


# Matching Matched Filtering with Deep Networks for Gravitational-Wave Astronomy

Hunter Gabbard,<sup>\*</sup> Michael Williams, Fergus Hayes, and Chris Messenger

*SUPA, School of Physics and Astronomy, University of Glasgow, Glasgow G12 8QQ, United Kingdom*

 (Received 16 December 2017; revised manuscript received 12 February 2018; published 6 April 2018)

We report on the construction of a deep convolutional neural network that can reproduce the sensitivity of a matched-filtering search for binary black hole gravitational-wave signals. The standard method for the detection of well-modeled transient gravitational-wave signals is matched filtering. We use only whitened time series of measured gravitational-wave strain as an input, and we train and test on simulated binary black hole signals in synthetic Gaussian noise representative of Advanced LIGO sensitivity. We show that our network can classify signal from noise with a performance that emulates that of match filtering applied to the same data sets when considering the sensitivity defined by receiver-operator characteristics.

DOI: [10.1103/PhysRevLett.120.141103](https://doi.org/10.1103/PhysRevLett.120.141103)

*Introduction.*—The field of gravitational-wave astronomy has seen an explosion of compact binary coalescence detections over the past several years. The first of these were binary black hole detections [1–3] and more recently the advanced detector network made the first detection of a binary neutron star system [4]. This latter event was seen in conjunction with a gamma-ray burst [5–7] and multiple postmerger electromagnetic signatures [8]. These detections were made possible by the Advanced Laser Interferometer Gravitational Wave Observatory detectors, as well as the recent joint detections of GW170814 and GW170817 with Advanced Virgo [4,9]. Over the coming years many more such observations, including binary black hole (BBH), binary neutron stars, as well as other more exotic sources are likely to be observed on a more frequent basis. As such, the need for more efficient search methods will be more pertinent as the detectors increase in sensitivity.

The algorithms used by the search pipelines to make detections [10–12] are, in general, computationally expensive. The methods used are complex, sophisticated processes computed over a large parameter space using advanced signal processing techniques. The computational cost to run the search analysis is due to the large parameter space, as well as analysis of the high frequency components of the waveform where high data sample rates are required. Distinguishing noise from signal in these search pipelines is achieved, in part, using a technique known as template-based matched filtering.

Matched filtering uses a bank [12–16] of template waveforms [17–20] each with different component mass

and/or spin values. A template bank spans a large astrophysical parameter space since we do not know *a priori* the true gravitational-waves parameter values. Waveform models that cover the inspiral, merger, and ringdown phases of a compact binary coalescence are based on combining post-Newtonian theory [20–23], the effective-one-body formalism [24], and numerical relativity simulations [25].

Deep learning is a subset of machine learning, which has gained in popularity in recent years [26–31] with the rapid development of graphics-processing-unit technology. Some successful implementations of deep learning include image processing [26,32,33], medical diagnosis [34], and microarray gene expression classification [35]. There has also been some recent success in the field of gravitational-wave astronomy in the form of glitch classification [36–38] and notably for signal identification [39,40] where it was first shown that deep learning could be a detection tool [39]. Deep learning is able to perform analyses rapidly since the method's computationally intensive stage is precomputed during the training prior to the analysis of actual data [41]. This could result in low-latency searches that have the potential to be orders of magnitude faster than other comparable classification methods.

A deep learning algorithm is composed of stacked arrays of processing units, called neurons, which can be from one to several layers deep. A neuron acts as a filter, whereby it performs a transformation on an array of inputs. This transformation is a linear operation between the input array and the weight and bias parameters associated with the neuron. The resulting array is then typically passed to a nonlinear activation function to constrain the neuron output to be within a set range. Deep learning algorithms typically consist of an input layer, followed by one to several hidden layers and then one to multiple output neurons. The scalars produced from the output neurons can be used to solve classification problems, where each output neuron corresponds to the probability that an input sample is of a certain class.

