UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS FÍSICAS

Máster en Física Teórica

TRABAJO FIN DE MÁSTER

Aprendizaje activo profundo aplicado a ondas gravitacionales

Deep Active Learning applied to gravitational waves

Johanna Stammer Goldaracena

Director/es:

Miguel Cárdenas Montes

Carlos Delgado Méndez

Curso Académico 2025-26

UNIVERSIDAD
COMPLUTENSE
MADRID

Declaración responsable sobre autoría y uso ético de herramientas de Inteligencia Artificial (IA)

Yo, Johanna Stammer Goldaracena,

con DNI/NIE/PASAPORTE: 72854214N,

declaro de manera responsable que el presente Trabajo Fin de Máster (TFM) titulado:

Deep Active Learning applied to gravitational waves

es el resultado de mi trabajo intelectual personal y creativo, y ha sido elaborado de acuerdo con los principios éticos y las normas de integridad vigentes en la comunidad académica y, más específicamente, en la Universidad Complutense de Madrid.

Soy, pues, la persona autora del material aquí incluido y, cuando no ha sido así y he tomado el material de otra fuente, lo he citado o bien he declarado su procedencia de forma clara —incluidas, en su caso, herramientas de inteligencia artificial—. Las ideas y aportaciones principales incluidas en este trabajo, y que acreditan la adquisición de competencias, son mías y no proceden de otras fuentes o han sido reescritas usando material de otras fuentes.

Asimismo, aseguro que los datos y recursos utilizados son legítimos, verificables y han sido obtenidos de fuentes confiables y autorizadas. Además, he tomado medidas para garantizar la confidencialidad y privacidad de los datos utilizados, evitando cualquier tipo de sesgo o discriminación injusta en el tratamiento de la información.

En Madrid, a 8 de septiembre de 2025.

Firmado por STAMMER GOLDARACENA JOHANNA - ***5421**
el día 08/09/2025 con un certificado emitido por AC
FNMT Usuarios

FIRMA

Deep Active Learning applied to gravitational waves

Johanna Stammer Goldaracena

Supervisors: Miguel Cárdenas Montes and Carlos Delgado Méndez
CIEMAT, Av. Complutense 40, 28040 Madrid, Spain

The detection of gravitational waves (GW) has opened a new window to progress in our understanding of astrophysical events and objects. The instruments used for direct detection (interferometers) require high sensitivity due to the tiny signals these GWs generate. In addition, proper characterisation of the detector is crucial for identifying noise sources and enhancing the performance. This thesis explores the implementation of Deep Active Learning (DAL) to identify and characterise short duration transient noise in the GW signal stream. We employ a convolutional neural network (CNN) combined with the DBSCAN clustering algorithm to classify glitches detected by interferometers. Moreover, an Attention Layer is implemented to highlight the relevant areas of the images for the final classification. Our approach recognises patterns similar to previously identified signals and detects anomalous ones that could correspond to previously unseen phenomena.

CONTENTS

I. INTRODUCTION

The concept of gravitational waves is a consequence of Albert Einstein's Theory of General Relativity [1]. In this theory, gravity is described as the deformation of the spacetime due to the presence of matter (or energy). Its motion can generate small perturbations in this geometry, known as gravitational waves, which propagate across the cosmos. As these waves traverse space, their influence is experienced by other objects.

LIGO interferometers detected GWs for the first time in 2015 [2]. When a gravitational wave passes through Earth it distorts spacetime, altering the time required for light to travel a given distance within the detector. Therefore, these detectors use lasers to measure minute deviations of the interferometer geometry as a function of time. The typical deviations in length that need to be measured are of one part in 10^{-21}, requiring sophisticated mechanisms and analysis methods that minimise noises which reduce the sensibility and detection efficiency.

One important family of sources of noise in the collected data is the presence of short-duration fluctuations known as glitches. These degrade the quality of the data and, therefore, their identification and characterisation is essential for achieving optimal results. To complete this task, Deep Active Learning (DAL) [3] techniques have been employed, giving highly effective outcomes [4, 5].

In particular, the combination of a convolutional neural network (CNN) and a DBSCAN algorithm enables the classification of glitches while also identifying previously unknown signals. Since the unidentified signals must be manually examined by scientists to assess their category, this procedure allows minimising the need of treating with spurious cases. Also, the implementation of a Spatial Attention layer in this process might help highlighting the main characteristics for each label so that the identification is more straightforward.

For this study, we use the publicly available *Gravity Spy Training Set* from Kaggle [6], a collection of labelled spectrogram QT images recorded by the LIGO and Virgo interferometers. It contains around 8000 event images, each representing waveforms across 4 time intervals from 0.5s to 4s. One label corresponds to signals from merging black holes, while the other 21 classes represent different detector glitches. The first training is done on a subset of about 250 images from 5 labels: *1080Lines*, *1400Ripples*, *Air Compressor*, *Scratchy* and *Paired Doves*. Later, an additional label, *None of the Above*, is introduced as unknown signals and used throughout the study, along with the remaining images of the previously known labels.

This final procedure constitutes the primary focus of this thesis. We first present the relevant theoretical aspects of the key topics explored in this study in Section II and Section III, as well as the characteristics and processing of the employed data, with emphasis on the information flow throughout the procedure. Finally, the obtained results are shown in Section IV, followed by the conclusions achieved from this research in Section V.

