Early warning of coalescing neutron-star and neutron-star-black-hole binaries from the nonstationary noise background using neural networks

Hang Yu*

TAPIR, Walter Burke Institute for Theoretical Physics, California Institute of Technology, MC 350-17, Pasadena, California 91125, USA

Rana X. Adhikari and Ryan Magee

LIGO Laboratory, California Institute of Technology, MC 100-36, Pasadena, California 91125, USA

Surabhi Sachdev

Department of Physics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA and Institute for Gravitation and the Cosmos, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

Yanbei Chen

TAPIR, Walter Burke Institute for Theoretical Physics, California Institute of Technology, Mailcode 350-17, Pasadena, California 91125, USA

(Received 20 April 2021; accepted 4 August 2021; published 7 September 2021)

The success of the multimessenger astronomy relies on gravitational-wave observatories like LIGO and Virgo to provide prompt warning of merger events involving neutron stars (including both binary neutron stars and neutron-star-black-hole binaries), which further depends critically on the low-frequency sensitivity of LIGO as a typical binary neutron star stays in this band for minutes. However, the current sub-60 Hz sensitivity of LIGO has not yet reached its design target and the excess noise can be more than an order of magnitude below 20 Hz. It is limited by nonlinearly coupled noises from auxiliary control loops which are also nonstationary, posing challenges to realistic early warning pipelines. Nevertheless, machine-learning-based neural networks provide ways to simultaneously improve the low-frequency sensitivity and mitigate its nonstationarity, and detect the real-time gravitational-wave signal with a very short computational time. We propose to achieve this by inputting both the main gravitational-wave readout and key auxiliary witnesses to a compound neural network. Using simulated data with characteristic representing the real LIGO detectors, our machine-learning-based neural networks can reduce nonlinearly coupled noise by about a factor of 5 and allows a typical binary neutron star (neutron-star black hole) to be detected 100 s (10 s) before the merger at a distance of 40 Mpc (160 Mpc). If one can further reduce the noise to the fundamental limit, our neural networks can achieve detection out to a distance of 80 and 240 Mpc for binary neutron stars and neutron-star-black-hole binaries, respectively. It thus demonstrates that utilizing machine-learning-based neural networks is a promising direction for the timely detection of the coalescence of electromagnetically bright LIGO/Virgo sources.

DOI: 10.1103/PhysRevD.104.062004

I. INTRODUCTION

The current generation of ground-based gravitational-wave interferometers [1–3] firmly established a new way to observe our cosmos. Since the first detection of gravitational waves (GWs) from a binary black hole (BBH) merger [4], Advanced LIGO (aLIGO [1]) and Advanced Virgo [2] have gone on to document dozens of gravitational-wave candidates [5,6] that have been confirmed and added to by the broader astrophysical community [7–14]. One of the most spectacular discoveries made by Advanced LIGO and Virgo is the first observed binary neutron star (BNS) coalescence, GW170817. GW170817 was jointly detected in low-latency in gravitational waves [15] and by Fermi-GBM in gamma rays [16]. The subsequent discovery and followup of kilonova AT 2017gfo led to a concerted followup effort across the electromagnetic (EM) spectrum [17]. The resulting multimessenger observations enabled an abundance of new science: constraints on the maximum NS mass [18], better understandings of neutron star mode coupling and equation of state [19–21], as well as tests of general relativity [22].

*hangyu@caltech.edu

Despite the successes surrounding GW170817, there is still much to be learned about compact binary mergers containing at least one neutron star. In particular, there are various astrophysical processes that can generate precursor and/or early stage signals that are yet to be detected. For example, tidal interactions might shatter crusts of neutron stars and lead to short gamma-ray burst [23]. The property of the final merger product may be better revealed with prompt x-ray and optical observations [24]. In the radio band, precursor magnetosphere interactions might cause radio emissions [25,26] and could be a potential mechanism leading to fast radio bursts [27,28]. See, e.g., Ref. [29] for further discussions on potential early warning signals as well as a nice summary of the follow-up capacity of various EM observatories.

To detect the prompt signatures of these processes, LIGO and Virgo would need to be able to identify the existence of a GW event and then determine its sky location in a timely manner. This is especially important for binaries where at least one component is a neutron star. The GW alert for GW170817 was not sent out until ∼40 minutes after the merger and the sky location was not released until another 4 hours later [17]; in principle, this information can be obtained minutes prior to the final merger as a typical BNS event will stay in the sensitivity band of LIGO and Virgo for minutes if it can be detected at a GW frequency of around 20 Hz [29].

There are presently four low-latency, matched-filter based pipelines that produce near real-time gravitational-wave alerts for BNS and NSBH mergers: GSTLAL [30–32], PYCBC [33], MBTA [34], and SPIIR [35,36]. Several of these pipelines have already developed analyses capable of early warning detection [29,30,36]. See also Ref. [37] for a summary of current efforts carried out by the LIGO and Virgo collaborations during the second observing run to low-latency warnings. The prospect of premerger detection is ultimately limited by latencies surrounding data acquisition, handling, and analysis. Ref. [38] recently demonstrated that even at present latencies, the LIGO-Virgo collaboration is capable of identifying, localizing, and broadcasting GW candidates prior to merger.

