UNIVERSITY OF BIRMINGHAM

Early detection of massive black hole mergers from gravitational wave signals using machine learning

Laurence Allen*
School of Physics and Astronomy,
The University of Birmingham

(Word count: 7548)
(Project partner: Sam Scivier)
(Academic supervisor: Alberto Vecchio)
(Dated: May 5, 2022)

The recent detection of gravitational waves from binary black hole mergers has emphasized the importance of multi-messenger astrophysics to probe the properties of astrophysical objects and phenomena. To enable cross-discipline analysis of such events, the importance of low-latency, early warning 'alerts' have become apparent. Here we use deep learning with neural networks on time series data for rapid detection of binary black holes signals from detector based and galactic background noise sources as well as simulated transient noise glitches. The networks produced highly reliable identification of the events above a threshold signal-to-noise value of approximately 30, distinguishing the signals from noise and showing the feasibility of such a method for real-time detection, enabling follow-up observations.

I. INTRODUCTION

The detection of gravitational waves (GWs) emitted from binary black hole (BBH) mergers in 2015 [1] using the Advanced Laser Interferometer Gravitational-Wave Observatory (LIGO) has enabled new phenomenological tests to Einstein's General Theory of Relativity in great detail. As the sensitivity of GW observatories increases, the frequency of detection of signals is only going to rise. LIGO is currently limited to a detectable frequency band of $10^1 - 10^3 Hz$ [2] due to the length of the interferometers. Additionally, ground-based GW detectors are affected by seismic background noise [3], to negate this and detect lower frequency events, a space based interferometer has been proposed. The Laser Interferometer Space Antenna (LISA) will have a sensitivity band of approximately $10^{-4} - 10^{-1} Hz$ [3], enabling the detection of events that are both more massive, and at higher redshifts. The combination of LIGO and LISA will provide physicists with the tools to observe astronomical events that span large ranges of frequencies, therefore encapsulating a wide variety values of the intrinsic and extrinsic parameters that describe such objects. For this project we are specifically interested in detecting the GW signals from the coalescence of massive to super massive BBHs ($10^5 M_\odot - 10^7 M_\odot$), such as those thought to be at the centre of galaxies.

These massive binary merging events occur relatively frequently at high redshifts [4] where the signal strength is relatively small in comparison to the intrinsic detector noise as well as the background of other GW sources in the Universe. For this reason, the successful detection of a GW signal is limited by the signal-to-noise ratio (SNR). In fact, our galaxy contains many short period (high frequency) stellar-mass binaries such as white dwarfs that will be unresolved by LISA due to the upper limit on the observable frequency band [5]. Within the frequency band this will dominate over the intrinsic detector noise, defining the limitation on the types of other GW sources that are detectable. The LISA observable frequency band

* LGA834@bham.student.ac.uk

constrains the masses of the detectable BBHs to a range of $10^4 M_\odot < M(z+1) < 10^7 M_\odot$ for redshifts ranges, $z < 10$ [5]. These large masses result in massive GW signals of SNR on the order, $S/N \sim 10^3$ [5]. These 'bright' signals will be easier to differentiate from the noise than for the higher frequency signals of LIGO which have considerably lower SNRs.

LISA is comprised of three satellites and therefore can be treated as two statistically independent detectors consisting of two laser arms each. This enables the measurement of two different polarizations of the GW [5] as well as a method to cross-check the signal to confirm that the same behaviour is observed in each detector, ruling out transient noise glitches that have similar waveforms to the desired signals. The position of the binary is calculated from the time series data. As LISA orbits the Sun, the distance from the source to the detector, also known as the luminosity distance D_L, changes, causing the relative velocity of the source to change, creating a Doppler shift. The amplitude and phase of the Doppler shift as well as the relative polarizations yield the angular position of the source.

During the inspiral of the two black holes (BHs), the system loses energy as it emits GWs, reducing the orbital period and hence increasing the observed frequency of the GWs. Therefore, the time spent in the observable frequency band directly depends on the masses of the BHs. For a $10^5 M_\odot$ equal mass BBH the observable time is on the order of a year. However, for $10^7 M_\odot$ this dramatically reduces to the order of a day [3]. This highlights the importance of a rapid, low-latency method for alerting the presence of a coalescence event, enabling the coordinated observation by both gravitational and electromagnetic methods. This multi-messenger astronomy [6] is vital to provide deeper insights about properties of astrophysical objects such as BBHs.

One such signal analysis method is the use of machine learning techniques such as neural networks (NNs). These are algorithms that learn trends and patterns in training data to make predictions about previously unseen data. A supervised NN learns from samples of labelled data to perform either classification or regression [7]. Both these predictions are important; deducing if the signal detected is noise or due to an extra-galactic source such as a BH is essential. Once the identification has been carried out, regression could be used to determine the particular physical parameters such as the chirp masses. This projected investigated the former; a classification algorithm.

The current method of detecting a GW signal is by matched filtering[8], which utilises a large set of waveform templates consisting of different physical parameters that are compared to the detected signal. Therefore, the determination of the physical parameters of a binary system is a very computationally expensive operation, requiring super computers and high-latency times for results. Thus a fast, albeit potentially less accurate, method using NNs is incredibly important to enable near-instant analysis of large quantities of data.

In this paper, we investigate a deep-learning approach to rapidly classify BBH gravitational wave signals from both the galactic background and detector noise as well as transient noise glitches that resemble the waveform of the GW signals. This project encompasses the generation of data sets containing all these forms of noise and signal, as well as training and optimising the design of a neural network. Analysing the performance of the network enabled the affect of the varied and numerous combinations of the model parameters to be understood. All these components were combined into a self developed code pipeline and binary classification NN that was designed without prior reference code. The code pipeline and trained networks are available on GitHub [9].

II. METHOD

A. Generating The Data Set

The aim of the project was to create a proof-of-concept working NN that would perform a binary classification to determine if a data sample was from a massive BBH merger GW signal or noise. To provide a signal that was fully understood and easily manipulated to examine different parameter behaviour, it was decided to produce a model of the signal itself rather than using an existing generated data set. Additionally, a circular binary system is fully described by 15 intrinsic and extrinsic parameters [10], so training a NN on data with this large a parameter space would require vast computational time and resources.

1. Modelling the GW signal

To avoid this unnecessary complexity, the black hole masses were the only intrinsic parameter variables used for the signal generation. The two individual black hole masses give the chirp mass of the system

$$\mathcal{M} = \frac{(m_1 m_2)^{3/5}}{(m_1 + m_2)^{1/5}} \tag{1}$$

For the case where the respective masses of the two black holes are equal, $m_1 = m_2 = m$, the chirp mass can be expressed as

$$\mathcal{M} = \frac{m}{2^{1/5}} \tag{2}$$

All the extrinsic parameters describing the position of the source with respect to the observer were randomly generated within appropriate ranges. Additionally, all values were taken to be in the detector frame rather than in the frame of the source.

Curt Cutler's [5] mathematical derivation of the required equations to describe the physical system was followed for this project and some of the most important equations have been compiled below.

The leading order post Newtonian expansion for the frequency evolution, $f(t)$, of the waveform as a function of time [4] is given by

$$f(t) = \frac{\mathcal{M}^{-5/8}}{8\pi}\left(\frac{t_c - t}{5}\right)^{-3/8} \tag{3}$$

where t_c is the time at the coalescence of the two black holes. This can be rearranged to calculate the maximum coalescence time that enforces the system to pass through the frequency band within the observation time, T_{obs}

$$t_c^{max} = 5(8\pi f_{min})^{-8/3}\mathcal{M}^{-5/3} + T_{obs} \tag{4}$$

where $f_{min} = 10^{-4}$ Hz is the lower limit on the LISA frequency band.