---

*Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.*

In this Letter we investigate the simplest case of establishing whether a signal is present in the data or if the data contain only detector noise. We propose a deep learning procedure requiring only the raw data time series as input with minimal signal preprocessing. We compare the results of our network with the widely used matched-filtering technique and show how a deep learning approach can be pretrained using simulated data sets and applied in low latency to achieve the same sensitivity as established matched-filtering techniques.

*Simulation details.*—In order to make a clean comparison between the deep learning approach and matched filtering, we distinguish between two cases, BBH merger signals in additive Gaussian noise (signal + noise) and Gaussian noise alone (noise only). We choose to focus on BBH signals rather than including binary neutron star systems for the reason that BBH systems are higher mass systems and have shorter duration signals once the inspiralling systems have entered the Advanced LIGO frequency band. They typically then merge on the timescale of  $\mathcal{O}(1s)$ , allowing us to use relatively small data sets for this study.

The input data sets consist of “whitened” simulated gravitational-wave time series where the whitening process uses the detector noise power spectral density (PSD) to rescale the noise contribution at each frequency to have equal power. Our noise is initially generated from a PSD equivalent to the Advanced LIGO design sensitivity [42].

Signals are simulated using a library of gravitational-wave data analysis routines called LALSuite. We use the IMRPhenomD-type waveform [43,44], which models the inspiral, merger, and ringdown components of BBH gravitational-wave signals. We simulate systems with component black hole masses in the range from 5 to  $95M_{\odot}$ ,  $m_1 > m_2$ , with zero spin. Training, validation, and testing data sets contain signals drawn from an astrophysically motivated distribution where we assume  $m_{1,2} \sim \log m_{1,2}$  [45]. Each signal is given a random right ascension and declination assuming an isotropic prior on the sky, the polarization angle and phase are drawn from a uniform prior on the range  $[0, 2\pi]$ , and the inclination angle is drawn such that the cosine of inclination is uniform on the range  $[-1, 1]$ . The waveforms are then randomly placed within the time series such that the peak amplitude of each waveform is randomly positioned within the fractional range  $[0.75, 0.95]$  of the time series.

The waveform amplitude is scaled to achieve a predefined optimal signal-to-noise ratio (SNR) defined as

$$\rho_{\text{opt}}^2 = 4 \int_{f_{\text{min}}}^{\infty} \frac{|h(\tilde{f})|^2}{S_n(f)} df, \quad (1)$$

where  $h(\tilde{f})$  is the frequency domain representation of the gravitational-wave strain and  $S_n(f)$  is the single-sided detector noise PSD [46]. The simulated time series were chosen to be 1 s in duration sampled at 8192 Hz. Therefore,

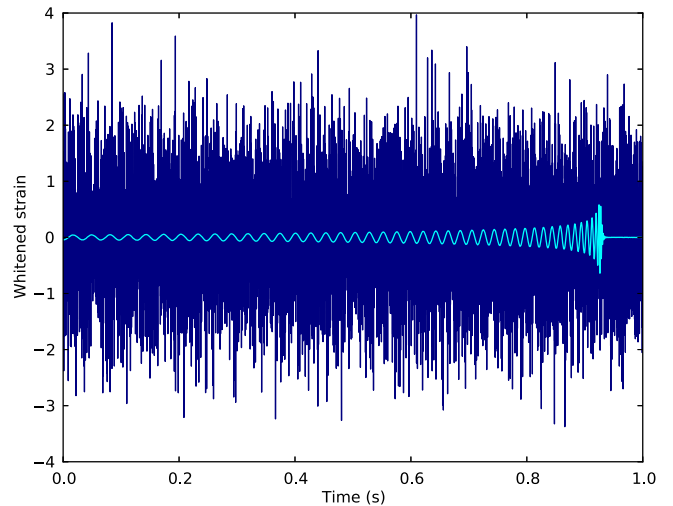


FIG. 1. A whitened noise-free time series of a BBH signal with component masses  $m_1 = 41.86M_{\odot}$  and  $m_2 = 6.65M_{\odot}$  with optimal SNR  $\rho_{\text{opt}} = 8$  (cyan). The dark blue time series shows the same gravitational-wave signal with additive whitened Gaussian noise of unit variance. This latter time series is representative of the data sets used to train, validate, and test the deep neural network.

we consider  $f_{\text{min}}$  as the frequency of the gravitational-wave signal at the start of the sample time series. An example time series can be seen in Fig. 1.