Machine-learning (ML) based convolutional neural networks (CNNs) offer yet another attractive alternative to achieve the early warning of BNSs/NSBHs. Instead of individually computing the overlap between a time series of GW readout and each waveform template from a large template bank, a trained CNN would only need to do the computation once to predict the existence and the property of the source. It can therefore serve as the first step for existing pipelines and further accelerate their computational efficiency.

Indeed, various authors have considered the possibility of detecting GW events using ML-based CNNs. Refs. [39–43] showed that it is possible to input real-time GW readout and then use CNNs to detect massive black hole binaries (BBHs) and later Refs. [44,45] considered the possibility of detecting BNSs with longer signal duration. Recently, Ref. [46] further considered detecting BNS events tens of seconds prior to the final merger. In addition to using full time series of the data, Refs. [47,48] also considered the possibility of achieving the early warning by searching for chirp patterns in the spectrograms.

However, many of the the analyses above assume a stationary Gaussian noise background and often at the designed sensitivity of aLIGO [49]. While this is a decent approximation in the $f > 100$ Hz frequency band, at the lower frequencies that matter most for the early warning the detector noise not only exceeds the designed level by orders of magnitude but also exhibits nonstationarity [10,50,51]. Therefore, it would be crucial to take into account these features of realistic detector noise in order to design a CNN to achieve early warning in practice.

Our work thus extends the field by considering the detection of GW events from a nonstationary noise background representative of realistic LIGO detectors. In addition to the main GW readout, we further show that in principle one can also input to the CNN some key auxiliary channels witnessing the sources of contamination to hence enhance the low-frequency sensitivity. As the contamination typically involves nonlinear and nonstationary coupling mechanisms, it cannot be mitigated by standard signal processing techniques assuming linear and stationary noise coupling. We demonstrate that, with CNNs involving nonlinear activations, one can nonetheless tackle the challenges of nonlinearity and nonstationarity and achieve simultaneous noise mitigation and signal detection both in real time.

The rest of the paper is organized as follows. In Sec. II we briefly overview the LIGO sensitivity during its third observing run (O3) and discuss the main source of contamination to the low frequency band of interest. In Sec. III we then describe the properties of the GW signal. This is followed by Sec. IV in which we provide the details of the construction of training of our early warning CNN. Specifically, we describe the preparation of our training datasets in Sec. IV A and then in Secs. IV B–IV D the procedures we adopt for the network training. The performance of our CNN is examined in Sec. V. Lastly, we conclude and discuss our results in Sec. VI.

II. OVERVIEW OF LIGO SENSITIVITY

While LIGO has achieved a great success, its sensitivity can still be further improved as we demonstrate in Fig. 1. Here the orange trace is the representative sensitivity at the LIGO Hanford observatory during the third observing run (O3) [51] and the red trace is its fundamental limit set by quantum and thermal fluctuations at the O3 configuration (which actually closely matches the designed sensitivity of aLIGO [1]). While the two traces overlaps at $f \gtrsim 100$ Hz, at lower frequencies the excess noise can be significant.

At 30 Hz (20 Hz), the fundamental limit is a factor of 3 (10) below the current sensitivity, indicating a large room of improvement. Opening up the sensitivity in the low-frequency band can be especially rewarding for multimessenger astronomy and astrophysics, as it allows a coalescing BNS (whose strain we show in the purple trace) to be detected at a lower frequency and hence a much earlier time prior to the merger; see the discussion in Sec. III.

A major source of contamination to the current low-frequency sensitivity is the control noises of auxiliary degrees of freedom [51] (see also, e.g., Refs. [50,53]). For instance, while it is necessary to engage an active angular control system to maintain the alignment of test masses at below a few Hz during LIGO's observation, the system also inevitably feeds back the sensing noise in the 10–30 Hz band and causes excess angular perturbation $\theta(t)$. The angular perturbation further couples with the off-pivot beam spot motion and leads to a longitudinal displacement that contaminates the GW readout as

$$\delta x(t) = x_{\rm spot}(t)\theta(t) = [x^{\rm (dc)}_{\rm spot} + x^{\rm (ac)}_{\rm spot}(t)]\theta(t). \qquad (1)$$

where in the second equality, we have decomposed the spot motion into a direct current (dc) part $x^{\rm (dc)}_{\rm spot}$ and an alternating current (ac) part $x^{\rm (ac)}_{\rm spot}$. Such a contamination can be mitigated by both online feed-forward cancellation and off-line signal regressions (see, e.g., Ref. [54]). However, standard signal processing techniques [such as computing the Wiener filter from $\theta(t)$ to $\delta x(t)$] assume the coupling is linear and stationary and therefore can only remove the constant coupling part $\propto x^{\rm (dc)}_{\rm spot}$ but not the fluctuating piece $\propto x^{\rm (ac)}_{\rm spot}(t)$. In fact, it is exactly due to the temporal variability of couplings like $x^{\rm (ac)}_{\rm spot}(t)$ that the current LIGO noise background at low frequencies is nonstationary [50,51], dramatically complicating the data analysis process [10].