The dimensionless GW strain signal $h(t)$ measured by the LISA interferometer is [5]

$$h(t) = A(t)A_p(t)\cos[\phi(t) + \varphi_p(t) + \varphi_d(t)] \tag{5}$$

where $t = t(f)$, $A_p(t)$ is the polarization amplitude, D_L is the luminosity distance, $\phi(f)$ is the phase, $\varphi_p(t)$ is the polarization phase and $\varphi_d(t)$ is the Doppler phase. The amplitude of the GW signal is

$$A(t) = 2\frac{\mathcal{M}^{5/3}}{D_L}[\pi f(t)]^{2/3} \tag{6}$$

where D_L is the luminosity distance of the binary system. The frequency domain is preferred over the time domain so the strain signal was Fourier transformed to calculate $\tilde{h}(f)$ using the Scipy signal function of Short Time Fourier Transform (STFT)[11]. Due to the sampling parameters of the Fourier transform, the output spectogram from the STFT had dimensions of 100 by 1000 segments, corresponding to the times and frequencies respectively.

The frequency of the smallest innermost stable circular orbit (isco) of the binary is given by

$$f_{isco} = \frac{1}{\pi 6^{3/2}} \cdot \frac{1}{M} \tag{7}$$

where M is the total mass of the binary, $M = m_1 + m_2$. For equal mass black holes the total mass is related to the chirp mass as follows: $M = 2^{6/5}\mathcal{M}$. This equal mass assumption was used when generating the signal.

When the frequency surpasses f_{isco}, the orbit becomes perturbed and is no longer circular. The waveform model did not encapsulate this behaviour and therefore the frequency of the data was limited to $f \leq f_{isco}$. The following analytic Fourier transformed strain signal was used when calculating the SNR [5]

$$\tilde{h}(f) \simeq \begin{cases} A_p(t)\frac{\mathcal{M}^{5/6}}{D_L} f^{-7/6} e^{i[\phi(f) - \varphi_p(t) - \varphi_d(t)]} & \text{if } 0 < f \leq f_{isco} \\ 0 & \text{if } f > f_{isco} \end{cases}$$

where all parameters are those defined in equation 5 and equation 6. Once again this expression is to leading order in the post Newtonian expansion ($\mathcal{O}(v/c)$).

To generate the signal data, an observation time of $T_{obs} = 10^6 s$ ($\sim$ 10 days) was defined. As high chirp mass BBHs only remains in band for timescales on the order of a single day, the ideal observation time would be less than this, to enable detection and allow follow-up observations. However, such a short observation time would also mean that the GW signals would have little time to evolve into different amplitude regimes, hence being harder for the NN to detect. For these reasons the compromised time of $10^6 s$ was selected. Once more is known about the freqeuncy of data downloads from LISA, the observation time can be selected more specifically to match the real mission requirements. The time array of points was generated from Δt up to T_{obs}, consisting of $T_{seg} = 10000$ time segment points. Δt is the time between subsequent data samples and set to a value of 100s.

The chirp mass was randomly drawn from a logarithmic uniform distribution in the range $10^5 M_\odot < \mathcal{M} < 10^7 M_\odot$. With the same approach, the luminosity distance was drawn in the range $1Gpc < D_L < 10Gpc$. These ranges were chosen as they generate the signal-to-noise (SNR) values that would be on the limit of the detectable region as seen in other research [12].

The time of coalescence was drawn uniformly from a distribution in the range $T_{seg} < t_c < t_c^{max}$. This range was chosen because during the LISA mission, the majority of the signals will not coalesce during the mission lifetime so would not be captured in the observation time. Secondly, using this distribution ensured the network was being exposed to the less extreme amplitude regime of the inspiral rather than just the large, rapidly increasing amplitude in proximity to the coalescence point.

Within this total range, the coalescence times were drawn at equal probabilities from two sub-ranges: $T_{seg} < t_c \leq T_{obs}$ and $T_{obs} < t_c < t_c^{max}$. The first range creates a uniform distribution but, for the second, the maximum coalescence time has a non-linear relationship to chirp mass, $t_c^{max} \propto \mathcal{M}^{-5/3}$, as seen by equation 4. Therefore, for the smallest chirp mass binaries of $10^5 M_\odot$, $t_c^{max} = 140 T_{obs}$, whereas for $10^7 M_\odot$, $t_c^{max} = 1.06 T_{obs}$. This resulted in the second t_c range producing a very non-uniform distribution, strongly peaked near T_{obs}. The true t_c distribution could be calculated from the chirp mass distributions of binary black hole systems in the universe and would favour almost all the signals coalescing outside of the observation time. Hopefully these mass

distributions will be quantified with the collection of data on high mass systems during the LISA mission, enabling more accurate distributions for t_c to be generated.

2. Noise

The LISA sensitivity curves were constructed to sample the noise associated with the three detectors forming the interferometer. The signal response function is well-described by [13]

$$\mathcal{R}(f) = \frac{3}{10}\frac{1}{\left(1 + 0.6\,(f/f_*)^2\right)} \tag{8}$$

where $f_* = 19.09$ mHz. The total noise in a Michelson-style LISA detector is

$$P_n(f) = \frac{P_{OMS}}{L^2} + 2(1+\cos^2(f/f_*))\frac{P_{acc}(f)}{(2\pi f)^4 L^2} \tag{9}$$

where P_{OMS} is the single test mass acceleration noise, P_{acc} is the single-link optical metrology noise and $L = 2.5$ Gm. Dividing $P_n(f)$ by $\mathcal{R}(f)$ gives the mathematical fit of the LISA sensitivity curve

$$S_n(f) = \frac{P_n(f)}{\mathcal{R}(f)} \tag{10}$$

The galactic confusion noise is well described by the function

$$S_c(f) = Af^{-7/3}e^{-f\alpha+\beta f\sin(\kappa f)}\left[1+\tanh(\gamma(f_k - f))\right] \tag{11}$$

As the LISA mission progresses, more foreground sources are removed causing the galactic confusion noise to decrease. For this model the parameters were used for a mission duration of a single year and shown by table II in appendix A. The characteristic strains of the sensitivity and confusion noise curves are plotted in figure 1 alongside a GW signal. Combining the two different noise curves gives the complete sensitivity curve

$$S(f) = S_c(f) + S_n(f) \tag{12}$$

The noise data was then generated by sampling a stationary Gaussian distribution of the complete sensitivity curve, $S(f)$.

