN, Z., AND YUAN, J. Sibnet: Sibling convolutional encoder for video captioning. *IEEE transactions on pattern analysis and machine intelligence 43*, 9 (2020), 3259–3272.

[57] LUO, W., LI, Y., URTASUN, R., AND ZEMEL, R. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems 29* (2016).

[58] MADDERN, W., PASCOE, G., LINEGAR, C., AND NEWMAN, P. 1 year, 1000 km: The Oxford RobotCar dataset. *The International Journal of Robotics Research 36*, 1 (2017), 3–15.

[59] MANNA, A., KUNDU, R., KAPLUN, D., SINITCA, A., AND SARKAR, R. A fuzzy rank-based ensemble of cnn models for classification of cervical cytology. *Scientific Reports 11*, 1 (2021), 14538.

[60] MANOCHA, A., AND BHATIA, M. A novel deep fusion strategy for covid-19 prediction using multimodality approach. *Computers and Electrical Engineering 103* (2022), 108274.

[61] MARSHALL, P., CLARKSON, W., SHEMMER, O., BISWAS, R., DE VAL-BORRO, M., RHO, J., JONES, L., ANGUITA, T., RIDGWAY, S., BIANCO, F., IVEZIC, Z., LOCHNER, M., MEYERS, J., VIVAS, K., GRAHAM, M., CLAVER, C., DIGEL, S., KASLIWAL, V., MCGEHEE, P. M., GAWISER, E., BELLM, E., WALKOWICZ, L., OLSEN, K., YOACHIM, P., BELL, K., NIDEVER, D., LUND, M., CONNOLLY, A., ARCAVI, I., AND AWAN, H. LSST Science Collaborations Observing Strategy White Paper: "Science-driven Optimization of the LSST Observing Strategy", Aug. 2017.

[62] MARSHALL, P. J., HOGG, D. W., MOUSTAKAS, L. A., FASSNACHT, C. D., BRADAČ, M., SCHRABBACK, T., AND BLANDFORD, R. D. Automated detection of galaxy-scale gravitational lenses in high-resolution imaging data. *The Astrophysical Journal 694*, 2 (2009), 924.

[63] MILANI, F., PINCIROLI VAGO, N. O., AND FRATERNALI, P. Proposals generation for weakly supervised object detection in artwork images. *Journal of Imaging 8*, 8 (2022), 215.

[64] MORGAN, R., NORD, B., BECHTOL, K., GONZÁLEZ, S., BUCKLEY-GEER, E., MÖLLER, A., PARK, J., KIM, A., BIRRER, S., AGUENA, M., ET AL. DeepZipper: A novel deep-learning architecture for lensed supernovae identification. *The Astrophysical Journal 927*, 1 (2022), 109.

[65] MORGAN, R., NORD, B., BECHTOL, K., MÖLLER, A., HARTLEY, W., BIRRER, S., GONZÁLEZ, S., MARTINEZ, M., GRUENDL, R., BUCKLEY-GEER, E., ET AL. Deepzipper ii: Searching for lensed supernovae in dark energy survey data with deep learning. *arXiv preprint arXiv:2204.05924* (2022).

[66] MORGAN, R., NORD, B., BIRRER, S., LIN, J. Y.-Y., AND POH, J. deeplenstronomy: A dataset simulation package for strong gravitational lensing. *arXiv preprint arXiv:2102.02830* (2021).

[67] NISHIMORI, M., KIUCHI, K., NISHIMURA, K., KUSANO, K., YOSHIDA, A., ADACHI, K., HIRAYAMA, Y., MIYAZAKI, Y., FUJIWARA, R., SOMMER, P., ET AL. Accessory pathway analysis using a multimodal deep learning model. *Scientific Reports 11*, 1 (2021), 8045.

[68] OGURI, M. Strong gravitational lensing of explosive transients. *Reports on Progress in Physics 82*, 12 (2019), 126901.

[69] OORD, A., LI, Y., BABUSCHKIN, I., SIMONYAN, K., VINYALS, O., KAVUKCUOGLU, K., DRIESSCHE, G., LOCKHART, E., COBO, L., STIMBERG, F., ET AL. Parallel wavenet: Fast high-fidelity speech synthesis. In *International conference on machine learning* (2018), PMLR, pp. 3918–3926.