Furthermore, there are no direct witnesses for the spot position on the test masses, $x^{\rm (ac)}_{\rm spot}(t)$, over the entire frequency band of interests. Instead, it has to be reconstructed from multiple sensors through complicated geometrical conversions as well as signal filtering and blending, with each step subject to its own calibration uncertainties.

Nonetheless, CNN using ML offers an attractive way to tackle this problem. By inputting sufficient auxiliary witness channels, a deep CNN [55] would be able to figure out the correct, frequency-dependent combinations of the witness that reconstructs the contamination. Moreover, as each layer typically involves a nonlinear activation function, it would be able to capture nonlinear couplings like Eq. (1) that classical, linear signal processing techniques fail (see also Refs. [56–59] for some recent efforts to mitigate nonlinear noises in the LIGO detectors). Furthermore, as a CNN is trained directly on time series, it is especially suitable to be implemented in real time and has the potential to be integrated into a low-latency detection pipeline.

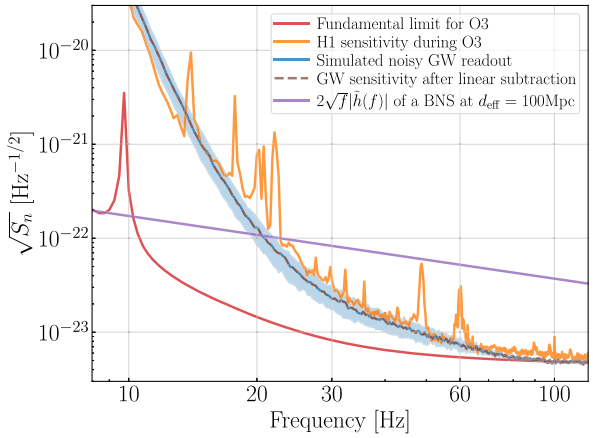

FIG. 1. Comparison of the strain sensitivities. The orange trace is the realistic sensitivity of the LIGO Hanford detector during O3 [51] and the red trace its fundamental limit. We simulate noise according to the mechanism described in Eq. (1) and a typical realization is shown in the blue trace. Note that it is in fact nonstationary and its spectrum can vary within the blue-shaded band. Because of the nonlinear nature of the noise coupling, a linear, coherence-based subtraction cannot mitigate the noise as shown in the dashed-brown trace. As a reference, we also show the strain of a typical coalescing BNS event with $M_1 = M_2 = 1.4\,M_\odot$ and located at an effective distance [52] of $d_{\rm eff} = 100$ Mpc in the purple trace. The event can be detected at a much lower frequency (hence a much earlier time) if the contamination at low frequencies can be mitigated.

To demonstrate this point, we simulate excess noise according to the mechanism described in Eq. (1) and combine it with the fundamental limit to form the blue trace in Fig. 1. The $x_{\rm spot}(t)$ and $\theta(t)$ as well their witness channels are simulated with similar characteristics as in realistic LIGO detectors, with one exception that we reduce the roll off of $\theta(t)$ in the 25–80 Hz band so that the entire O3 sensitivity can be approximated by this mechanism (see Sec. IV A for more details). In reality, the noise in the 25–80 Hz are dominated by other noise sources [51] which we ignore here for simplicity. Note that we have assumed that the constant coupling piece is already removed (i.e., $x^{\rm (dc)}_{\rm spot} = 0$), and linear subtraction cannot further mitigate the contamination. This is illustrated by the brown-dashed curve in Fig. 1 where we compute the multi-input-single-output coherence between all the auxiliary witness and the gain GW channel and then subtract out the coherent component in the frequency domain. To further simulate the nonstationarity on timescales longer than the length of each realization (256 s), we allow the overall root-mean-square (rms) of $x_{\rm spot}(t)$ to be a random variable. Thus the blue trace in Fig. 1 is just the amplitude spectral density (ASD) of a typical realization; the noises we simulate in fact have their spectra vary within the shaded blue region (see also Fig. 6).

III. GW SIGNAL

Having described the noise and how we may use ML techniques to mitigate it, we now turn to the discussion about detecting the astrophysical GW events. Specifically, our goal is to detect a GW event minutes before the final merger and further classify its type (NS vs BH) to assist the EM follow up strategies.

For the early warning purpose, we can approximate the waveform using only the leading-order quadrupole formula and write (with $G = c = 1$; see, e.g., [60])

$$h(t) = \frac{\mathcal{A}}{d}\mathcal{M}_c^{5/4}\left(\frac{5}{t_m}\right)^{1/4}\cos[\Phi(t)], \tag{2}$$

$$\Phi(t) = -2\left(\frac{t_m}{5\mathcal{M}_c}\right)^{5/8} + \Phi_c, \tag{3}$$

where $t_m = t_c - t$ is the time to merger and t_c and Φ_c are time and phase of coalescence. The time t_m is further related to the GW frequency f according to

$$t_m(f) = 86.7\ \mathrm{s}\left(\frac{\mathcal{M}_c}{1.22\ M_\odot}\right)^{-5/3}\left(\frac{f}{25\ \mathrm{Hz}}\right)^{-8/3}, \tag{4}$$

$$f(t_m) = 23.7\ \mathrm{Hz}\left(\frac{\mathcal{M}_c}{1.22\ M_\odot}\right)^{-5/8}\left(\frac{t_m}{100\ \mathrm{s}}\right)^{-3/8}. \tag{5}$$