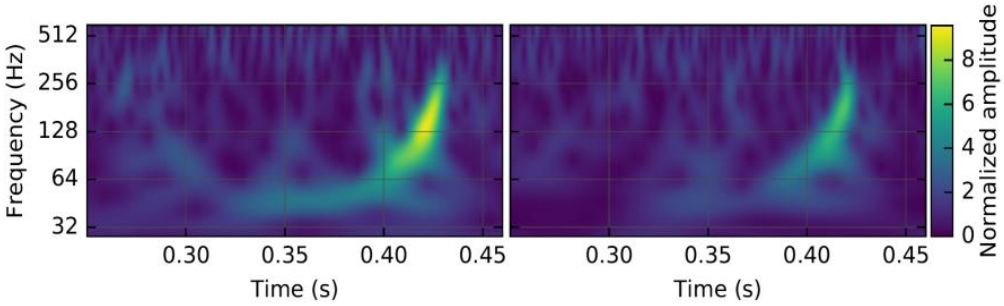

FIG. 1. The gravitational-wave event GW150914 observed by the LIGO Hanford (H1, left) and Livingston (L1, right) detectors [2]. These images are obtained from the time series measured by the detectors after applying a Q-transform. This is a time-frequency representation similar to the Fourier transform with adaptive resolution, where different frequency ranges are represented in different time intervals. The signal can also be decomposed into frequency modes. The horizontal axis represents the time of the signal, while the vertical axis shows the frequency (Hz). The colour encodes the intensity of the signal. Thus, the initially detected strain time series is shown as the evolution of the intensity of frequency modes, making the characteristic *chirp* form of a binary system merger visible.

II. GRAVITATIONAL WAVES PHYSICS

A. Description and sources

As already said, GWs are distortions in spacetime caused by the movement of massive objects. In general, any acceleration of a mass with non vanishing quadrupolar moment generates this radiation, though not all cases produce detectable effects. The sources can be classified based on their motion characteristics as it follows [7]:

- Burst sources: Typically associated with core-collapse supernovae, which are highly unpredictable and their evolution is complex. The resulting waveforms depend on the specific dynamics of the individual object.
- Continuous sources: Primarily emitted by spinning neutron stars, these waves arise from rotational motion. Even if they experience energy loss due to gravitational radiation (leading to a gradual decrease in amplitude over time) the change is negligible on short timescales, allowing them to be considered effectively continuous.
- Stochastic sources: Formed by the cumulative effect of several small perturbations that cannot be analysed individually. These may originate from the early universe or by the combined influence of multiple astronomical systems. The detected signal resembles white noise.
- Inspiral sources: Generated during the coalescence of compact binary systems. As the objects spiral closer together, both the frequency and amplitude of the emitted waves increase, producing a characteristic signal known as a **chirp**.

This work focuses on inspiral sources, as they are the primary targets of current interferometers (LIGO, Virgo, GEO300 and Kagra) and because our methodology has been specifically developed for the classification of glitches in their signals.

Inspiral sources. They are composed of two compact objects, which accordingly to its kind are classified as binary neutron stars (BNS), binary black holes (BBH) or neutron star-black hole (NSBH). Each configuration produces a distinct signal in terms of amplitude, frequency and time to coalescence, although the general waveform remains consistent.

Mathematically, following the derivations in [8] (section 4.1), a binary source with masses m_1 and m_2 ($m_1 + m_2 = m_{\text{total}}$) and separation R can be modelled as an equivalent single-particle system (in the centre-of-mass frame) with a reduced mass $\mu = \frac{m_1 m_2}{m_1+m_2}$ moving in a circular orbit of radius R. Following the calculations, the evolution of the emitted GWs' frequency can be derived:

$$f_{\text{gw}}(\tau) = \frac{1}{\pi}\left(\frac{5}{256}\frac{1}{\tau}\right)^{3/8}\left(\frac{GM_c}{c^3}\right)^{-5/8}, \qquad (1)$$

where $\tau = t_{\text{coal}} - t$ (t_{coal} being the time of coalescence) is the time to coalescence and $M_c = \mu^{3/5} m_{\text{total}}^{2/5}$ is the *chirp mass*. With this expression we are able to understand the change in f_{gw} through time. So, when time t increases, $\tau = t_{\text{coal}} - t$ decreases and f_{gw} increases, producing the characteristic chirping form of the signal, as it is illustrated in Fig. 1.

B. Detection: Interferometers

Gravitational waves coming from binaries of compact objects can be detected by interferometers, which are Michelson-Morley interferometers enhanced with Fabry-Perot cavities to improve sensitivity in the typical frequency range of emission of these objects.

As illustrated in Fig. 2, the detector splits a laser beam into two paths using a beam splitter. These rays travel

along the interferometer arms until they reach mirrors at the ends, where they are reflected back. Upon recombination, the total signal is received by the system. When a GW passes through, it alters the arm lengths (due to spacetime distortion), leading to changes in the beams' travel time. So the detection is performed by measuring the phase shift between the split beams through interference.

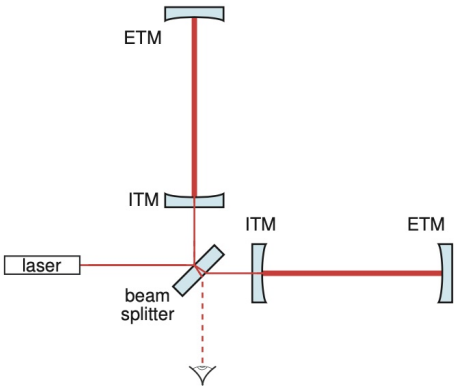

FIG. 2. Scheme of a Michelson-Morley interferometer with Fabry-Perot cavities implemented [9].