[70] ORAMAS, S., BARBIERI, F., NIETO CABALLERO, O., AND SERRA, X. Multimodal deep learning for music genre classification. *Transactions of the International Society for Music Information Retrieval. 2018; 1 (1): 4-21.* (2018).

[71] PAN, J., FANG, W., ZHANG, Z., CHEN, B., ZHANG, Z., AND WANG, S. Multimodal emotion recognition based on facial expressions, speech, and eeg. *IEEE Open Journal of Engineering in Medicine and Biology* (2023).

[72] PARK, J. W. Strongly-lensed quasar selection based on both multi-band tabular data. Project report of the "CS230 Deep Learning" (2018 edition) course at Stanford.

[73] PARK, J. W., VILLAR, A., LI, Y., JIANG, Y.-F., HO, S., LIN, J. Y.-Y., MARSHALL, P. J., AND ROODMAN, A. Inferring black hole properties from astronomical multivariate time series with bayesian attentive neural processes. *arXiv preprint arXiv:2106.01450* (2021).

[74] PEARSON, J., PENNOCK, C., AND ROBINSON, T. Auto-detection of strong gravitational lenses using convolutional neural networks. *Emergent Scientist 2* (2018), 1.

[75] PELLETIER, C., WEBB, G. I., AND PETITJEAN, F. Temporal convolutional neural network for the classification of satellite image time series. *Remote Sensing 11*, 5 (2019), 523.

[76] PETRILLO, C., TORTORA, C., CHATTERJEE, S., VERNARDOS, G., KOOPMANS, L., VERDOES KLEIJN, G., NAPOLITANO, N. R., COVONE, G., KELVIN, L., AND HOPKINS, A. Testing convolutional neural networks for finding strong gravitational lenses in kids. *Monthly Notices of the Royal Astronomical Society 482*, 1 (2019), 807–820.

[77] PETRILLO, C., TORTORA, C., CHATTERJEE, S., VERNARDOS, G., KOOPMANS, L., VERDOES KLEIJN, G., NAPOLITANO, N. R., COVONE, G., SCHNEIDER, P., GRADO, A., ET AL. Finding strong gravitational lenses in the kilo degree survey with convolutional neural networks. *Monthly Notices of the Royal Astronomical Society 472*, 1 (2017), 1129–1150.

[78] PETRILLO, C., TORTORA, C., VERNARDOS, G., KOOPMANS, L., VERDOES KLEIJN, G., BILICKI, M., NAPOLITANO, N. R., CHATTERJEE, S., COVONE, G., DVORNIK, A., ET AL. Links: discovering galaxy-scale strong lenses in the kilo-degree survey using convolutional neural networks. *Monthly Notices of the Royal Astronomical Society 484*, 3 (2019), 3879–3896.

[79] PINCIROLI VAGO, N. O., MILANI, F., FRATERNALI, P., AND DA SILVA TORRES, R. Comparing cam algorithms for the identification of salient image features in iconography artwork analysis. *Journal of Imaging 7*, 7 (2021), 106.

[80] POURRAHMANI, M., NAYYERI, H., AND COORAY, A. LensFlow: A convolutional neural network in search of strong gravitational lenses. *The Astrophysical Journal 856*, 1 (mar 2018), 68.

[81] POUYANFAR, S., TAO, Y., TIAN, H., CHEN, S.-C., AND SHYU, M.-L. Multimodal deep learning based on multiple correspondence analysis for disaster management. *World Wide Web 22* (2019), 1893–1911.

[82] QUIDER, A. M., PETTINI, M., SHAPLEY, A. E., AND STEIDEL, C. C. The ultraviolet spectrum of the gravitationally lensed galaxy 'the Cosmic Horseshoe': a close-up of a star-forming galaxy at z 2. *Monthly Notices of the Royal Astronomical Society 398*, 3 (2009), 1263–1278.

[83] RAHMAN, M., ABEDIN, T., PROTTOY, K. S., MOSHRUBA, A., SIDDIQUI, F. H., ET AL. Semantically Sensible Video Captioning (SSVC). *arXiv preprint arXiv:2009.07335* (2020).

[84] RAMACHANDRAM, D., AND TAYLOR, G. W. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine 34*, 6 (2017), 96–108.