In this work, we do not include the detailed antenna responses (which are encoded in the quantity $\mathcal{A}$) nor the joint detection by multiple detectors. Instead, we set $\mathcal{A} = 1$ and simply replace the distance to the source d in Eq. (2) by $d_{\mathrm{eff}}/\sqrt{N_{\mathrm{det}}}$, where $d_{\mathrm{eff}} \simeq 2.3d$ is the averaged effective distance [52,61] and N_{det} is the number of detectors observing. For the rest of the work, we will use $N_{\mathrm{det}} = 3$ as the default value.

From the above equations we see that the waveform depends only on one intrinsic parameter of the source, the chirp mass $\mathcal{M}_c$, defined as

$$\mathcal{M}_c = \frac{(M_1M_2)^{3/5}}{(M_1+M_2)^{1/5}}, \tag{6}$$

with $M_{1,2}$ the component masses. Therefore, we put a GW event into three categories according to its chirp mass.

We define the first category as events with $1\ M_\odot \leq \mathcal{M}_c < 1.8\ M_\odot$ and label such an event as a "BNS" event. Note that a BNS with $M_1 = M_2 = 2\ M_\odot$ (which is the mass of the heaviest NS observed to date [62]) will have $\mathcal{M}_c = 1.74\ M_\odot$. Therefore, we would expect that most astrophysical BNS events will fall into this category (including GW170817with $\mathcal{M}_c = 1.19\ M_\odot$ [15] and GW190425 with $\mathcal{M}_c = 1.44\ M_\odot$ [63]).

We also define a "BBH" category as sources with $4.5\ M_\odot \leq \mathcal{M}_c < 10\ M_\odot$. The lower boundary is inspired by noticing a BBH with $M_1 = M_2 = 5\ M_\odot$ would have $\mathcal{M}_c = 4.35\ M_\odot$. In principle, the upper boundary of $\mathcal{M}_c < 10\ M_\odot$ for this category is not necessary (or it should be set to a much greater value). We nonetheless put it to $10\ M_\odot$ for the training simplicity. Moreover, more massive systems merge in only a few seconds or even less in duration [Eq. (4) and Fig. 2], and therefore they are not the main target of our study here.

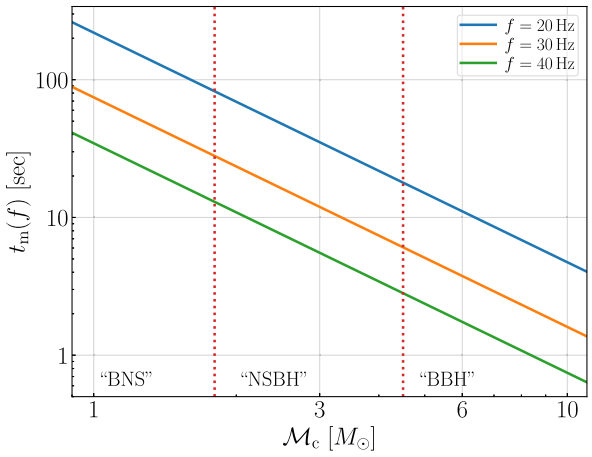

FIG. 2. Merger time as a function of the chirp mass. The vertical lines correspond the boundaries of the three signal classes we consider here. Note that the label for each class is put under quotation marks because here we only loosely define each class by its chirp mass, the information that is best constrained at the early inspiral stage. The source's properties can be better refined by a follow-up analysis utilizing data at higher GW frequencies.

Lastly, we refer to sources with $1.8\ M_\odot \leq \mathcal{M}_c < 4.5\ M_\odot$ as the "NSBH" category, as it covers events with $(M_1, M_2) = (8\ M_\odot, 1.4\ M_\odot)$, or $\mathcal{M}_c = 2.7\ M_\odot$. We nonetheless point out that this category may also contain a binary of NSs both more massive than $2\ M_\odot$ or a pair of light BHs both in the lower "mass gap" with $M_{1,2} < 5\ M_\odot$. While it is possible to refine our knowledge of the source if we include dynamics at high post-Newtonian orders and/or potential tidal interactions, these effects are encoded at higher GW frequencies and therefore is beyond the scope of our work targeting the early warning using only the low-frequency portion of the signal. Indeed, at the frequency range we are interested in here, the corrections we drop at most is on the order $\sim v^2 \simeq 1.1\%[(M_1+M_2)/3\ M_\odot]^{2/3}(f/25\ \mathrm{Hz})^{2/3}$. While potentially important for future detectors with improved sensitivity (see, e.g., Ref. [64]), for the current LIGO detectors corrections at this level can be dropped. Nonetheless, we may imagine our work here would serve as a first step of a future, integrated early warning pipeline, and once an event is detected here, it can then trigger further analysis on the signal to refine its property.