A measurable property of light is the power of the radiation $P \propto |E_{\rm tot}|^2$; the detected power is expressed as:

$$P = \frac{P_0}{2}[1 - \cos \Delta\phi_{\rm Mich}(t)]. \tag{2}$$

To maximise the variable P, $\Delta\phi_{\rm Mich}$ (a phase shift related to the geometry of the system) should be as large as possible. It can be expressed as $\Delta\phi_{\rm Mich} = 2h(t - \frac{L}{c})k_L L$, where $h = \frac{\delta L}{L}$ is the amplitude of the GW strain, k_L is the wave-vector of the laser and L represents the arms' length of the interferometer. Because of the small scale of GWs ($h \sim 10^{-21}$ and $f \sim 10^2$ Hz for typical detected compact binaries [8]) these arms should be around 700 km long, which is impractical on Earth dimensions. Employing Fabry-Perot cavities solves the issue. They consist of two highly reflective mirrors with a transmissive substrate designed to trap the light beam inside for the largest duration possible, effectively increasing the optical path length of the laser.

The improvement obtained by implementing the cavity thanks to its impact on the storage time (with a high reflection coefficient, light is easily reflected and stays inside for a longer time) is described by the finesse $\mathcal{F}$:

$$\mathcal{F} = \frac{\pi\sqrt{r_1 r_2}}{1 - r_1 r_2}, \tag{3}$$

where r_1, r_2 are the reflection coefficients of both mirrors composing the cavity ($0 < r_1, r_2 < 1$). The phase shift measured in the interferometer incorporating Fabry-Perot cavities is:

$$\Delta\phi_{\rm FP}(t) = \frac{2\mathcal{F}}{\pi}\Delta\phi_{\rm Mich}(t). \tag{4}$$

The phase shift of the interferometer can be increased by a factor of $\frac{2\mathcal{F}}{\pi}$, as large as possible based on the cavity mirrors reflectivity. This makes it possible to detect GWs in terrestrial dimensions by combining both systems.

1. *Noises*

Interferometers have complex machinery with numerous subsystems that must operate in sync to detect signals. However, this sophisticated network also introduces many noise sources due to the physical limitations of the design, requiring a wide range of attenuation methods. In general, most detector noises are stationary and frequency-dependent, meaning that for distinct frequency ranges different sources will be dominant. The main contributions of noise are [10]:

- Seismic noise: originated by ground vibrations. It dominates around $\sim$ 1 Hz and it is reduced by suspending the mirrors with multiple pendulums, which decreases the noise by a factor of $\propto f^{-2}$ with each additional pendulum.
- Thermal noise: generated by dissipation in the mirrors and their suspension chains. It is the most important effect around $\sim 1 - 100$ Hz.
- Shot noise: caused by fluctuations in the number of photons from the laser hitting the detector in a given time interval (since photons arrive in discrete "packets" rather than continuously). It dominates above 100 Hz.

There are many others, such as gravity gradient noise, radiation pressure, elastic noise... These also contribute and must be mitigated using different techniques in the instrumentation. Furthermore, there is another type

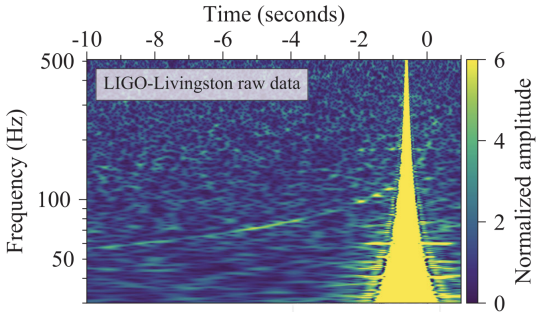

FIG. 3. Time-frequency representation of the raw LIGO-Livingston data for GW170817 event. A glitch is visible 1 s before the coalescence time of the GW signal [11].

of noise that compromises the data: transient, short-duration disturbances produced by the interaction between detector subsystems or by the surroundings that may mimic GW or overlap them. These are called **glitches** (an example is shown in Fig. 3). Their presence