3. Signal-to-noise ratio (SNR)

The squared amplitude of the optimal SNR for the Fourier transformed strain signal $\tilde{h}(f)$ [13] is given by

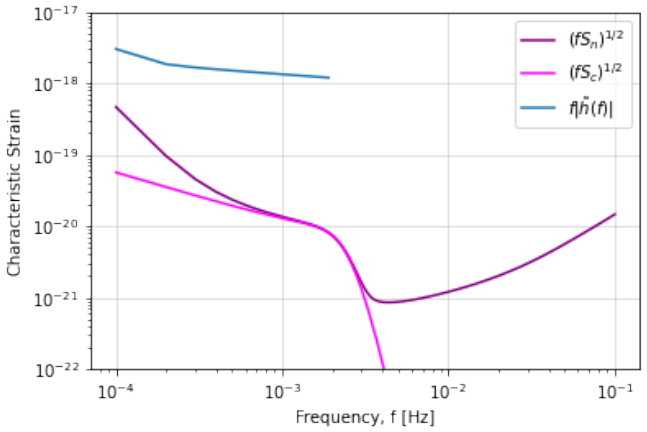

FIG. 1. Dimensionless characteristic strain for the sensitivity curve ($\sqrt{fS_n(f)}$), galactic confusion noise curve ($\sqrt{fS_c(f)}$) and GW signal ($f|\tilde{h}(f)|$). The parameters of the binary used to generate the GW signal were $\mathcal{M} = 10^6 M_\odot$, $D_L = 5Gpc$ and $t_c = 0.5T_{obs}$.

$$\rho^2 = 4\int \frac{|\tilde{h}(f)|^2}{P_n(f)}df \tag{13}$$

where $P_n(f)$ is the noise in a LISA detector. The bounds of the integral are set by the LISA observable frequency band. The lower frequency limit was the maximum of either f_{min} or the frequency at the start of the observation time, $f_{lower} = max\,(f_{min}, f(t = \Delta t))$. If the t_{isco} drawn exceeded the observation time, $t_{isco} > T_{obs}$, then the upper frequency limit was set as $f_{upper} = f(t = T_{obs})$. If this criterion was not met, then $f_{upper} = f(t = t_{isco}) = f_{isco}$. Due to the chosen range of the $\mathcal{M}$ and D_L parameters, $f_{isco} < f_{max}$, so f_{isco} was always within the upper limit of the LISA frequency band.

Figure 2 is a logarithmic plot of SNR as a function of luminosity distance and therefore shows the power law relationship. The amplitude of the strain signal is proportional to $1/D_L$ as shown by equation 6. SNR is proportional to the magnitude of the strain signal and therefore the following relationship of SNR $\propto 1/D_L$ is reached. This explains the gradient of -1 shown in the graph. As mass increases, the SNR decreases. This is because the upper integration limit is determined by f_{isco} for all these signals as they coalesce within the observation time. f_{isco} is inversely proportional to $1/\mathcal{M}$ so as the mass increases, the region over which the frequency is integrated diminishes, resulting in a lower SNR.

4. Glitches

Glitches are transient (short duration) non-Gaussian noise events present in data from GW observatories [14]. These glitches resemble the GW signature of massive bi-

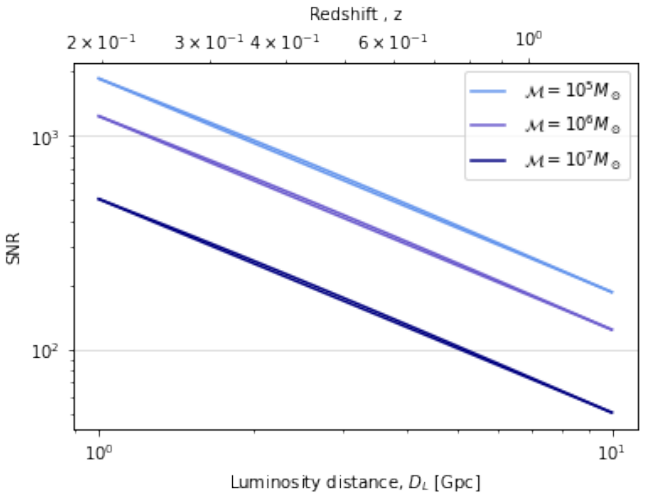

FIG. 2. SNR as a function of luminosity and redshift for three different chirp mass binaries with $t_c = 0.5T_{obs}$. Notable SNR valuers were: $\text{SNR}_{min} \sim 50$, $\text{SNR}_{max} \sim 1850$, $\text{SNR}_{mean} \sim 310$. The gradient of the data is $1/D_L$.

nary black hole mergers. Therefore, it is crucial to correctly classify these objects as noise rather than GW signal to mitigates the adverse affects that the glitches would otherwise cause on the network and increase the sensitivity of this rapid detection system. Secondly, due to the length of the LISA mission, the probability that these glitches will be observed is very high and therefore essential that they are included in the data.

A sine-Gaussian waveform was used to model these glitches. This consists of a sinusoid modulated by a Gaussian envelope

$$g(t) = A_g \sin(2\pi f_g t + \phi_g) \exp\left(-\frac{(t-t_g)^2}{2\tau_g^2 \cdot SF}\right) \tag{14}$$

where A_g is the amplitude of the glitch, f_g is the frequency of the glitch, ϕ_g is the initial phase, t_g is the time at which the glitch is centred, $\tau_g = 1/f_g$ and SF is the scaling factor that controls the width of the Gaussian envelope. A_g was scaled to match the amplitude of the strain signal ($\sim 10^{-18}$). The glitch frequency was chosen in the range $10^{-5} < f_g < 10^{-1} Hz$ to probe the region associated with the LISA observable frequency band. Note the lower bound of $10^{-5} Hz$ is lower than the stated lower bound of LISA. This was to push the network into the region where some parts or all of the glitch would not be in band to make the model more realistic. All the glitches were decoupled from the frequency by setting the scaling factor to a value that was drawn randomly in the range $10^3 < SF < 10^5$.

5. *Data types*

LISA consists of three satellites which enables the interferometer to be configured as a pair of two-arm detectors which output two linearly independent signals. This second detector channel still has stationary and Gaussian noise sampled from the complete sensitivity curve. The noise in the three arms of LISA will be correlated [5], but for this paper the noise was assumed to be symmetric. This simplification is reasonable as the satellites are effectively identical and hence the instrumental noise will therefore also be approximately identical. Similarly, the galactic confusion noise can be considered symmetric due to the isotropic nature of the background.

This results in the strain of the second detector, $h_{II}(t)$, being equivalent to the output of the first detector $h_I(t)$ rotated by $\pi/4$ radians. The parameter that this rotation was applied to was $\phi_s(t)$, which describes the source location in the detector coordinate system.

The three different components; GW signal, noise and glitch were combined in different permutations to create the complete data set. The four types of data subsets that were investigated were as follows:

$$d_k^1(t) = h_k(t) + n_k(t) + g_k(t) \tag{15a}$$
$$d_k^2(t) = h_k(t) + n_k(t) \tag{15b}$$
$$d_k^3(t) = n_k(t) + g_k(t) \tag{15c}$$
$$d_k^4(t) = n_k(t) \tag{15d}$$

where the subscript k denotes detector channel, $k = \{1, 2\}$. From these four data subsets there are two distinct classes: those that contain GW signal and therefore will be classified as "signal", and those that only contain noise and hence classified as "noise". Therefore, it is clear to see that a perfect binary classifier neural network would class d^1 and d^2 as signal and d^3 and d^4 as noise. Note the removal of the k subscript here as the classification is independent of the detector channel.

B. The neural network

This project used a supervised deep artificial NN that emulates the structure, or architecture, found inside the brain by stacking layers of artificial neurons [7] to form layers. For a deep NN there is an input layer, followed by one or more hidden layers leading to a final output layer.

The input layer feeds the data into the network and the output layer yields the associated probability that a specific data sample is associated to one of the two classes, either signal or noise. The neurons in the hidden layers perform a number of linear operations [12] to learn the intricate behaviour of the input training data. The computation for a single neuron can be simplified to a description of a linear relationship

$$z = wx + b \tag{16}$$

where x is the input data or features, w is the weight associated to the neuron, b is the bias which the network adds to after multiplying the data by the weight and z is the predicted output or label for the data. As the network trains on the data, it iteratively updates these parameters as it tries to optimize them to better describe the data.