In Fig. 2 we show the merger time t_m for binaries with different chirp masses $\mathcal{M}_c$ at three different GW frequencies. The two vertical, dotted lines indicate the boundaries

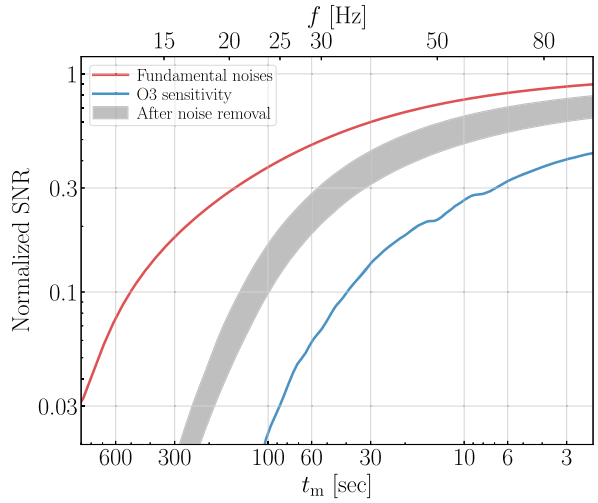

FIG. 3. Cumulative SNR as a function of t_m (bottom x axis) and f (top x axis) for a BNS with $\mathcal{M}_c = 1.22\ M_\odot$. The red trace is computed using the fundamental sensitivity and the blue one using the realistic O3 sensitivity. Also shown in the gray band is the SNR using the residual noises after the CNN's cleaning (Sec. IV B). At 100 s prior to the merger, one could in principle integrate about 40% of the total SNR if we reach the design sensitivity, yet currently only about 1% is accumulated in this band.

between the three categories defined in our study. From the plot we see that if the event can be detected by ≃30 Hz, then for the three categories ("BNS," "NSBH," "BBH"), we should in principle be able to detect the signal $[\mathcal{O}(100), \mathcal{O}(10), \mathcal{O}(1)]$ s prior to the merger.

In reality, the situation may be more challenging because the current LIGO low-frequency sensitivity is orders of magnitude above its fundamental limit as we have already seen in Fig. 1. We illustrate this point further in Fig. 3 where we show the cumulative signal-to-noise ratio (SNR) ρ for a BNS with $M_1 = M_2 = 1.4\ M_\odot$ as a function of t_m (bottom x axis) and f (top x axis). Specifically, we define the cumulative SNR through

$$\rho^2[f < f(t_m)] = 4\mathrm{Re}\left[\int^{f(t_m)} \frac{\tilde{h}^*(f)\tilde{h}(f)}{S_n(f)} df\right], \quad (7)$$

where $\tilde{h}(f) = \int h(t) \exp(i2\pi f t) dt$ and S_n the detector's power spectral density. In the plot, we further normalize the curves by the total SNR assuming the fundamental O3 sensitivity (the red trace in Fig. 1).

As can be seen from Fig. 3, with the current O3 sensitivity (blue trace), to accumulate to a normalized SNR of 0.2, we have to integrate the signal to around 40 Hz or $t_m \simeq 20$ s. Such a time window might not be sufficient especially if one wants to catch potential precursor signals of BNS mergers given various realistic delays in the information communication and decision making. In contrast, if LIGO can reach its fundamental limit, one would only need to integrate to 15 Hz, which is 300 s prior to the merger. It thus demonstrates the great scientific reward of enhancing the low-frequency sensitivity, which we propose to achieve via ML-based nonlinear noise regression.

IV. NEURAL NETWORK

A cartoon illustrating the proposed CNN structure is shown in Fig. 4. Here we input both the main GW readout and a few key auxiliary witness channels to simultaneously achieve noise mitigation and signal classification.

To assist the convergence of the network, we adopt a compound structure. We first use the network "CNN noise" to perform noise reconstruction and then subtract its output from the noisy GW readout to form a cleaned strain signal. This is then fed to the network "CNN class" to achieve signal detection and classification. Both subnetworks can be first trained individually and then combined together to perform a global optimization.

Our ML training is performed using KERAS [65], a PYTHON-based interface running on top of the ML platform TENSORFLOW [66]. The details of data generation and network training is presented below.

A. Data preparation

We describe in this section how we generate the data we used for training the CNN.

We generate the GW signal from Eq. (2) with the distance d replaced by $2.3 d_{\mathrm{eff}}/\sqrt{N_{\mathrm{det}}}$ and $N_{\mathrm{det}} = 3$. For training the CNN, it is not necessary to sample the masses following a specific astrophysical distribution. Instead, we sample M_c from a normal distribution with a mean of $1.22\ M_\odot$ and standard deviation of $0.3\ M_\odot$ and truncate the distribution at $[1, 1.8) M_\odot$, the predefined range of the "BNS" class. This choice emphasizes slightly BNS systems with $M_1 \simeq M_2 \simeq 1.4\ M_\odot$ that are most abundant according to the Galactic BNS population [67] while still allowing