[85] SAITO, N., OGATA, T., FUNABASHI, S., MORI, H., AND SUGANO, S. How to select and use tools?: Active perception of target objects using multimodal deep learning. *IEEE Robotics and Automation Letters 6*, 2 (2021), 2517–2524.

[86] SALEKIN, M. S., ZAMZMI, G., GOLDGOF, D., KASTURI, R., HO, T., AND SUN, Y. Multimodal spatio-temporal deep learning approach for neonatal postoperative pain assessment. *Computers in biology and medicine 129* (2021), 104150.

[87] SAVARY, E., ROJAS, K., MAUS, M., CLÉMENT, B., COURBIN, F., GAVAZZI, R., CHAN, J., LEMON, C., VERNARDOS, G., CAÑAMERAS, R., ET AL. Strong lensing in unions: Toward a pipeline from discovery to modeling. *Astronomy & Astrophysics 666*, ARTICLE (2022), A1.

[88] SAVARY, E. M. C. Teaching machines how to find strongly lensed galaxies in cosmological sky surveys. Tech. rep., EPFL, 2022.

[89] SCHAEFER, C., GEIGER, M., KUNTZER, T., AND KNEIB, J.-P. Deep convolutional neural networks as strong gravitational lens detectors. *Astronomy & Astrophysics 611* (2018), A2.

[90] SCHNEIDER, D., GUNN, J., AND HOESSEL, J. CCD photometry of Abell clusters. I-magnitudes and redshifts for 84 brightest cluster galaxies. *The Astrophysical Journal 264* (1983), 337–355.

[91] SHAIKH, R., BANERJEE, P., PAUL, S., AND SARKAR, T. Strong gravitational lensing by wormholes. *Journal of Cosmology and Astroparticle Physics 2019*, 07 (2019), 028.

[92] SHEN, J., PANG, R., WEISS, R. J., SCHUSTER, M., JAITLY, N., YANG, Z., CHEN, Z., ZHANG, Y., WANG, Y., SKERRV-RYAN, R., ET AL. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (2018), IEEE, pp. 4779–4783.

[93] SHIMIZU, H., AND NAKAYAMA, K. I. Artificial intelligence in oncology. *Cancer science 111*, 5 (2020), 1452–1460.

[94] STAHLSCHMIDT, S. R., ULFENBORG, B., AND SYNNERGREN, J. Multimodal deep learning for biomedical data fusion: a review. *Briefings in Bioinformatics 23*, 2 (01 2022). bbab569.

[95] STALDER, B., REIL, K., CLAVER, C., LIANG, M., TSAI, T. W., LANGE, T., HAUPT, J., WIECHA, O., LOPEZ, M., POCZULP, G., ET AL. Rubin commissioning camera: integration, functional testing, and lab performance. In *Ground-based and Airborne Instrumentation for Astronomy VIII* (2020), vol. 11447, SPIE, pp. 86–98.

[96] STERN, D., DJORGOVSKI, S., KRONE-MARTINS, A., SLUSE, D., DELCHAMBRE, L., DUCOURANT, C., TEIXEIRA, R., SURDEJ, J., BOEHM, C., DEN BROK, J., ET AL. Gaia gral: Gaia dr2 gravitational lens systems. vi. spectroscopic confirmation and modeling of quadruply imaged lensed quasars. *The Astrophysical Journal 921*, 1 (2021), 42.

[97] SUEL, E., BHATT, S., BRAUER, M., FLAXMAN, S., AND EZZATI, M. Multimodal deep learning from satellite and street-level imagery for measuring income, overcrowding, and environmental deprivation in urban areas. *Remote Sensing of Environment 257* (2021), 112339.

[98] SUMMAIRA, J., LI, X., SHOIB, A. M., LI, S., AND ABDUL, J. Recent advances and trends in multimodal deep learning: A review. *arXiv preprint arXiv:2105.11087* (2021).

[99] TANG, Z., SHI, Y., WANG, D., FENG, Y., AND ZHANG, S. Memory visualization for gated recurrent neural networks in speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2017), IEEE, pp. 2736–2740.