Because of the requirements of the matched-filtering comparison it was necessary to add padding to each time series so as to avoid nonphysical boundary artefacts from the whitening procedure. The Gaussianity of the noise and smoothness of the simulated advanced LIGO PSD allows the use of relatively short padding. Therefore, each 1 s time series has an additional 0.5 s of data prior to and after the signal. The signal itself has a Tukey window ( $\alpha = 1/8$ ) applied to truncate the signal content to the central 1 s. The convolutional neural network (CNN) approach only has access to this central 1 s of data. Similarly, the optimal SNR is computed considering only the central 1 s.

Supervised deep learning requires data sets to be subdivided into training, validation, and testing sets. Training sets are the data samples that the network learns from, the validation set allows the developer to verify that the network is learning correctly, and the test set is used to quantify the performance of the trained network. In a practical scenario the training and validation sets are used to train the network prior to data taking. This constitutes the vast majority of computational effort and is a procedure that needs to be computed only once. The trained network can then be applied to test data at a vastly reduced cost in comparison to the training stage [41]. Of the data set generated we use 90% of these samples for training, 5% for validation, and 5% for testing. A data set was generated for each predefined optimal SNR value ranging from 1–10 in integer steps.

Our training data sets contain  $5 \times 10^5$  independent time series with 50% containing signal + noise and 50% noise

only. For each simulated gravitational-wave signal (drawn from the signal parameter space) we generate 25 independent noise realizations from which 25 signal + noise samples are produced. This procedure is standard within machine-learning classification and allows the network to learn how to identify individual signals under different noise scenarios. Each noise-only sample consists of an independent noise realization and in total we therefore use 10000 unique waveforms in the  $m_1, m_2$  mass space. Each data sample time series is then represented in the form of a  $1 \times 8192$  pixel image with the gray-scale intensity of each pixel proportional to the gravitational-wave amplitude.

*The deep network approach.*—In our model, we use a variant of a deep learning algorithm called a CNN [47] composed of multiple layers. The input layer holds the raw pixel values of the sample image, which, in our case, is a one-dimensional time series vector. The weight and bias parameters of the network are also in one-dimensional vector form. Each neuron in the convolutional layer computes the convolution between the neuron’s weight vector and the outputs from the layer below it, and then the result is summed with the bias vector. Neuron weight vectors are updated through an optimization algorithm called back propagation [48]. Activation functions apply an elementwise nonlinear operation, rescaling their inputs onto a specific range and leaving the size of the previous layer’s output unchanged. Pooling layers perform a down-sampling operation along the spatial dimensions of their input. Finally we have a hidden layer connected to an output layer that computes the inferred class probabilities. These values are input to a loss function, chosen as the binary cross entropy [49], defined as

$$f(\theta) = -\sum_{i \in S} \log(\theta_i^S) - \sum_{i \in N} \log(\theta_i^N), \quad (2)$$

where  $\theta_i^{S/N}$  is the predicted probability of class signal + noise ( $S$ ) or noise-only ( $N$ ) for the  $i$ th training sample. The loss function is minimized when input data samples are assigned the correct class with the highest confidence.

In order to optimize a network, multiple hyperparameters must be tuned. We define hyperparameters as parameters we are free to choose. Such parameters include the number and type of network layers, the number of neurons within each layer, size of the neuron weight vectors, max-pooling parameters, type of activation functions, preprocessing of input data, learning rate, and the application (or otherwise) of specific deep learning techniques. We begin the process with the simplest network that provides a discernible level of effective classification. In most cases this consists of an input, convolutional, hidden, and logistic output layer. The optimal network structure was determined through multiple tests and tunings of hyperparameters by means of trial and error.

During the training stage an optimization function (back propagation) works by computing the gradient of the loss function [Eq. (2)], then attempting to minimize that loss function. The errors are then propagated back through the network while also updating the weight and bias terms accordingly. Back propagation is done over multiple iterations called epochs. We use adaptive moment estimation with incorporated Nesterov momentum [50] with a learning rate of 0.002,  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ ,  $\epsilon = 10^{-8}$  and a momentum schedule of 0.004. We outline the structure of the final neural network architecture in Table I.