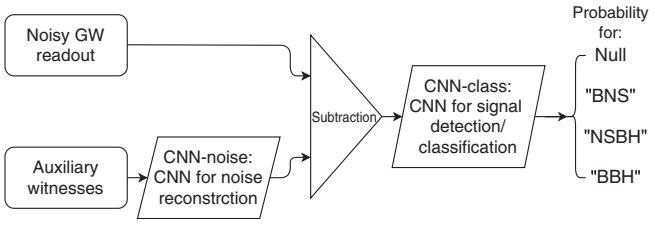

FIG. 4. A compound CNN we propose to use to detect GW events. In the "CNN-noise" CNN model we first reconstruct the noise that limits the low-frequency sensitivity using auxiliary channels. We then subtract its output from the main GW readout and then pass the residuals to "CNN-class" model which outputs the probability of the input time series belonging to each one of the classes we defined. The two CNNs can be first trained individually and then combined and optimized globally.

TABLE I. Summary of the three signal categories considered in our study. The first row is the chirp mass defining each category. The second row is the frequency at which we cut off the signal; only signal prior to this cutoff is used for detection. The third row is the typical time to merger at the cutoff frequency and the last row is the integration time of each signal. The chirp masses are given in $[M_\odot]$, times are in [sec], and frequencies are in [Hz]. All waveforms are sampled at a rate of 256 Hz.

Label	"BNS"	"NSBH"	"BBH"
$\mathcal{M}_c$	[1, 1.8)	[1.8, 4.5)	[4.5, 10)
$f_{\rm cut}$	[24, 25)	[28, 32)	[35, 40)
$t_{\rm m}^{\rm typ}\vert_{f_{\rm cut}}$	[97, 87)	[17, 12)	[4, 3)
$t_{\rm int}$	256	$\min[256, t_m\vert_{10} - t_m\vert_{f_{\rm cut}}]$	$t_m\vert_{10} - t_m\vert_{f_{\rm cut}}$

more massive BNSs to be adequately sampled in the training set. For "NSBH" and "BBH," the masses are simply sampled from uniform distributions.

To achieve early warning, we do not use the entire waveform up to the merger, but truncate the high-frequency end of the waveform at a cutoff frequency $f_{\rm cut}$. For the "BNS" class, we randomly sample $f_{\rm cut}$ between 24 and 25 Hz. For typical BNS event with $M_1 = M_2 = 1.4\ M_\odot$ ($\mathcal{M}_c = 1.2\ M_\odot$), this corresponds to $t_{\rm m}(f_{\rm cut}) = 97 - 87$ s. The starting frequency is chosen such that the integration time $t_{\rm int}$ of the signal is 256 s. Because more massive systems can evolve to higher frequencies for a given amount of integration time, we set $f_{\rm cut}$ to slightly higher values for the "NSBH" and the "BBH" classes; we sample $f_{\rm cut}$ from [28, 32) and [35, 40) Hz. For an NSBH event with $(M_1, M_2) = (8\ M_\odot, 1.4\ M_\odot)$ this leads to 17 to 12 s of premerger warning time, and for a BBH with $M_1 = M_2 = 5\ M_\odot$, it is 4 to 3 s prior to the merger. The integration time $t_{\rm int}$ is the minimum of 256 s and $t_m(10\ {\rm Hz}) - t_m(f_{\rm cut})$. In all cases, the phase at coalescence Φ_c is always sampled randomly from $[0, 2\pi)$. Because we consider $f_{\rm cut} < 40$ Hz, we only need to sample each waveform at a rate of 256 Hz. Such a relatively low sampling rate is the key allowing us to integrate the signal for a duration as long as 256 s. In Table I we summarize the key parameters of the three signal classes we consider. Hereafter, we will use the word "typical" to stand for a "BNS" binary with $M_1 = M_2 = 1.4\ M_\odot$, an "NSBH" binary with $(M_1, M_2) = (8\ M_\odot, 1.4\ M_\odot)$, and a "BBH" binary with $M_1 = M_2 = 5\ M_\odot$.

Once a waveform is generated, we then inject it to a noise background of 256 s long, containing both a stationary part due to the fundamental noise limit (the red trace in Fig. 1), and an additional low-frequency contamination represented by the blue stripe in Fig. 1 (see the description shortly after). As one may imagine continuously passing the strain data to the CNN we trained here (specifically, "CNN class" in Fig. 4), we align the signal so that it reaches $f_{\rm cut}$ at the end of the time series. This process is further illustrated in Fig. 5 [68].