The perfect weights and biases are not calculated as this would be extremely computationally costly due to the size and intricacy of the unknown parameter space. Instead, the problem of learning is tackled by an optimization algorithm where "good enough" predictions of the weight and bias are reached quickly. The optimization algorithm used is stochastic gradient descent which enables the weights to be calculated using the back propagation of an error algorithm. This error process involves moving along the error gradient curve in steps, determined by the learning rate of the network, which then reduces the error at the next evaluation.

When referring to the error evaluation of the data, the metric of loss defined by a loss function was used. Loss is a number which describes how inaccurate the model's prediction was in comparison to the true data. For this project the loss function used was cross entropy, which describes the cross entropy between two probability distributions. The first distribution is the known probability of each class label. As this is a binary classification problem, this probability will be 1 for the correct label class and 0 for the other. The second distribution is the predicted probability that the data belongs to specific label class[15]. This predicted probability was produced in the output layer which consisted of just two neurons, one associated with each class, and each activated with a softmax function.

The batch size was set as 10; this determines the number of data samples the network works through before updating the parameters. The learning algorithm was run for 100 epochs, which means the network was exposed to the complete data set 100 times.

A supervised learning network requires that data sets are divided into training, validation and testing data. The training set is the input examples which the network learns from by fitting the parameters mentioned previously. The validation set provides an evaluation of the model fit and updates the model's parameters before the next epoch [16]. The test set is held back from the network to remain unseen and provides evaluation on the final model fit. The split ratio between the three data subset types within the network training data was 8:1:1. The validation and test subsets enabled the performance of the trained network to be quantified before it was fed a secondary test set containing data which had $\mathcal{M}, D_L, t_c$ parameters that differed from those which it was trained on.

The design of the network was optimized by tuning multiple hyper-parameters such as the number and type of layers within the network, the number of neurons within each layer and the associated activation functions. There is no easy way to pick the values of the hyper-parameters because there are essentially infinite combinations. Instead, the performance of the network was compared graphically by plotting loss and accuracy for a number of different values and selecting the best parameters based on these results. This means that the chosen values are by no means the best possible, but are adequate to achieve meaningful results.

A major issue that is prevalent in NNs is that of over-fitting. This occurs when the model learns the details of the training data to the extent that the performance of the model on new unseen data is negatively impacted. To stop such over-fitting, dropout regularization was used between each hidden layer at a rate of 0.25. This randomly sets the input units of the hidden layers to 0 with a frequency of 25% so they do not contribute to the learning for that epoch [17]. Dropout has the effect of making the training process noisy, forcing neurons within a hidden layer to probabilistically take on more or less responsibility for the inputs. This in turn stops the network layers co-adapting to the data together, increasing the robustness of the model. The other main way to reduce the effect of overfitting is to have very large training data sets. Therefore, the size of the training data sets were maximised to the capacity of the computational memory available.

The size of the input features was reduced from the original spectogram which had dimensions (1000, 100) to (500, 50) via array compression to further increase the operating speed of the network. The input layer fed into three dense layers of neurons which were fully connected to each neuron in the preceding layer. The dense layers were comprised of 250, 100 and 25 neurons respectively and were activated by a ReLU function. The learning rate was initialised at a value of 10^{-2} but scheduled to decrease exponentially after the first 10 epochs, increasing the rate the model learnt the data.

To construct and train the network, the Python library Keras was used, which provides an API to access TensorFlow [18]. The Adam optimizer was used, which applied the Adam algorithm [19], a type of stochastic gradient descent.

III. RESULTS

Initially the network was trained without glitches, on a data set of 20000 data samples comprising an equal split of d^2 and d^4 data subsets. The coalescence time was drawn in the range $T_{seg} < t_c < t_c^{max}$. The generation time to create this training set was approximately 10 minutes using 16 4GB CPU cores on the University of Birmingham BlueBEAR supercomputer.

The neural network was then trained on this data set,

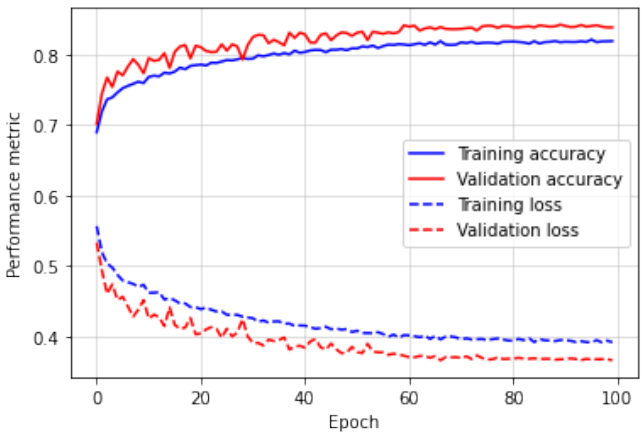

FIG. 3. Performance metrics of accuracy and loss as a function of epoch for training and validation data subsets. The network was trained on 20000 samples, evenly split between d^2 and d^4 data type subsets. Test accuracy and loss were 84.0% and 0.379 respectively

taking approximately three minutes (30 mins CPU time). Figure 3 shows the performance metrics of accuracy and loss which both plateau at approximately 80 epochs. This plateauing shows that an adequate number of epochs were used, enabling convergence to consistent accuracy and loss values. The validation values follow that of the test set, showing that the regularization used to reduce over-fitting worked successfully. There is a slight seperation of the two, suggesting that more data samples should be used. The evaluation of the network for the 2000 test data samples took less than a second, showing that once the network is trained, the classification of new data has incredibly low-latency.

This trained network was then tested on a different unseen data set consisting of 5000 data samples for each detector channel with all the same parameters as the training data set except that the coalescence time was now drawn in the range $T_{seg} < t_c < T_{obs}$. This range was chosen as an initial check to make sure that the effect of the glitches could be isolated from the performance of the network. Channel one comprised of noise and signal data, d^2, the same as that of the training data. Note this can be written more compactly in the "channel notation", d^2_1. Channel two ,however, had the addition of glitches, d^1_2.

The accuracies of channel one and two were 86.1% and 99.4% respectively. The reason the second channel outperformed the first is because the trained network had "no concept" of glitches, only of signals. Therefore, when a glitch appeared in the spectogram, the network classified it as signal as they have similar characteristics. This false positive results in an improved accuracy, but for an undesired reason.

Sometimes the network can struggle to accurately predict which class a specific data sample belongs to. The main factor that determines the network's ability to perform such a classification correctly is the SNR. To negate the effect of a false positive in a single detector channel, the two channel data streams were compared for concordant results. If the two channels classed the data differently, then they were rejected as a possible signal, leaving only the concordant signals. This process is referred to as cross checking the channels. For the case above, the cross check percentage was 86.6%.

This false positive effect was magnified when a test set containing no signals, only noise and glitches (d^4_1 & d^3_2), was analysed. The accuracies in the two channels were 98.0% and 6.74% respectively. This shows exactly the same behaviour as before, however now the glitches hinder the classification rather than aid it. The cross check percentage was 8.50%, at least providing some indication that there was a major discrepancy between the two channels.

To rectify the negative effects glitches had on the classification process of the test data, a new improved network was trained that included the glitches. The training set consisted of all four data subsets; d^1, d^2, d^3, d^4. Due to now only having 25% of each data subset, the total number of data samples was increased to 40000. Having such a large data set required more memory and therefore the number of cores used was increased to 32. All other parameters were kept the same as before and the generation time was approximately 16 minutes.