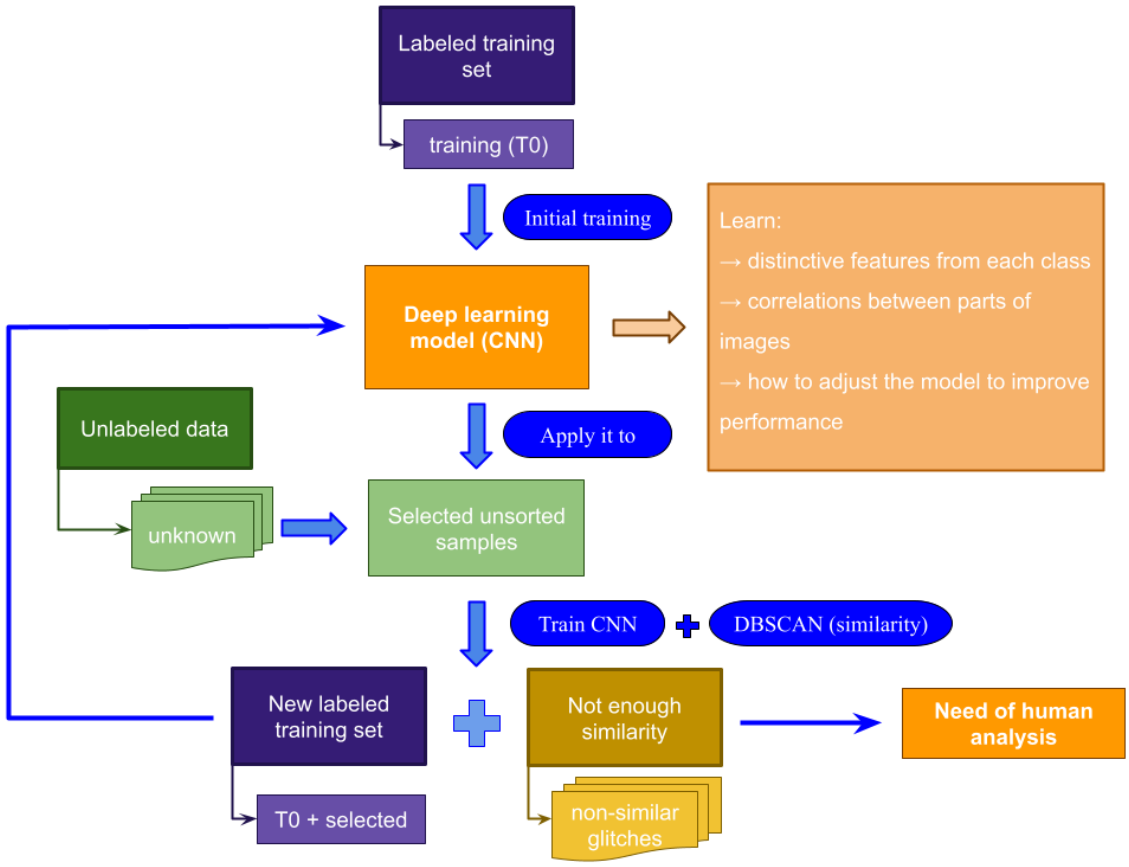

FIG. 4. Deep Active Learning scheme. The main figures and information flux are represented.

can bias estimates of GWs properties: the median rate of glitches in the LIGO and Virgo interferometers through the third observing runs (O3) were above 1 per minute. So, at this rate, the probability of a glitch to compromise a signal from a BBH is $\sim 10\,\%$ and it is almost certain that a glitch will overlap a BNS signal [12].

If we take into account that GW signals are very weak compared to all the remaining noises, this issue needs of a proper solution. In the case of glitches, there is no direct way of avoiding their presence through improvements in instrumentation, so it is important to be able to identify and classify them.

To complete this task, many promising options have been proposed; but the one of our interest is the classification of glitches based on machine learning. This method eases the human workload by quickly recognising and labelling signals, saving hours of manual work, as the algorithm effectively understands and attenuates known glitches, allowing to focus on unexamined or anomalous data that might lead to the discovery of new unwanted noise patterns.

III. DEEP ACTIVE LEARNING

The main objective of machine learning is to derive general patterns from a limited data set, in order to obtain a classification of the analysed objects. The idea is to generate a mapping process from the inputs to the distinct labels that the mechanism is able to distinguish. Deep Active Learning (DAL) is an iterative training paradigm that enables to do so by combining the data efficiency of Active Learning with the representational power of Deep Learning models [3].

Concerning its application, DAL is particularly suited for many scientific domains where labelling processes require expert validation. This is the case of gravitational wave detection, in which manual annotation is not only highly time-consuming but also difficult to adapt when the data includes new unknown glitch types. By focusing the labelling effort on the most informative samples, DAL allows human labour to prioritise the anomalous glitch images that emerge in the analysis.

The general workflow of DAL is illustrated in Fig. 4. The whole dataset consists of two main subsets: (1) a group of known data (images previously annotated by experts with known labels) and (2) an additional pool of unlabelled images, which may contain samples from the

previous classes or entirely new/unseen ones.

The process starts with a small, labelled subset of the known data to form the initial training set T0. From here, the workflow proceeds as follows:

1. A CNN is initially trained using the labelled data T0. The model learns to distinguish between the classes present in the training images.

2. The trained CNN is then applied to an unseen dataset, which might have examples of classes already present in the training set, but also some samples of new, previously unseen classes.

3. Next, the whole combined dataset, containing both T0 and the newly added images, is passed to the DBSCAN clustering algorithm. DBSCAN groups the data into clusters, unifying samples that correspond to known classes and identifying outliers or "anomalous" images, those not sufficiently similar to any known category.

4. The unseen examples clustered by DBSCAN as belonging to previously known labels are incorporated into the training set T0. In the next iteration, the CNN is retrained using the expanded set, now including both the original T0 and the newly incorporated examples. Meanwhile, the "anomalous" cases are kept aside for human inspection.

5. A new iteration of the process with the updated dataset begins.

As shown in Fig. 4 and just explained, the success of the DAL framework in this context relies on two main components: **1)** A deep neural network, implemented as a Convolutional Neural Network (CNN) and responsible for learning the main features from input images. **2)** A clustering algorithm, DBSCAN, used to identify similar samples and detect outliers. This allows the selection of anomalous unlabelled signals for expert review.

A. CNN

Convolutional Neural Networks (CNN) are a technique specifically designed to work with structured data such as images, videos or audio spectrograms. Their goal is to detect patterns and recognise complex structures hierarchically in the inputs without human intervention. This is done by applying **filters** (or *kernels*), which are small windows that scan the image. For this end, this method's network is composed of multiple layers which combine several techniques:

- Convolutional Layers: they extract local patterns and spatial features by applying the *kernels* all over the input image.

- Pooling Layers: they reduce the spatial dimensions of the input, producing a smaller and more robust representation that lowers computational cost.

- Flattening: it transforms multidimensional objects into 1D vectors.

- Dense (Fully Connected) Layer: it takes the flattened vector and learns high-level combinations of the extracted features. Each neuron is connected to all values from the previous layer and the model makes predictions based on the most relevant features for each label.

Some additional mechanisms can be integrated into the CNNs to enhance the classifier performance:

a. ***Early stopping.*** It is a strategy commonly used to prevent overfitting. The training is halted when the model's performance stops improving on the validation set, preventing it from memorising the training samples too precisely and becoming less effective on unseen data. The stopping is performed after a given patience parameter, 2 epochs of worsening in our case.

b. ***Spatial Attention.*** In our architecture this layer is placed between the input and the first convolutional block. This mechanism helps the model focus on the main features/patterns or the most relevant parts of the images by assigning higher weights to important regions and suppressing less informative ones. It allows the network to pay attention to certain areas of the spatial inputs. In other words, it teaches the CNN model where to look in the image.