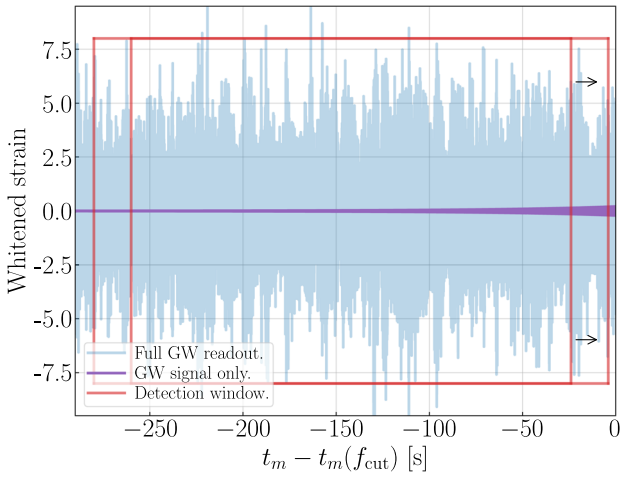

FIG. 5. A cartoon illustrating the early warning process. We continuously feed 256-s-long data segments (i.e., data enclosed in the red box, or the "detection window") to the CNN to check for the existence of GW events. Effectively, the window slides towards right as time progresses and the GW frequency of the signal at the right end of the window increases. We train the CNN to output an alert when the signal frequency at the right end reaches a predefined upper cutoff value $f_{\rm cut}$.

Together with the three signal classes, we also consider a "null" class containing only the nonstationary detector noise. The goal of CNN class is then to output the probability of a 256-second data series belonging to one of the four classes.

To simulate the excess low-frequency contamination [blue stripe in Fig. 1 and Eq. (1)], we generate noises with similar characteristics as in realistic LIGO detectors.

Specifically, we simulate four independent time series of the fast ($\gtrsim$10 Hz) angular motion $\theta(t)$, corresponding to sensing noises in the four high-bandwidth angular feedback loops (for controlling pitch and yaw motions of the two arm cavities) [69]. Instead of using realistic spectral shapes for $\theta(t)$ as in the LIGO system, we design them so that the contamination has a spectral shape similar to the full O3 sensitivity [51]. In other words, we give $\theta(t)$ extra power in the > 25 Hz band and ignore other sources of contamination in this band. This does not affect the main results of our study though because we choose $f_{\rm cut}$ between 24 and 25 Hz for the "BNS" class.

Meanwhile, we also simulate eight independent spot-position motions $x_{\rm spot}(t)$ for the four test masses and two angular degrees of freedom (pitch and yaw). Their motions are mostly induced by the microseismic motion and peak in the 0.1–0.3 Hz band. This is the main source of nonstationarity on timescale of 10 s. At longer timescales, the overall rms value of $x(t)$ drifts and shows seasonal dependence: during winter times the microseimic motion is typically higher than in the summer. To simulate this, on top of a typical value ${\rm rms}[x(t)] \simeq 0.3$ mm, we additionally sample an overall scale factor uniformly from [0.7, 1.4] and apply it to $x(t)$ for each realization.

In order to sense the true spot motions, we assume the information is contained in two sets of witness sensors. The first set of sensors probe the spot motion by exciting each mirror in angle and looking for length fluctuations at the excitation frequency. The angle-to-length conversion factor directly gives us the spot motion at each test mass [see Eq. (1)]. However, they have very limited SNR and can only trace the long-term ($\lesssim$0.1 Hz) drift of the spot motion. The other set of sensors are optical levers placed locally at each test mass. They senses the angular motion of each test mass relative to its local ground, which can then be converted to the spot motion using the cavity's geometry. They provide information in the $\gtrsim$0.1 Hz band but are polluted by seismic and thermal drifts at lower frequencies and therefore are not coherent with the true spot motion at <0.1 Hz. Consequently, we would need two sensors (one dithering-based sensor and one optical lever) for the spot motion per test mass per direction. In total, we thus need 20 auxiliary witness channels [4 for $\theta(t)$ and 16 for $x_{\rm spot}(t)$] to reconstruct the low-frequency contamination [71]. Same as the main GW readout, all of the auxiliary channels are sampled at a rate of 256 Hz. As a caveat, we note that for the real LIGO detectors, more low-frequency channels (e.g., monitors of the seismic motion and sensors of the longitudinal motions, etc.) may be needed to fully reconstruct the spot-position motion, as various cross-couplings exist at low frequencies, which are beyond the complexity capture by our current simulation.

To reduce the complexity of the problem, we first train the two sub models, "CNN noise" and "CNN class," individually, which we will describe in Secs. IV B and IV C. After each submodel's convergence, we then load their weights into the compound model as the initial condition and perform a global optimization (Sec. IV D).

B. Noise subtraction

Our first step is to construct a CNN that mitigates the excess low-frequency contamination to the GW readout in real time. We will refer to this CNN specifically as "CNN noise." It takes the 20 auxiliary witness channels we simulate as the input and estimates their nonlinear contamination to the main GW readout as the output (see also Fig. 4). To achieve supervised learning, we use time series from the noisy GW readout as training targets for this step.

Because for most of the observation time there will be no GW signal present in the data, the training is thus performed on signal-free data series only in this step. We also do not need to use the full 256 seconds of data for each training segment for noise mitigation because the contamination relies only on the instantaneous spot and angle [Eq. (1)]. The time series only needs to be long enough to capture the microseismic motion (with a characteristic period of ∼10 s) which is the main cause of fluctuations in the spot motion. Consequently, we use 64 seconds of data from 21 channels (20 auxiliary witnesses as the input and 1 noisy GW readout as the target) for each segment (i.e., "batch" in the ML literatures), and train "CNN noise" over 128 segments for each training epoch. In total, we train "CNN noise" on 8192 seconds of data and we use an additional 2048 seconds of data for validation.