The architecture of the neural network remained the same, with the exception of the batch size which was reduced to 20. This value was chosen as it ensured the validation loss followed the training, meaning the network wasn't overfitting. Upon testing with the test subset of the training data, the network performed well, achieving an accuracy of 78.9% and a loss of 0.467. These metrics were worse than the previously trained network due to the addition of the glitches, which require more intricate connections between the neurons in each layer. The training time was 8 minutes (2 hours 15 mins CPU time). This training time was more than 2 times greater than the previous network, which one would expect from an increase in the data size by a factor of 2, due to the reduced batch size which results in the model spending longer updating its internal parameters.

This network was then tested on the previous data set containing d^2_1 & d^1_2. The accuracy of the two channels was 80.3% and 81.0% respectively with a cross check percentage of 87.2%. To an appropriate degree of precision these two results are the same, showing that the glitches no longer have an impact upon the network's ability to classify signals. Testing with the d^4_1 & d^3_2 data that did not contain signals, resulted in accuracies of 99.0% and 86.8% respectively with a cross check percentage of 86.0%. Once again, the channel containing glitches has worse performance than the channel without. However, this is a vast improvement from the accuracy of 6.74% previously.

From this point on, the network was pushed into progressively more difficult regimes by changing different parameters of the test data to analyse the performance for

signals. As the noise is stationary and Gaussian it is independent of the choice of these parameters, meaning the performance of the network would remain the same as stated above. Therefore, only evaluation of test data containing signals was pursued. The network was extremely effective at correctly classifying the set of signals discussed above because if they coalesce within the observation time, then by definition they have high SNRs as well as having sharp features in the spectogram that the network can learn from. To push the network to detect signals with lower SNRs, and hence determine the minimum effective SNR threshold for accurate detection, a new test data set was formed containing signals with coalescence times that were drawn in the range $T_{seg} < t_c < t_c^{max}$, the same as the training set.

The channels of the test set contained 5000 samples of each data type; d_1^2 & d_2^1, achieving accuracies of 60.1% and 60.7% respectively with a cross check percentage of 89.3%. To further understand the effect that the choice of the coalescence time had on the network's results, a figure was generated of the outputted probabilities from the output layer that were associated with classifying the data samples as signals. This was plotted as a function of t_c.

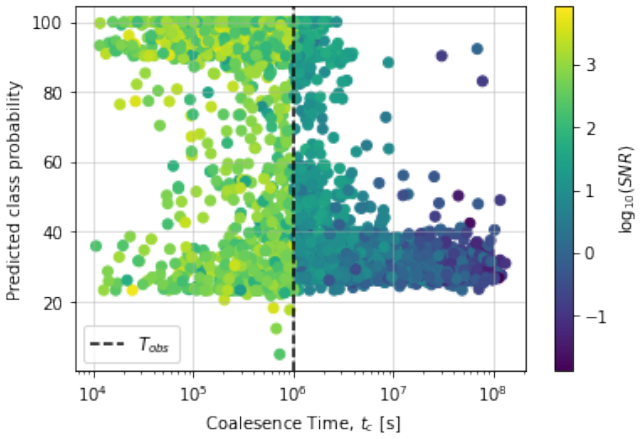

FIG. 4. Predicted signal class probability as a function of coalescence time with a color scale displaying $\log_{10}(SNR)$. Test data contained signals and glitches (d^1), with coalescence times drawn in the range $T_{seg} < t_c < t_c^{max}$.

Figure 4 shows that signals which coalesce outside the observation time have a much lower probability of being correctly classified by the network. This is because only the inspiral of the signal is captured, hence resulting in a low SNR. Because the type of classification carried out by the network is binary it means that if the predicted class probability exceeds 50% then the data will fall into that class. So almost all the events which coalesce outside T_{obs} are being incorrectly classified.

The SNR is not only dependent on the coalescence time but also on the chirp mass and luminosity distance. Figure5 shows that the network was most accurate at classifying the events which had large chirp

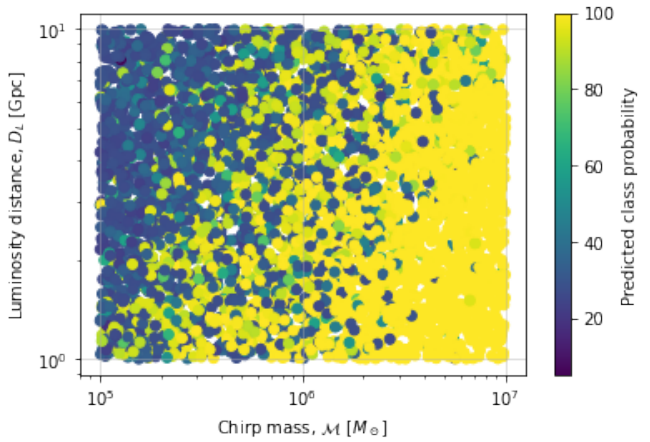

FIG. 5. Signal predicted class probability dependency on chirp mass and luminosity distance. Test data contained signals and glitches (d^1) with coalescence times drawn in the range $T_{seg} < t_c < t_c^{max}$.

masses and were at small distances. This is because the SNR is proportional to $D_L^{-1/2}\mathcal{M}^{5/6}$. There is a subtle additional complexity to this behaviour. The upper integration limit for the SNR calculation, equation 13, is the frequency f_{isco} which is inversely proportional to chirp mass. Therefore, as chirp mass changes there is a playoff between the two proportionalities so as a consequence, the predicted signal class probability is not actually directly proportional to $\mathcal{M}^{5/6}$.

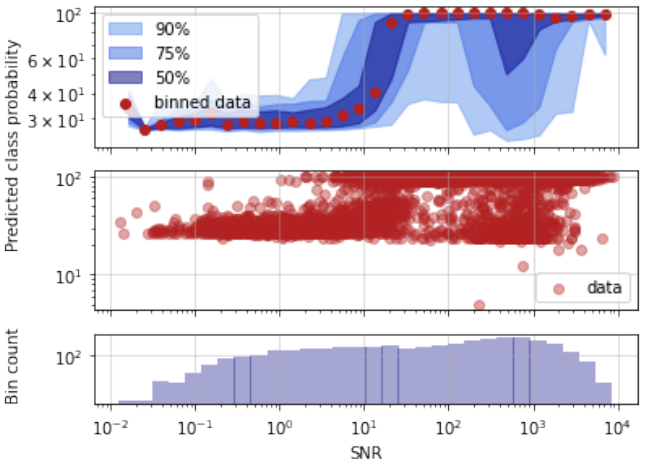

FIG. 6. Sensitivity curve illustrating the ability of the neural network to detect BBH GW signals from noise and transient glitches. Data statistically binned into 30 bins and the median value taken. The shading shows the nth central percentile which shows the spread of the data. Bin count histogram represents the statistical weight that each binned data point carries.

To quantify the network's ability to determine BBH GW signals from noise and transient glitches, a sensitivity curve was constructed. The first panel of figure 6

shows the sensitivity curve for statistically binned data into 30 bins where the median value of each bin is then taken and displayed. In each bin the central nth percentile is taken and displayed as a shaded area to visualize the spread in data and hence the associated error. For high SNRs the median probability is approximately 100%. However, it carries a very large error and even dips below 100% for very high SNRs. The reason for this is due to the drawing of the coalescence time. As t_c is drawn from two different distributions it creates a dual band structure in the data, as seen in the second panel of figure 6.