The final ranking statistic that we extract from the CNN analysis is taken from the output layer, composed of two neurons, where each neuron produces a probability value between 0 and 1 with their sum being unity. Each neuron gives the inferred probability that the input data belong to the noise or signal + noise class, respectively. The computational time spent on training the network for each SNR is  $\mathcal{O}(1)$  hour on a single GPU. This one-time cost can be compared to the  $\mathcal{O}(1s)$  spent applying the trained network to all 25,000 1 s test data samples also using a single GPU. Therefore, at the point of data taking this particular analysis can be run  $10^4$  times faster than real time.

*Applying matched filtering.*—In order to establish the power of the deep learning approach we must compare our results to the standard matched-filtering process used in the

TABLE I. The optimized network consisting of six convolutional layers ( $C$ ), followed by three hidden layers ( $H$ ). Max pooling is performed on the first, fifth, and eighth layer, whereas dropout is only performed on the two hidden layers. Each layer uses an exponential linear unit (Elu) activation function (with range  $[-1, \infty)$ ) while the last layer uses a Softmax (SMax) activation function in order to normalize the output values to be between 0 and 1 so as to give a probability value for each class.

Parameter (Option)	Layer								
	1	2	3	4	5	6	7	8	9
Type	$C$	$C$	$C$	$C$	$C$	$C$	$H$	$H$	$H$
No. Neurons	8	8	16	16	32	32	64	64	2
Filter size	64	32	32	16	16	16	Not applicable	Not applicable	Not applicable
Max pool size	Not applicable	8	Not applicable	6	Not applicable	4	Not applicable	Not applicable	Not applicable
Drop out	0	0	0	0	0	0	0.5	0.5	0
Activation function	Elu	Elu	Elu	Elu	Elu	Elu	Elu	Elu	SMax

detection of compact binary coalescence signals [46,51]. The ranking statistic used in this case is the matched-filter SNR numerically maximized over arrival time, phase, and distance. By first defining the noise weighted inner product as a function of a time shift  $\Delta t$  between the arrival time of the signal and the template,

$$(a|b)[\Delta t] = 4 \int_{f_{\min}}^{\infty} \frac{\tilde{a}(f)\tilde{b}^*(f)}{S_n(f)} e^{2\pi i f \Delta t} df, \quad (3)$$

we can construct the matched-filter SNR as

$$\rho^2[\Delta t] = \frac{(s|h)^2[\Delta t] + i(s|h)^2[\Delta t]}{(h|h)}, \quad (4)$$

where  $s$  is the data containing noise and a potential signal, and  $h$  is the noise-free gravitational-wave template [52]. For a given template this quantity is efficiently computed using the FFT and the SNR time series maximized over  $\Delta t$ . The subsequent step is to further numerically maximize this quantity over a collection of component mass combinations. In this analysis a comprehensive template bank is generated in the  $m_1, m_2$  mass space covering our predefined range of masses. We use a maximum mismatch of 3% and a lower frequency cutoff of 20 Hz using the PyCBC geometric nonspinning template bank generation tool [10,53]. This template bank contained 8056 individual templates.

When generating a SNR time series for an input data set we select  $f_{\min}$  according to the conservative case (lowest  $f_{\min}$ ) in which the signal merger occurs at the 0.95 fraction of 1 s time series. We therefore select only maximized SNR time series values recovered from within the [0.75, 0.95] fractional range since this is the parameter space on which the CNN has been trained. For the practical computation of the matched-filtering analysis we take each of the data samples from the testing data set to compute the matched-filter ranking statistic.

*Results.*—After tuning the multiple hyperparameters (Table I) and training the neural network, we present the results of our CNN classifier on a noise versus signal + noise sample set. With values of statistics now assigned to each test data sample from both the CNN and matched-filtering approaches, and having knowledge of the true class associated with each sample, we may now construct receiver operator characteristic (ROC) curves.