[100] TEIMOORINIA, H., TOYONAGA, R. D., FABBRO, S., AND BOTTRELL, C. Comparison of multi-class and binary classification machine learning models in identifying strong gravitational lenses. *Publications of the Astronomical Society of the Pacific 132*, 1010 (2020), 044501.

[101] TORRES, R. N., FRATERNALI, P., AND ROMERO, J. ODIN: An object detection and instance segmentation diagnosis framework. In *European Conference on Computer Vision* (2020), Springer, pp. 19–31.

[102] TORRES, R. N., MILANI, F., AND FRATERNALI, P. ODIN: Pluggable meta-annotations and metrics for the diagnosis of classification and localization. In *International Conference on Machine Learning, Optimization, and Data Science* (2021), Springer, pp. 383–398.

[103] TREU, T. Strong lensing by galaxies. *Annual Review of Astronomy and Astrophysics 48*, 1 (2010), 87–125.

[104] TYSON, J. A., VALDES, F., AND WENK, R. Detection of systematic gravitational lens galaxy image alignments-mapping dark matter in galaxy clusters. *The Astrophysical Journal 349* (1990), L1–L4.

[105] VAKULIK, V., SCHILD, R., DUDINOV, V., NURITDINOV, S., TSVETKOVA, V., BURKHONOV, O., AND AKHUNOV, T. Observational determination of the time delays in gravitational lens system Q2237+ 0305. *Astronomy & Astrophysics 447*, 3 (2006), 905–913.

[106] VÁSQUEZ-CORREA, J. C., ARIAS-VERGARA, T., OROZCO-ARROYAVE, J. R., ESKOFIER, B., KLUCKEN, J., AND NÖTH, E. Multimodal assessment of parkinson's disease: a deep learning approach. *IEEE journal of biomedical and health informatics 23*, 4 (2018), 1618–1630.

[107] VELIOGLU, R., AND ROSE, J. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *arXiv preprint arXiv:2012.12975* (2020).

[108] VICEDOMINI, M., BRESCIA, M., CAVUOTI, S., RICCIO, G., AND LONGO, G. *Statistical Characterization and Classification of Astronomical Transients with Machine Learning in the era of the Vera C. Rubin Observatory*. Springer International Publishing, Cham, 2021, pp. 81–113.

[109] WEI, R., MI, L., HU, Y., AND CHEN, Z. Exploiting the local temporal information for video captioning. *Journal of Visual Communication and Image Representation 67* (2020), 102751.

[110] WEI, Y., WANG, L., CAO, H., SHAO, M., AND WU, C. Multi-attention generative adversarial network for image captioning. *Neurocomputing 387* (2020), 91–99.

[111] WITTEN, I. H., AND FRANK, E. Data mining: practical machine learning tools and techniques with java implementations. *Acm Sigmod Record 31*, 1 (2002), 76–77.

[112] WOJTAK, R., HJORTH, J., AND GALL, C. Magnified or multiply imaged? – Search strategies for gravitationally lensed supernovae in wide-field surveys. *Monthly Notices of the Royal Astronomical Society 487*, 3 (06 2019), 3342–3355.

[113] XU, T., ZHANG, H., HUANG, X., ZHANG, S., AND METAXAS, D. N. Multimodal deep learning for cervical dysplasia diagnosis. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016* (Cham, 2016), S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds., Springer International Publishing, pp. 115–123.

[114] YANG, S., YU, X., AND ZHOU, Y. LSTM and GRU neural network performance comparison study: Taking yelp review dataset as an example. In *2020 International Workshop on Electronic Communication and Artificial Intelligence (IWECAI)* (2020), pp. 98–101.

[115] ZANGRANDO, N. The ODIN framework, a tool for image classification diagnosis, 2021.

[116] ZHAO, B., LU, H., CHEN, S., LIU, J., AND WU, D. Convolutional neural networks for time series classification. *Journal of Systems Engineering and Electronics 28*, 1 (2017), 162–169.

[117] ZHAO, K., GAO, Q., HAO, S., SUN, J., AND ZHOU, L. Credible remote sensing scene classification using evidential fusion on aerial-ground dual-view images. *arXiv preprint arXiv:2301.00622* (2023).

[118] ZWICKY, F. On the probability of detecting nebulae which act as gravitational lenses. *Physical Review 51*, 8 (1937), 679.