B. DBSCAN

DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) is an **unsupervised** clustering algorithm that groups similar inputs into clusters without requiring prior knowledge of the number of categories. It is an appropriate approach for datasets with irregular cluster shapes and noise, as it can identify core structures while also marking unrelated points as *outliers*. This makes it particularly useful in gravitational wave detection, where different glitch types may appear in uneven amounts and unknown signals must be identified and treated separately, as explained through Fig. 4.

The idea behind this method is to map each image to a point in a high-dimensional Euclidean space, where the dimensionality depends on the size of the images (the number of pixels) and the number of channels (three in the case of RGB images). Through this representation, the algorithm studies how far apart these points are from one another in the feature space. DBSCAN then evaluates the density of points in a given local neighbourhood using two main parameters:

- ε (epsilon): the maximum distance for which two points are considered neighbours.

- `minPts`: the minimum number of neighbours a point must have within distance ε to be considered a *core point*.

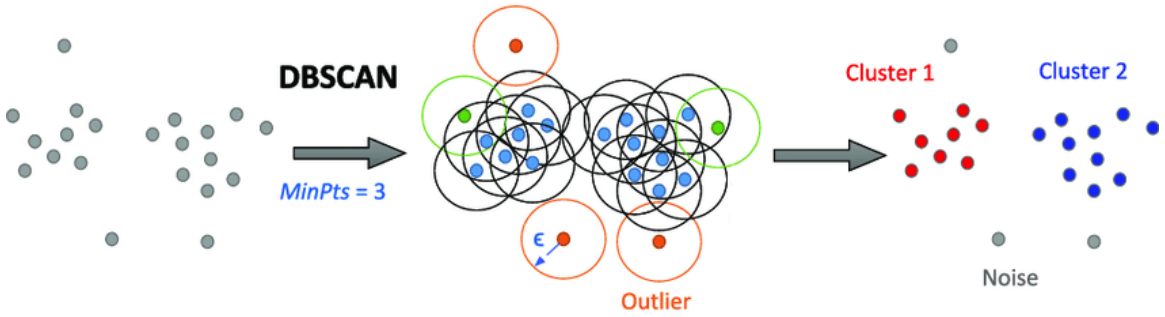

FIG. 5. Clusters obtained (right) through DBSCAN after evaluating ε and *MinPts* parameters [13].

In practice, images are grouped into clusters only if they have a sufficient amount of similar neighbours within the ε radius. Once a core point is found, all its directly density-reachable points are added to the cluster. If an input does not have enough nearby neighbours on its proximities (fewer neighbours than `minPts`), then it will be classified as an *outlier*. This methodology is visually represented in Fig. 5. However, it should be noted that it is highly sensitive to the choice of its parameters, which becomes more pronounced in high-dimensional spaces. In such cases, the Euclidean distance used by DBSCAN becomes less effective at distinguishing points, making it harder to find clusters of similar points.

IV. RESULTS

Data processing prior to analysis. The employed dataset [6] was initially divided into 3 subsets: training, validation and test. To make the images suitable for our DAL model, they are reorganized using an algorithm designed to select both the types of glitches and the number of samples for each case. More specifically, the user specifies the glitch type names and the desired amounts through a text file. In this way, the training set T0 described in Fig. 4 is explicitly defined according to the needs of the analysis. All samples are chosen randomly from the initially available images, ensuring the performance of the DAL mechanism is not biased by any specific data subset. After that, the components of the code related to the CNN and DBSCAN algorithms are incorporated and adapted to the requirements of this project.

In this work, the initially studied dataset includes the following classes and samples sizes: **1) Training set**: it contains 5 distinct labels, each of them with $80-120$ images. The selected glitches are: *1080Lines*, *1400Ripples*, *Air Compressor*, *Paired Doves* and *Scratchy*. **2) Target set**: this consists of an additional label named *None of the Above*, which represents the *unseen* data that was already classified as unknown in the original dataset. This category includes around 60 examples that could not be associated with any of the known glitch types, since they are neither similar to any of the studied labels nor show consistency among themselves.

This reduced dataset has been intentionally chosen for pedagogical purposes: to explain the procedure and to enable a clear evaluation and understanding of all the mechanisms of DAL. Once its correct functioning has been validated in this controlled setting, further evaluation on a larger scale is foreseen, outside the scope of the Master's project, in order to prepare the model for deployment in production.

a. ***Performance of the Supervised Learning component.*** A CNN has been built as explained in Section III, incorporating a Spatial Attention layer. The effect of this mechanism on our dataset is illustrated in Fig. 6. For example, it can be observed that the model without Spatial Attention does not classify properly this sample corresponding to the *Scratchy* class, while the version including the Spatial Attention layer properly identifies it, improving the "probability" parameter given to the correct/true class by 3.12.

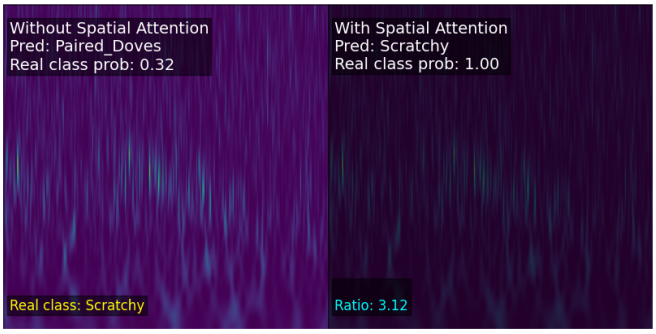

FIG. 6. Comparison between an original image (extracted from the working dataset) and the same sample after the Spatial Attention layer. At the same time, a model without the Spatial Attention layer (left text) and a model with this layer implemented (right text) analyse and classify the image, exposing the "probability" for corresponding to the real class.