In Fig. 2 we compare our CNN results to that of matched filtering. Given the ranking statistic from a particular analysis and defining a parametric threshold value on that statistic we are able to plot the fraction of noise samples incorrectly identified as signals (false alarm probability) versus the fraction of signal samples correctly identified (true alarm probability). These curves are defined as ROC curves and a ranking statistic is deemed superior to another if at a given false alarm probability it achieves a higher detection probability. Our results show that the CNN

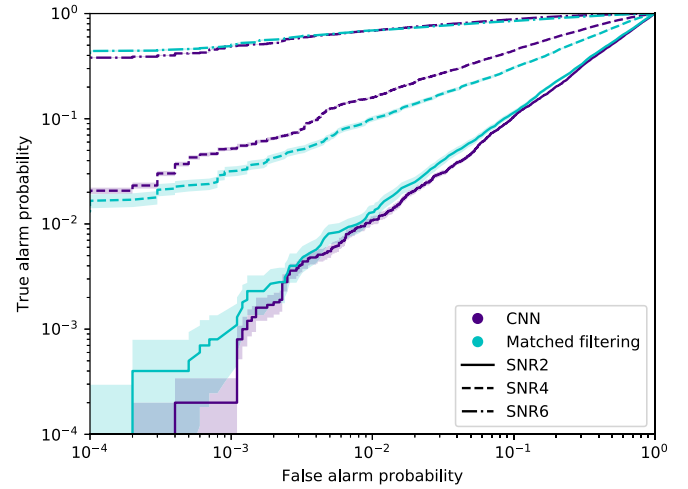


FIG. 2. The ROC curves for test data sets containing signals with optimal SNR,  $\rho_{\text{opt}} = 2, 4, 6$ . We plot the true alarm probability versus the false alarm probability estimated from the output of the CNN (purple) and matched-filtering (cyan) approaches. Uncertainties in the true alarm probability correspond to  $1\text{-}\sigma$  bounds assuming a binomial distribution.

approach closely matches the sensitivity of matched filtering for all test data sets across the range of false alarm probabilities explored in this analysis [54].

We can make an additional direct comparison between approaches by fixing a false alarm probability and plotting the corresponding true alarm probability versus the optimal SNR of the signals in each test data set. We show these efficiency curves in Fig. 3 at false alarm probabilities

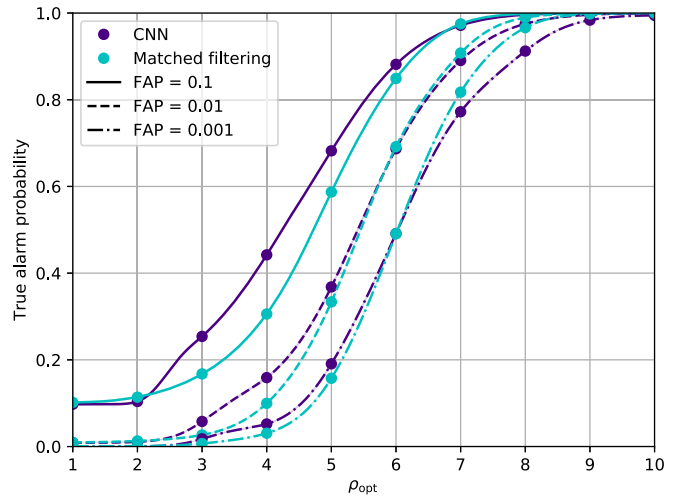


FIG. 3. Efficiency curves comparing the performance of the CNN and matched-filtering approaches for false alarm probabilities  $10^{-1}$  (solid),  $10^{-2}$  (dashed),  $10^{-3}$  (dot-dashed). The true alarm probability is plotted as a function of the optimal SNR for the CNN (purple) and the matched-filtering (cyan) analyses. Solid dots indicate at which SNR values analyses were performed and line thicknesses are indicative of the statistical uncertainties in the curves.