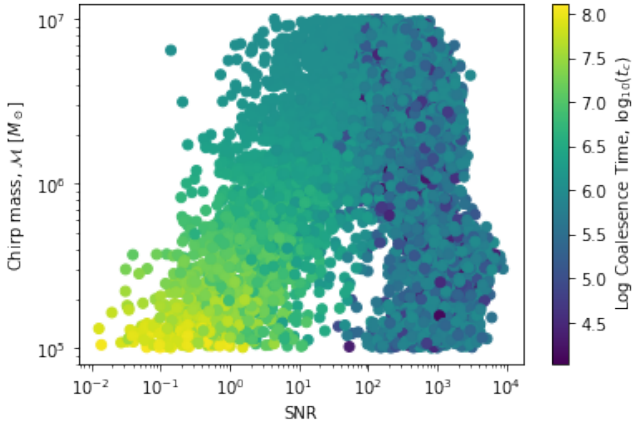

FIG. 7. Relationship between chirp mass and SNR for coalescence times drawn at equal probabilities from two distributions of ranges $T_{seg} < t_c < T_{obs}$ and $T_{obs} < t_c < t_c^{max}$. The color scale cut off between the two distributions is at $\log_{10}(T_{obs}) = 6$.

Figure 7 shows that for chirp masses in range $10^5 M_\odot < \mathcal{M} < 10^6 M_\odot$, there are two distinct SNR regions that they can fall into, depending on which distribution the coalescence time was drawn from. This bifurcation of the data inhibited the quality of results for the high SNR sensitivity curve, so new results were gathered from a test data set that contained all the coalescence times drawn outside of T_{obs}. Equally, the mass could have been limited to be above $10^6 M_\odot$, to avoid the bifurcation. However, during the LISA mission it is likely most merger events will coalesce outside of the observation time so this was a better compromise to make than loosing half the range of masses.

Figure 8 shows that the dual SNR region issue was resolved for the new t_c range. The statistical weight that each binned data point carries is determined by the total number of data points in that bin. Due to the uniform logarithmic drawing of $\mathcal{M}$ and D_L combined with the long tailed t_c distribution, the extremity regions in SNR have a low bin count as shown by the histogram. This is the reason for the slight increase in predicted class probability at SNR $= 10^{-2}$. This is not a major issue as the main region of interest, the minimum effective SNR threshold for accurate detection, has the highest bin

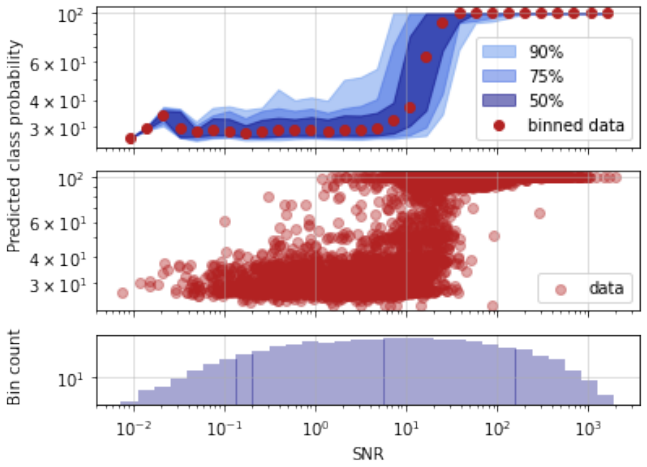

FIG. 8. Sensitivity curve for coalescence times drawn outside the observation time, from a single distribution in the range $T_{obs} < t_c < t_c^{max}$

count by design and therefore carries the most statistical confidence. Overall, this figure shows that the binned predicted signal class probability saturates to 100% for SNR ≥ 30 and that all signals were correctly identified above an SNR of approximately 100.

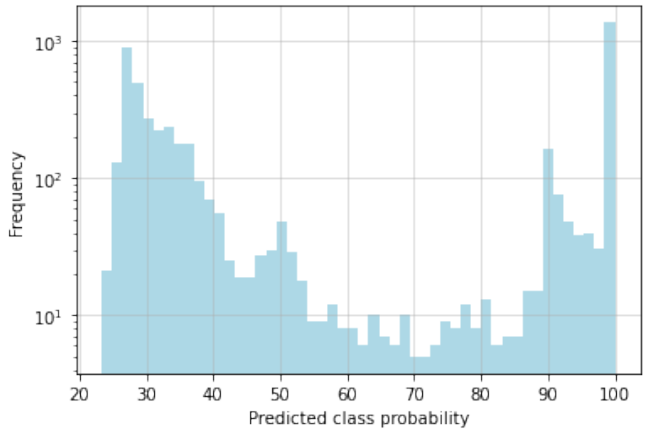

FIG. 9. Predicted class probability frequency distribution test data containing signals and glitches, d^1, with coalescence times drawn in the range $T_{obs} < t_c < t_c^{max}$.

For low SNR signals the binned class probability saturates at approximately 30 rather than falling to zero. The predicted class probability distribution shown by figure 9 is trimodal because of the three significant peaks. This structure is suspected to be from the network training process. For GW signals with a high t_c^{max}, the track is only in the LISA frequency band for a short period of time. This means that only a few, or even a single pixel segment of the spectogram contains the signal. Therefore, to the network, this training example looks no different to a noise example, except that the associated label is telling the network that it is a signal. All outputs of

the network are based on the training process minimizing the loss function, therefore this trimodal structure must be a compromise of this minimisation.

The trained network was then used to evaluate a number of different test data sets. Each of these data sets had slightly different ranges of the $\mathcal{M}$ and t_c parameters. These results are summarised in table I and for completeness the results discussed previously have been included in the entry for test data sets 4 and 7. Three different mass ranges were used; high, low and a combination of the two. Coalescence time was either drawn so that all the samples coalesced inside, outside or evenly split between the two. As expected, the channel containing the glitches generally outperformed the channel without glitches. Data sets with higher masses performed better than low masses due to higher SNR values. For all data sets the accuracies of the combined mass range was almost exactly the average of the sum of each high or low mass range as expected.

For all the parameter variations, the minimum effective median binned SNR threshold for accurate detection of signals remained approximately in the range 20 to 30. The performance of the network remained constant with median binned data yielding 100% predicted probability above the SNR threshold. This shows that the accuracies should really be thought of as the percentage of signals below said threshold rather than an evaluation of how well the network classifies the data. Therefore, for the majority of BBH mergers with chirp masses in the range $10^6 M_\odot < \mathcal{M} < 10^7 M_\odot$ at distances of $1Gpc < D_L < 10Gpc$, the network correctly classifies them, proving its use as a reliable rapid detection tool. This imposed lower mass limit suggests that the study of more massive objects using this method would yield even higher detection accuracies.

Apart from the SNR, the other main factor that effects the network's ability to detect signals is the amplitude of the glitches, A_g. The initial test data sets discussed above had a glitch amplitude on the order 10^{-18}. Four more test data sets were analysed with amplitudes of order 10^{-15}, 10^{-16}, 10^{-19} and 10^{-20} with masses and coalescence times drawn from $10^6 M_\odot < \mathcal{M} < 10^7 M_\odot$ and $T_{obs} < t_c < t_c^{max}$ respectively. The smaller amplitude glitches had no effect, causing both channels of d^2 & d^1 to have equal accuracies. For 10^{-15}, 10^{-16} the accuracies were 69.5% and 70.0% respectively compared to 71.2% without glitches, showing a decrease as the amplitude increased. This result was surprising, as up to this point the channel containing glitches had always been equal to or outperformed the channel without.