The training process uses $\sim$ 250 images (around 50 samples per label) highlighted by the Spatial Attention mechanism and typically takes around $2-10$ minutes for the CNN to learn how to classify the specific implemented glitch types with a final accuracy of $> 0.98\,\%$.

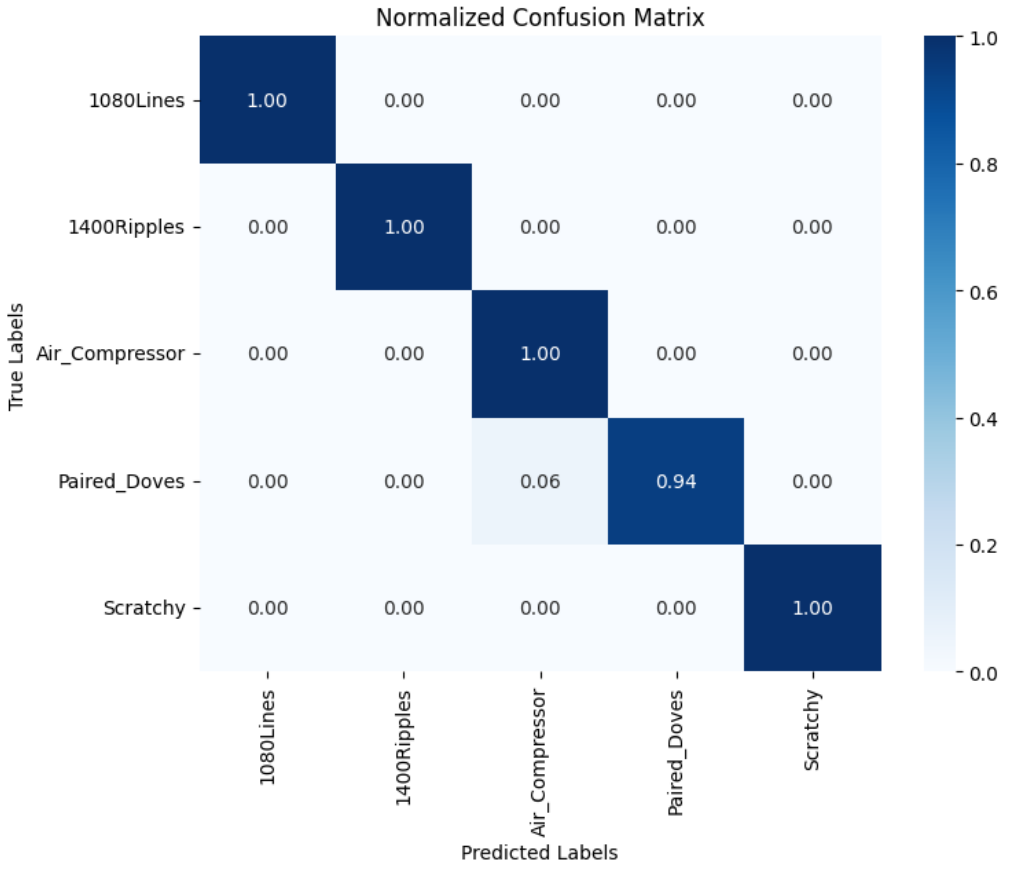

FIG. 7. Normalized confusion matrix for one trained model. The observed results are representative of the general performance and effectiveness of the CNN.

Once trained, the model is evaluated on the validation set that takes around $20 - 30$ objects from each class (~ 120 images in total) that were not used to train the CNN. With this set, the classifier's performance is evaluated through the **confusion matrix** (Fig. 7), which compares the predicted and true labels on the validation dataset. It provides a detailed picture of how well the model distinguishes between the different glitch types, offering insight into possible confusions during training.

As shown in Fig. 7, which corresponds to a representative result obtained with the trained model, the CNN successfully learns the distinguishing features of each class and is able to label (almost) all validation images correctly. This outcome is essential to ensure that the next step, DBSCAN clustering, can have an appropriate functioning and that images are correctly grouped.

Notably, the only misclassified image belongs to the glitch type with the least representation (42 samples) in the training set: *Paired Doves*. This is to be expected, as a lower number of training images for a class makes it harder for the model to capture its general and distinctive features effectively. It could even suggest that the training subset for this label is not comprehensive enough to represent the entire class variability and may require a wider range of examples.

b. ***Performance of the Unsupervised Learning component.*** First, the preprocessing of the images passed through the Spatial Attention layer takes approximately 5 minutes for the 600 employed samples in this second step (56 anomalous and 544 non-anomalous samples). These images are then processed with the DBSCAN algorithm and its clustering performance is analysed by observing the resulting *outliers*, which vary depending on the main parameters (ε and *MinPts*).

To better understand the performance of the method, three target configurations are considered:

1. **All *real* outliers detected**: All *None of the Above* samples are correctly classified as anomalies, though this often leads to many non-anomalous images also being misclassified.

2. **No *false* outliers detected**: This case avoids misclassification of known classes, but it might miss many *real* outliers. As it will be shown later, this ideal case could not be achieved, so the closest attainable case with minimal misclassification was considered instead.