$10^{-1}$ ,  $10^{-2}$ ,  $10^{-3}$  for both the CNN and matched-filtering approaches. We again see very good agreement between the approaches at all false alarm probabilities with the CNN sensitivity exceeding that of the matched-filter approach at low SNR and high false alarm probability. Conversely we see the matched-filter sensitivity marginally exceeds the CNN at high SNR and low false alarm probability. This latter discrepancy could be mitigated by increasing the number of training samples.

*Conclusions.*—We have demonstrated that deep learning, when applied to gravitational-wave time series data, is able to closely reproduce the results of a matched-filtering analysis in Gaussian noise. We employ a deep convolutional neural network with rigorously tuned hyperparameters and produce an output that returns a ranking statistic equivalent to the inferred probability that data contain a signal. Matched-filtering analyses are often described as the optimal approach for signal detection in Gaussian noise. By building a neural network that is capable of reproducing this optimality we answer a fundamental question regarding the applicability of neural networks for gravitational-wave data analysis.

In practice, searches for transient signals in gravitational-wave data are strongly affected by non-Gaussian noise artefacts. To account for this, standard matched-filtering approaches are modified to include carefully chosen changes to the ranking statistic [55,56] together with the excision of poor quality data [57,58]. Our analysis represents a starting point from which a deep network can be trained on realistic non-Gaussian data. Since the claim of matched-filtering optimality is applicable only in the Gaussian noise case, there exists the potential for deep networks to exceed the sensitivity of existing matched-filtering approaches in real data.

In this work we have presented results for BBH mergers; however, this method could be applied to other merger types, such as binary neutron star and neutron star-black hole signals. This supervised learning approach can also be extended to other well-modeled gravitational-wave targets such as the continuous emission from rapidly rotating nonaxisymmetric neutron stars [59]. Finally we mention the possibilities for parameter estimation [60] where in the simplest cases an output regression layer can return point estimates of parameter values. As was exemplified in the case of GW170817, rapid detection confidence coupled with robust and equally rapid parameter estimates is critical for gravitational-wave multimessenger astronomy.

We acknowledge valuable input from the LIGO-Virgo Collaboration specifically from T. Dent, R. Reinhard, I. Siong Heng, M. Cavalgia, and the compact binary coalescence and machine-learning working groups. The authors also gratefully acknowledge the Science and Technology Facilities Council of the United Kingdom. C.M. is supported by the Science and Technology Research Council (Grant No. ST/L000946/1).

\*Corresponding author.