Plotting the sensitivity curve showed that the low accuracy was due to a large spread in the predicted signal class probabilities of the high SNR signals. The fact that this bimodal distribution is only seen for high SNR suggests that the network struggles when the glitches are very 'bright' in the spectogram, eclipsing the signals. Additionally, the network was not trained on these glitches and therefore at these amplitudes they look significantly

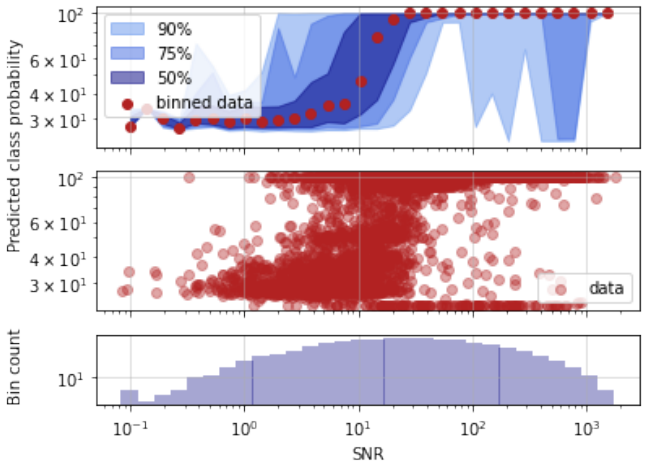

FIG. 10. Sensitivity curve for glitches with an amplitude drawn in the range $0.5 \times 10^{-15} < A_g < 1.5 \times 10^{-15}$.

different to either the signal or noise classes it understands. Further research is needed to understand the behaviour of high amplitude glitches and if such transient noise events would physically occur with such magnitude.

Finally, the test luminosity distance was increased so that it was drawn in the range $10Gpc < D_L < 100Gpc$. Note that this was the first time a test data set parameter was drawn so that it fell outside the range that the network was trained on. The chirp mass was in the range $10^6 M_\odot < \mathcal{M} < 10^7 M_\odot$ and coalescence time $T_{obs} < t_c < t_c^{max}$ with the original glitch amplitude of order 10^{-18}. Due to the parameters, the majority of the data samples had SNRs below the threshold, resulting in an accuracy of 23.9%. As shown in figure 11, the binned data sensitivity curve does not saturate at 100% until an SNR of approximately 60, yet it reaches 90% at the old threshold of 20. As the SNR is proportional to $1/D_L$, the data samples appear 'dimmer' in the spectogram that the network receives. However, the shape of the track remains similar to that of the training data due to the mass range. Therefore, the network does not classify the samples of SNR > 20 with 100% certainty as they have lower amplitudes then expected with those specific spectogram tracks. Once again the network has correctly classifies all the signals above an SNR of approximately 100.

IV. FUTURE WORK

Due to time constraints, not all areas of this research project were fully investigated and so for completeness some further research avenues and improvements have been listed.

The model of the GW signal is currently oversimplified. The amplitude increases up to t_c where it suddenly drops to zero. This is because only the inspiral behaviour of the BBH is being modelled. In fact there are two other stages: merger and ringdown [20]. Merging occurs once

	Parameter ranges		Detector channel 1		Detector channel 2		
Test data set	$\mathcal{M}$	t_c	Data type	Accuracy	Data type	Accuracy	Cross check
1	$10^5 M_\odot < \mathcal{M} < 10^7 M_\odot$	$T_{seg} < t_c < T_{obs}$	d^2	80.3%	d^1	81.0%	87.2%
2	$10^5 M_\odot < \mathcal{M} < 10^6 M_\odot$	$T_{seg} < t_c < T_{obs}$	d^2	62.4%	d^1	64.4%	79.5%
3	$10^6 M_\odot < \mathcal{M} < 10^7 M_\odot$	$T_{seg} < t_c < T_{obs}$	d^2	98.0%	d^1	97.8%	96.4%
4	$10^5 M_\odot < \mathcal{M} < 10^7 M_\odot$	$T_{seg} < t_c < t_c^{max}$	d^2	60.1%	d^1	60.7%	89.3%
5	$10^5 M_\odot < \mathcal{M} < 10^6 M_\odot$	$T_{seg} < t_c < t_c^{max}$	d^2	36.0%	d^1	36.6%	86.1%
6	$10^6 M_\odot < \mathcal{M} < 10^7 M_\odot$	$T_{seg} < t_c < t_c^{max}$	d^2	84.2%	d^1	85.7%	90.4%
7	$10^5 M_\odot < \mathcal{M} < 10^7 M_\odot$	$T_{obs} < t_c < t_c^{max}$	d^2	40.6%	d^1	40.8%	89.2%
8	$10^5 M_\odot < \mathcal{M} < 10^6 M_\odot$	$T_{obs} < t_c < t_c^{max}$	d^2	10.4%	d^1	10.8%	93.6%
9	$10^6 M_\odot < \mathcal{M} < 10^7 M_\odot$	$T_{obs} < t_c < t_c^{max}$	d^2	71.2%	d^1	71.0%	85.1%

TABLE I. Accuracies and cross check percentages of two differing channel data types consisting of a sample of 5000 data points for a total of nine test data sets with the $\mathcal{M}$ and t_c parameters each drawn from three distinct ranges. The D_L distribution was drawn from $1Gpc < D_L < 10Gpc$ for all data sets.

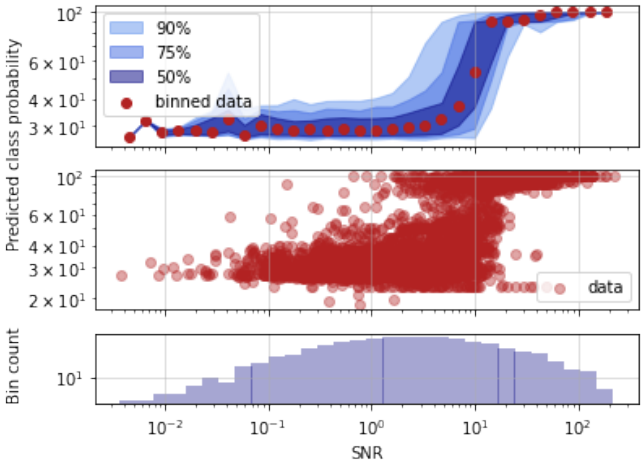

FIG. 11. Sensitivity curve for luminosity distances drawn in the range $10Gpc < D_L < 100Gpc$, outside those of the training data.

the BHs are within a certain separation of each other [21]. Once merged, the single BH dissipates more GWs to become stable through ringdown. Improving the model will remove sudden jumps in the signal that is fed into the network. This should decrease or remove the dependency with which the network uses these features to classify the data, providing more consistent analysis of the behaviour of the whole signal. This was not a major issue for the majority of the results due to the decision to draw t_c outside of T_{obs}. However, for a complete detection algorithm that did not contain simplifying assumptions, this would be useful.

In this paper, the only comparisons between the two statistically independent detector channels were made after the network's evaluation of each lone channel. Instead, both detector channels could be fed into the NN as a single input data stream. This would enable the NN to learn the relationship between two phase shifted signal wave forms and therefore hopefully result in better glitch identification. Unfortunately, this could also result in loss of the individual determination of a particular channel which could be useful when trouble shooting issues within specific detector arm configurations.

As the type of classification is binary, each data sample is assigned to a specific class if the probability predicted by the network exceeds 50%. Therefore, if the network is relatively unsure on which class the data belongs to, it could be incorrectly classified due to only just surpassing the 50% cutoff. To counteract this effect, the classification threshold for signals could be lowered to say 40%, where the data could then undergo further, albeit potentially slower, multi-method analysis. This is an important consideration as the number of BBH GW signals observed during the LISA mission could be fairly low, so it is vital that any signals are not missed.