3. **Intermediate case**: Approximately 90 % of *real* outliers are identified. This configuration has been

Category	Total Images	Detected as Anomalies	Anomaly Percentage (%)
Case 1 — $\varepsilon = 0.968$, *MinPts* = 19. 189 anomalies, 4 clusters			
1080Lines	124	6	4.84
1400Ripples	124	0	0.00
Air Compressor	100	10	10.00
None of the Above	56	56	**100.00**
Paired Doves	76	47	61.84
Scratchy	120	70	58.33
Case 2 — $\varepsilon = 4.827$, *MinPts* = 7. 10 anomalies, 1 cluster			
1080Lines	124	0	0.00
1400Ripples	124	0	0.00
Air Compressor	100	0	0.00
None of the Above	56	9	**16.07**
Paired Doves	76	1	1.32
Scratchy	120	0	0.00
Case 3 — $\varepsilon = 1.108$, *MinPts* = 24. 136 anomalies, 5 clusters			
1080Lines	124	11	8.87
1400Ripples	124	0	0.00
Air Compressor	100	17	17.00
None of the Above	56	51	**91.07**
Paired Doves	76	12	15.79
Scratchy	120	45	37.50

TABLE I. Detected anomalies per category for three distinct DBSCAN configurations, each for the different proposed target configurations. In every case, 56 anomalies and 544 non-anomalous samples were used.

chosen to represent a more realistic scenario in a real production environment.

The results for these three approaches are shown in detail in Table I and a high-level summary is provided in Table II. Each scenario uses a distinct DBSCAN configuration, exposing the significant impact of the parameter selection on the final outcome.

Paying attention to Table I, each scenario specifies the employed parameters ε and *minPts*, as well as the total anomalies found and the formed clusters. Apart from that, for each class the number of images employed is specified, and also those detected as anomalies (even if they are not) and their percentage.

On the one hand, **Case 1** corresponds to the scenario where all actual outliers (i.e., all samples in the **None of the Above** category) are correctly identified, achieving a full 100 % (56 out of 56) detection rate. In this configuration, 189 anomalies are found and, therefore, 133 false positives are classified as outliers in addition to the 56 real anomalies. More specifically, 47 samples from *Paired Doves* and 70 from *Scratchy* are labelled as anomalous, which represent 61.84 % and 58.33 % of all their samples, respectively. These percentages are significant and cause a large number of non-anomalous images to be wrongly labelled. The difficulty in classifying appropriately these sets can happen because of a smaller size of its dataset (for example, *Paired Doves* only contains 76 samples, whereas other labels are composed of around $100 - 120$ images), since the model has received less representative information from this category. Moreover, if the category has a high internal variability within its signals (which might happen to the *Scratchy* subset), it becomes more difficult for the clustering algorithm to group them accurately. Overall, Case 1 represents a restrictive strategy that prioritises capturing all *true* outliers, even at the expense of introducing a large volume of data wrongly assigned to the anomalous category and, therefore, increasing the dataset passed to experts for human analysis.

On the other hand, **Case 2** aims to minimise false positives as much as possible. It seeks the configuration with the lowest number of wrongly categorised outliers. As shown in Table I, a case with no *false* outliers could not be achieved. A total of 10 anomalies are found: 9 real ones and 1 false positive corresponding to the *Paired Doves* class. With these results, it can be said that only 16.07 % of the *None of the Above* samples are flagged as outliers. The rest have been wrongly clustered together with other signals and will not be analysed by human experts, leaving key outliers undetected.

Finally, **Case 3** represents a compromise. It does not require 100 % accuracy for the detection of *None of the Above*. In this way, the number of total flagged anomalies

Case	True anomalies detected (%)	False anomalies detected (%)	Data reduction rate (%)
1	100.00	24.45	31.50
2	16.07	0.18	1.67
3	91.07	15.63	22.67

TABLE II. Summary of anomaly detection performance across the three DBSCAN configurations. True anomalies detected refers to the proportion of *None_of_the_Above* images correctly flagged as anomalies (out of 56 samples), false anomalies detected refers to the proportion of non-anomalous images wrongly classified as anomalies (out of 544 samples, being 544 the total non-anomalous images in the analysis) and data reduction rate indicates the proportion of the total dataset flagged as anomalous for further human analysis (out of 600 samples).

is reduced from 189 (Case 1) to 136, at the expense of detecting 91.07 % of the real anomalies instead of 100 %. Nevertheless, this percentage still represents a high level of accuracy: 51 outliers out of 56 are correctly identified from *None of the Above*. The number of false positives is now 85 and, like in Case 1, the class with the highest number of false positives is again *Scratchy* with 45 samples out of 120, which corresponds to 37.50 % of its data. For other categories, the number of wrongly flagged anomalies is lower: 11 for *1080Lines*, 17 for *Air Compressor* and 12 for *Paired Doves* (all of them representing less than 20 % of all their samples).

A final observation can be made regarding the *1400Ripples* class, which seems to always be correctly categorised. As shown in Table I, it remains unaffected, regardless of the approach used, and its classification is always correct (0 out of 124 anomalies), suggesting a high internal coherence and separability from other classes.

With the aim of providing a high-level overview of the results obtained from the DBSCAN mechanism across the three proposed configurations, different percentages are shown in Table II. This analysis allows for a better understanding of the results explained in Table I.

It can be observed that only **Case 1** detects 100 % of the true anomalies, but it also shows a high percentage of *false* positives detected, with 24.45 % of non-anomalous samples labelled as outliers (133 false positives from 544 non-anomalous samples). These two figures lead to a reduction of the dataset to be analysed by human experts from the initial set to 31.50 %. **Case 2**, in contrast, shows very low percentages for both real outliers found, 16.07 % (9 out of 56), and non-anomalous samples misclassified as outliers, 0.18 % (1 out of 544). In this situation, only 1.67 % of the initial dataset is sent for human analysis but, as explained before, most of real anomalies are left behind and will not be studied by scientists. Finally, **Case 3** shows a more balanced behaviour: 91.07 % of real outliers are correctly detected and the percentage of non-anomalous images wrongly classified as outliers is notably reduced compared to Case 1. Here, only 15.63 % of non-anomalous data is labelled as outliers (85 out of 544). This reduction in misclassification also decreases the amount of data sent for human analysis: only 22.67 % of the initial dataset is classified as anomalous and, therefore, prepared to be studied by experts.