h.gabbard.1@research.gla.ac.uk

- [1] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), *Phys. Rev. Lett.* **116**, 061102 (2016).
- [2] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), *Phys. Rev. Lett.* **116**, 241103 (2016).
- [3] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), *Phys. Rev. Lett.* **118**, 221101 (2017).
- [4] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), *Phys. Rev. Lett.* **119**, 161101 (2017).
- [5] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), *Astrophys. J.* **848**, L13 (2017).
- [6] A. Goldstein *et al.*, *Astrophys. J.* **848**, L14 (2017).
- [7] V. Savchenko *et al.*, *Astrophys. J.* **848**, L15 (2017).
- [8] B. P. Abbott *et al.* (LIGO Scientific Collaboration, Virgo Collaboration, F. GBM, INTEGRAL, IceCube Collaboration, AstroSat Cadmium Zinc Telluride Imager Team, IPN Collaboration, The Insight-Hxmt Collaboration, ANTARES Collaboration, The Swift Collaboration, AGILE Team, The 1M2H Team, The Dark Energy Camera GW-EM Collaboration, the DES Collaboration, The DLT40 Collaboration, GRAWITA, GRAVitational Wave Inaf TeAm, The Fermi Large Area Telescope Collaboration, ATCA, A. Telescope Compact Array, ASKAP, A. SKA Pathfinder, Las Cumbres Observatory Group, OzGrav, DWF, AST3, CAASTRO Collaborations, The VINROUGE Collaboration, MASTER Collaboration, J-GEM, GROWTH, JAGWAR, C. NRAO, TTU-NRAO, NuSTAR Collaborations, Pan-STARRS, The MAXI Team, T. Consortium, KU Collaboration, N. Optical Telescope, ePESSTO, GROND, T. Tech University, SALT Group, TOROS, Transient Robotic Observatory of the South Collaboration, The BOOTES Collaboration, MWA, M. Widefield Array, The CALET Collaboration, IKI-GW Follow-up Collaboration, H. E. S. S. Collaboration, LOFAR Collaboration, LWA, L. Wavelength Array, HAWC Collaboration, The Pierre Auger Collaboration, ALMA Collaboration, Euro VLBI Team, Pi of the Sky Collaboration, The Chandra Team at McGill University, DFN, D. Fireball Network, ATLAS, H. Time Resolution Universe Survey, RIMAS, RATIR, and S. South Africa/MeerKAT), *Astrophys. J.* **848**, L12 (2017).
- [9] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), *Phys. Rev. Lett.* **119**, 141101 (2017).
- [10] S. A. Usman *et al.*, *Classical Quantum Gravity* **33**, 215004 (2016).
- [11] K. Cannon, R. Cariou, A. Chapman, M. Crispin-Ortuzar, N. Fotopoulos, M. Frei, C. Hanna, E. Kara, D. Keppel, L. Liao, S. Privitera, A. Searle, L. Singer, and A. Weinstein, *Astrophys. J.* **748**, 136 (2012).
- [12] 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, K. Wiesner, and J. L. Willis, *Phys. Rev. D* **90**, 082004 (2014).
- [13] D. A. Brown, I. Harry, A. Lundgren, and A. H. Nitz, *Phys. Rev. D* **86**, 084017 (2012).
- [14] T. D. Canton and I. W. Harry, [arXiv:1705.01845](https://arxiv.org/abs/1705.01845).
- [15] I. W. Harry, B. Allen, and B. S. Sathyaprakash, *Phys. Rev. D* **80**, 104014 (2009).