Finally, the model proved effective at detecting BBH GW signals from both galactic background noise as well as intrinsic detector and transient noises. To increase the complexity and realism of the model, other compact objects such as white dwarfs and neutron stars should be included. This would force the network to more intricately learn the specific behaviour of each different type of signal as these other compact objects would look more similar to the BBH waveform than glitches and noise. Additionally, these objects would have lower SNRs and therefore a NN trained on data including them could reduce the SNR threshold for accurate detection.

V. CONCLUSIONS

In this project, a complete model of the BBH merger inspiral GW signal was generated as well as stationary Gaussian noise sampled from the complete LISA sensitivity curve. These two data components were combined with transient non-Gaussian noise glitches to create a physically realistic training data set, representative of the LISA mission. A Python code pipeline was self developed to train and evaluate a deep NN on a varied range of parameters. The architecture and hyper-parameters of the network were optimised to have the highest accuracy possible on the training set.

Once trained, the NN was evaluated on multiple dif-

ferent test sets. The best performance was achieved for high mass coalescence events that occurred within the defined observation time as these had the highest SNRs. The network was then pushed into lower SNR regimes by varying the chirp mass, coalescence time, glitch amplitude and luminosity distance. Sensitivity curves were plotted to quantify the minimum effective SNR threshold for accurate detection of GW signals. For all the different parameter ranges within those of the training set, the network yielded 100% median binned detection probability for signals above a SNR threshold of 30 and correctly identified all signals with an SNR of 100 and above. This was the same order of magnitude as other research carried out [12], although our study included the additional complexity of glitches which are the expected cause of the slightly higher SNR threshold. This SNR of 100 limit for total detection shows that the network will be highly effective at detecting large mass ($> 10^6 M_\odot$) BBH mergers during the LISA mission. Although the training of the networks was computationally costly and relatively slow, once trained, the network could classify thousands of signals in less than a second, proving the viability of NNs as a rapid low-latency detection method. These early alerts will enable additional follow-up observations using both gravitational and electromagnetic methods.

To improve on the work developed here, it is important to create more complex data models. These would incorporate other GW binary sources such as neutron stars and white dwarfs, or increase the number of intrinsic and extrinsic parameters captured by the mathematical description of the BBH. Finally, increasing the size of the training data set will expose the network to a greater variety of signals, achieving better results.

Appendix A

Parameter	Value
α	0.171
β	292
κ	1020
γ	1680
f_k	0.00215

TABLE II. Parameters for the analytic fit of the confusion noise described by equation 11 [13].

[1] P. A. et al., Observation of gravitational waves from a binary black hole merger, Physical Review Letters **116**, 10.1103/physrevlett.116.061102 (2016).

[2] D. Martynov, E. Hall, B. Abbott, R. Abbott, T. Abbott, C. Adams, R. Adhikari, R. Anderson, S. Anderson, K. Arai, and et al., Sensitivity of the advanced ligo detectors at the beginning of gravitational wave astronomy, Physical Review D **93**, 10.1103/physrevd.93.112004 (2016).

[3] P. Amaro-Seoane, H. Audley, S. Babak, J. Baker, E. Barausse, P. Bender, E. Berti, P. Binetruy, M. Born, D. Bortoluzzi, J. Camp, C. Caprini, V. Cardoso, M. Colpi, J. Conklin, N. Cornish, C. Cutler, K. Danzmann, R. Dolesi, L. Ferraioli, V. Ferroni, E. Fitzsimons, J. Gair, L. G. Bote, D. Giardini, F. Gibert, C. Grimani, H. Halloin, G. Heinzel, T. Hertog, M. Hewitson, K. Holley-Bockelmann, D. Hollington, M. Hueller, H. Inchauspe, P. Jetzer, N. Karnesis, C. Killow, A. Klein, B. Klipstein, N. Korsakova, S. L. Larson, J. Livas, I. Lloro, N. Man, D. Mance, J. Martino, I. Mateos, K. McKenzie, S. T. McWilliams, C. Miller, G. Mueller, G. Nardini, G. Nelemans, M. Nofrarias, A. Petiteau, P. Pivato, E. Plagnol, E. Porter, J. Reiche, D. Robertson, N. Robertson, E. Rossi, G. Russano, B. Schutz, A. Sesana, D. Shoemaker, J. Slutsky, C. F. Sopuerta, T. Sumner, N. Tamanini, I. Thorpe, M. Troebs, M. Vallisneri, A. Vecchio, D. Vetrugno, S. Vitale, M. Volonteri, G. Wanner, H. Ward, P. Wass, W. Weber, J. Ziemer, and P. Zweifel, Laser interferometer space antenna (2017), arXiv:1702.00786 [astro-ph.IM].

[4] A. Vecchio, Lisa observations of rapidly spinning massive black hole binary systems, Phys. Rev. D **70**, 042001 (2004).

[5] C. Cutler, Angular resolution of the lisa gravitational wave detector, Phys. Rev. D **57**, 7089 (1998).

[6] B. P. Abbott, R. Abbott, T. D. Abbott, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari, V. B. Adya, and et al., Multi-messenger observations of a binary neutron star merger, The Astrophysical Journal **848**, L12 (2017).

[7] 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).

[8] T. Dal Canton, A. H. Nitz, A. P. Lundgren, A. B. Nielsen, D. A. Brown, T. Dent, I. W. Harry, B. Krishnan, A. J. Miller, K. Wette, and et al., Implementing a search for aligned-spin neutron star-black hole systems with advanced ground based gravitational wave detectors, Physical Review D **90**, 10.1103/physrevd.90.082004 (2014).

[9] L. Allen, Masters project, `https://github.com/laurenceallen/Masters-Project` (2022).

[10] A. Toubiana, S. Marsat, S. Babak, J. Baker, and T. D. Canton, Parameter estimation of stellar-mass black hole binaries with LISA, Physical Review D **102**, 10.1103/physrevd.102.124037 (2020).

[11] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pe-

dregosa, P. van Mulbregt, and SciPy 1.0 Contributors, SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, Nature Methods **17**, 261 (2020).

[12] P. G. Krastev, Real-time detection of gravitational waves from binary neutron stars using artificial neural networks, Physics Letters B **803**, 135330 (2020).

[13] T. Robson, N. J. Cornish, and C. Liu, The construction and use of lisa sensitivity curves, Classical and Quantum Gravity **36**, 105011 (2019).

[14] M. Cabero, A. Lundgren, A. H. Nitz, T. Dent, D. Barker, E. Goetz, J. S. Kissel, L. K. Nuttall, P. Schale, R. Schofield, and D. Davis, Blip glitches in advanced LIGO data, Classical and Quantum Gravity **36**, 155010 (2019).

[15] A gentle introduction to cross-entropy for machine learning, `https://machinelearningmastery.com/cross-entropy-for-machine-learning/`, accessed: 13/04/2022.

[16] What is the difference between test and validation datasets?, `https://machinelearningmastery.com/difference-test-validation-datasets/`, accessed: 13/04/2022.

[17] F. Chollet *et al.*, Keras (2015).

[18] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, TensorFlow: Large-scale machine learning on heterogeneous systems (2015), software available from tensorflow.org.

[19] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization (2017), arXiv:1412.6980 [cs.LG].

[20] A. K. Mehta, C. K. Mishra, V. Varma, and P. Ajith, Accurate inspiral-merger-ringdown gravitational waveforms for nonspinning black-hole binaries including the effect of subdominant modes, Physical Review D **96**, 10.1103/physrevd.96.124010 (2017).

[21] M. Mapelli, Binary black hole mergers: Formation and populations, Frontiers in Astronomy and Space Sciences **7**, 38 (2020).