This comparative analysis highlights the balance between thoroughness and precision. While **Case 1** succeeds in detecting all real outliers, it comes at the cost of an expensive review process with a high number of misleading candidates. In contrast, **Case 2** is extremely selective but misses a substantial portion of relevant anomalies. **Case 3** offers a more balanced outcome, capturing the majority of meaningful outliers while avoiding excessive misclassifications and, thus, reducing the human effort required.

V. CONCLUSIONS

With the results presented, it is shown that a Deep Active Learning approach offers significant advantages for data analysis in gravitational wave interferometers, particularly in the challenge of glitch classification. This method reduces the human effort involved in labelling all incoming samples by focusing on a smaller, more meaningful subset that potentially includes previously unseen glitch signals.

Focusing first on the supervised learning component, the classifier, it demonstrates an adequate ability to distinguish among different glitch images by effectively learning their main characteristics. The integration of the Spatial Attention layer enables the model to concentrate on the most relevant features within each class, enhancing accuracy and reducing computational time. Achieving an accuracy above 98 % supports the practical implementation of the classifier, although some glitch types, such as *Paired Doves*, may require additional attention due to their lower representation in the dataset.

Regarding the unsupervised learning component, its performance should be interpreted with a broader perspective. The implementation of a clustering mechanism helps identify anomalies throughout the dataset, although the specific outcome depends heavily on the selected configuration. Referring back to the obtained results, in scenarios where missing anomalies is not acceptable, **Case 1** would be preferred (even if it comes with the cost of a higher number of false positives and increased human analysis). However, if the goal is to avoid an overwhelming volume of false positives, a more balanced configuration, such as **Case 3**, might be more appropriate.

Overall, this type of mechanism proves useful in problems with a continuous appearance of new, unknown results and where expert human analysis is still necessary to classify each case. It has already been shown in other scientific areas, like medical image analysis, that a Deep Active Learning approach can help experts in identifying different pathologies while easing their workload. The same applies in the present context, where the detectors (interferometers) may produce previously unseen glitch signals (often due to maintenance issues) which are precisely the ones experts should be reviewing to recognise new types of spurious signals. This would lead to improvement and acceleration in the investigation area of gravitational waves, since this approach avoids the need to manually examine the whole dataset, which may include many already known glitches.

All things considered, in this work, the next step would be to improve the performance of the unsupervised component. As mentioned before, small changes in the parameters can lead to very different results. This high sensitivity may be due to the high dimensionality of the feature vectors (i.e., the image representations fed into DBSCAN), as the Euclidean distance becomes less effective at distinguishing nearby samples. Therefore, to overcome the present limitations in DBSCAN, such as its pronounced sensitivity to parameter variations and the struggle with clusters of distinct densities, a more refined clustering mechanism like HDBSCAN (Hierarchical DBSCAN) or similar alternatives should be considered.

REFERENCES

[1] A. Einstein, H. A. Lorentz, H. Minkowski, and H. Weyl, *The Principle of Relativity* (Dover Publications, 1920).

[2] B. P. Abbott, R. Abbott, and T. D. Abbott (LIGO Scientific Collaboration and Virgo Collaboration), Observation of Gravitational Waves from a Binary Black Hole Merger, Phys. Rev. Lett. **116**, 061102 (2016).

[3] C. M. Bishop and H. Bishop, *Deep Learning: Foundations and Concepts* (Springer, Cham, 2024).

[4] M. Razzano and E. Cuoco, Image-based Deep Learning for Classification of Noise Transients in Gravitational Wave Detectors, Classical and Quantum Gravity **35**, 095016 (2018).

[5] S. Bini, G. Vedovato, M. Drago, F. Salemi, and G. A. Prodi, An autoencoder neural network integrated into gravitational-wave burst searches to improve the rejection of noise transients, Classical and Quantum Gravity **40**, 135008 (2023).

[6] Kaggle, Gravity Spy (Gravitational Waves) Dataset, `https://www.kaggle.com/datasets/tentotheminus9/gravity-spy-gravitational-waves` (2023).

[7] LIGO Scientific Collaboration, Gravitational Wave Sources, `https://www.ligo.caltech.edu/page/gw-sources`.

[8] M. Maggiore, *Gravitational Waves, Vol. 1: Theory and Experiments* (Oxford University Press, 2007).

[9] M. Pitkin, S. Reid, S. Rowan, and J. Hough, Gravitational Wave Detection by Interferometry (Ground and Space), Living Reviews in Relativity **14**, 10.12942/lrr-2011-5 (2011).

[10] G. Cella and A. Giazotto, Invited Review Article: Interferometric gravity wave detectors, Review of Scientific Instruments **82**, 101101 (2011).

[11] B. P. Abbott, R. Abbott, T. D. Abbott, F. Acernese, K. Ackley, and C. Adams, GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral, Physical Review Letters **119**, 10.1103/physrevlett.119.161101 (2017).

[12] D. Davis, T. B. Littenberg, I. M. Romero-Shaw, M. Millhouse, J. McIver, F. Di Renzo, and G. Ashton, Subtracting glitches from gravitational-wave detector data during the third LIGO-Virgo observing run, Classical and Quantum Gravity **39**, 245013 (2022).

[13] I. Khater, I. Nabi, and G. Hamarneh, A Review of Super-Resolution Single-Molecule Localization Microscopy Cluster Analysis and Quantification Methods, Patterns **1**, 100038 (2020).