- [16] P. Ajith, N. Fotopoulos, S. Privitera, A. Neunzert, N. Mazumder, and A. J. Weinstein, *Phys. Rev. D* **89**, 084041 (2014).
- [17] B. S. Sathyaprakash and S. V. Dhurandhar, *Phys. Rev. D* **44**, 3819 (1991).
- [18] A. Taracchini, A. Buonanno, Y. Pan, T. Hinderer, M. Boyle, D. A. Hemberger, L. E. Kidder, G. Lovelace, A. H. Mroué, H. P. Pfeiffer, M. A. Scheel, B. Szilágyi, N. W. Taylor, and A. Zenginoglu, *Phys. Rev. D* **89**, 061502 (2014).
- [19] S. Privitera, S. R. P. Mohapatra, P. Ajith, K. Cannon, N. Fotopoulos, M. A. Frei, C. Hanna, A. J. Weinstein, and J. T. Whelan, *Phys. Rev. D* **89**, 024003 (2014).
- [20] L. Blanchet, *Living Rev. Relativity* **17**, 2 (2014).
- [21] K. G. Arun, A. Buonanno, G. Faye, and E. Ochsner, *Phys. Rev. D* **84**, 049901(E) (2011).
- [22] A. Buonanno, B. R. Iyer, E. Ochsner, Y. Pan, and B. S. Sathyaprakash, *Phys. Rev. D* **80**, 084043 (2009).
- [23] C. K. Mishra, A. Kela, K. G. Arun, and G. Faye, *Phys. Rev. D* **93**, 084054 (2016).
- [24] A. Buonanno and T. Damour, *Phys. Rev. D* **59**, 084006 (1999).
- [25] F. Pretorius, *Phys. Rev. Lett.* **95**, 121101 (2005).
- [26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, in *Advances in Neural Information Processing Systems 25*, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Curran Associates, Inc., 2012), pp. 1097–1105.
- [27] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, [arXiv:1406.2661](https://arxiv.org/abs/1406.2661).
- [28] K. Simonyan and A. Zisserman, [arXiv:1409.1556](https://arxiv.org/abs/1409.1556).
- [29] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, [arXiv:1412.7062](https://arxiv.org/abs/1412.7062).
- [30] M. D. Zeiler and R. Fergus, [arXiv:1311.2901](https://arxiv.org/abs/1311.2901).
- [31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, [arXiv:1409.4842](https://arxiv.org/abs/1409.4842).
- [32] R. Zhang, P. Isola, and A. A. Efros, [arXiv:1603.08511](https://arxiv.org/abs/1603.08511).
- [33] A. Karpathy and L. Fei-Fei, [arXiv:1412.2306](https://arxiv.org/abs/1412.2306).
- [34] I. Kononenko, *Artif. Intell. Med.* **23**, 89 (2001).
- [35] M. Pirooznia, J. Y. Yang, M. Q. Yang, and Y. Deng, *BMC Genomics* **9**, S13 (2008).
- [36] N. Mukund, S. Abraham, S. Kandhasamy, S. Mitra, and N. S. Philip, *Phys. Rev. D* **95**, 104059 (2017).
- [37] M. Zevin, S. Coughlin, S. Bahaadini, E. Besler, N. Rohani, S. Allen, M. Cabero, K. Crowston, A. K. Katsaggelos, S. L. Larson, T. K. Lee, C. Lintott, T. B. Littenberg, A. Lundgren, C. sterlund, J. R. Smith, L. Trouille, and V. Kalogera, *Classical Quantum Gravity* **34**, 064003 (2017).
- [38] D. George, H. Shen, and E. A. Huerta, [arXiv:1706.07446](https://arxiv.org/abs/1706.07446).
- [39] D. George and E. Huerta, *Phys. Lett. B* **778**, 64 (2018).
- [40] T. Gebhard, N. Kilbertus, G. Parascandolo, I. Harry, and B. Schölkopf, in *Workshop at the 31st Conference on Neural Information Processing Systems (NIPS), Long Beach, CA* (2017).
- [41] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, MA, 2016).
- [42] B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. X. Adhikari *et al.*, *Living Rev. Relativity* **19**, 1 (2016).
- [43] S. Husa, S. Khan, M. Hannam, M. Pürrer, F. Ohme, X. J. Forteza, and A. Bohé, *Phys. Rev. D* **93**, 044006 (2016).
- [44] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. J. Forteza, and A. Bohé, *Phys. Rev. D* **93**, 044007 (2016).
- [45] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), *Phys. Rev. X* **6**, 041015 (2016).
- [46] B. Allen, W. G. Anderson, P. R. Brady, D. A. Brown, and J. D. E. Creighton, *Phys. Rev. D* **85**, 122006 (2012).
- [47] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, *Proc. IEEE* **86**, 2278 (1998).
- [48] Y. LeCun, L. Bottou, G. B. Orr, and K. R. Müller, in *Neural Networks: Tricks of the Trade*, edited by G. B. Orr and K.-R. Müller (Springer Berlin Heidelberg, Berlin, Heidelberg, 1998), pp. 9–50.
- [49] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, [arXiv:1603.04467](https://arxiv.org/abs/1603.04467).
- [50] T. Dozat, *ICLR Workshop, San Juan, Puerto Rico* (2016).
- [51] S. Babak *et al.*, *Phys. Rev. D* **87**, 024033 (2013).
- [52] S. Babak, R. Balasubramanian, D. Churches, T. Cokelaer, and B. S. Sathyaprakash, *Classical Quantum Gravity* **23**, 5477 (2006).
- [53] A. Nitz *et al.*, Pycbc software (2017).
- [54] We are limited to a minimal false alarm probability of  $\sim 10^{-4}$  due to the limited number of testing samples used.
- [55] B. Allen, *Phys. Rev. D* **71**, 062001 (2005).
- [56] A. H. Nitz, T. Dent, T. D. Canton, S. Fairhurst, and D. A. Brown, *Astrophys. J.* **849**, 118 (2017).
- [57] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), *Classical Quantum Gravity* **35**, 065010 (2018).
- [58] B. P. Abbott *et al.*, *Classical Quantum Gravity* **33**, 134001 (2016).
- [59] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), [arXiv:1707.02669](https://arxiv.org/abs/1707.02669).
- [60] D. George and E. A. Huerta, *Phys. Rev. D* **97**, 044039 (2018).