Moreover, for the convenience of the subsequent signal classification, we would like the "cleaned" GW readout to have a nearly white spectral shape. Therefore, we precondition the noisy GW readout before it is passed for training. Since the current O3 detector noise is orders of magnitude greater than the fundamental noise limit (the ideal output of noise cleaning) at $f \lesssim 20$ Hz, the precondition is done in an iterative way.

In the first iteration, we whiten the GW readout according to the fundamental O3 noise limit. The spectrum of the residual after noise subtraction is then used to design the preconditioning filter for whitening the GW readout in the next iteration. Note that in each iteration, the preconditioning filter is determined before the training and the same filter is applied to all the samples. Because of the nonlinearity involved in the noise coupling, a CNN trained to estimate $\delta x(t)$ from $[\theta(t), x_{\rm spot}(t)]$ according to Eq. (1) does not apply for approximating $\mathcal{L}\{\delta x(t)\}$ from $[\mathcal{L}\{\theta(t)\}, x_{\rm spot}(t)]$ with $\mathcal{L}$ denoting a generic linear operator. Therefore, the weights in "CNN noise" need to be updated once the preconditioning filter changes. Nevertheless, we find the residual are similar for the first and second iterations, and therefore we do not to iterate further.

The same preconditioning filter is also applied to the witness channels for $\theta(t)$. While this does not preserve the exact coupling as we argued above, we nonetheless find it helps the CNN to converge faster numerically.

As for the witness channels for the spot motion, we only apply an overall calibration factor so that each channel's numerical values are of order unity. Specifically, we calibrate the dithering-based sensors to output spot motion in millimeter and the optical levers to output the low-frequency (<1 Hz) angular motion [72] in microradians. Note that the overall rms of each channel contains physical meaning [the coupling strength from $\theta(t)$ to $\delta x(t)$] and should not be normalized out. Similarly, for each GW readout we apply a fixed normalization constant.

Once the data are generated and preconditioned, we then pass them to "CNN noise" to learn the nonlinear noise coupling from the auxiliary channels to the main GW readout. The best performing network structure is summarized in Table II. Specifically, the data CNN noise takes is a three-dimensional array. The first dimension is the "batch dimension" and it corresponds to different realization of the training sample. The second dimension is a temporal dimension and it has a size given by the product of the sampling rate and duration of the data per segment. In the code, we leave the size of this dimension to be flexible (i.e., it is set to "None" when implemented in KERAS [65]) so that the noise subtraction can be performed on time series of

TABLE II. Network structure for noise subtraction. The network includes about 640,000 trainable parameters in total.

Layer	Output dimension	Kernel size	Activation
Conv1D	256	64	ELU
Conv1D	32	16	ELU
Dropout	⋯	⋯	⋯
Conv1D	256	8	ELU
BatchNormalization	⋯	⋯	⋯
Dense	256	⋯	ELU
Dropout	⋯	⋯	⋯
Dense	128	⋯	ELU
Dense	128	⋯	ELU
Dense	16	⋯	ELU
BatchNormalization	⋯	⋯	⋯
Dense	⋯	⋯	Linear

arbitrary duration. In other words, while we train CNN noise using segments of 64 seconds, when it is applied to a segment of 256 seconds (the duration of our detection window), the entire segment is cleaned. Indeed, this is possible because the noise coupling mechanism we consider here is instantaneous [all quantities are evaluated at the same time t; Eq. (1)]. This temporal axis is also the axis along which the convolution is preformed. Lastly, the final dimension corresponds the number of channels. It starts from 20 (the number of input channels we used in our simulation), goes through a series of matrix operations to change size according to the "output dimension" column in Table II, and eventually reaches a size of 1 as there is a single target channel (the main GW readout) in our study.

We construct a custom loss function for the training. Specifically, we compute the loss as

$$\text{Loss} = \int_{f_{\text{low}}}^{f_{\text{high}}} w S_n^{(\text{resi})} df, \tag{8}$$

where $S_n^{(\text{resi})}$ is the power spectral density of the residual (i.e., target − prediction), and w is a weighting function defined as

$$w = \mathcal{C} \frac{f^{\alpha}}{S_n^{(\text{trgt})}}, \tag{9}$$

where $S_n^{(\text{trgt})}$ is the power spectral density of the target and $\mathcal{C}$ is an overall constant so that the initial loss is of order unity. Empirically, we set $(f_{\text{low}}, f_{\text{high}}) = (7.5, 75\ \text{Hz})$ and $\alpha = -0.5$. In addition, we also sum a small contribution of the standard mean squared error (about 0.1 to the total loss) to the custom loss defined in Eq. (8) to avoid artificial offsets at zero frequency due to numerical overfitting.

We note that the loss function defined above aims to achieve a broad-band noise mitigation so that the results of "CNN noise" can be applied for various purpose (signal detection, sky localization, etc.). The optimization for the specific purpose of this work (detecting and classifying BNS ∼ 100 s prior to the merger) is left for the final step where we combine CNN noise and CNN class to preform global training.

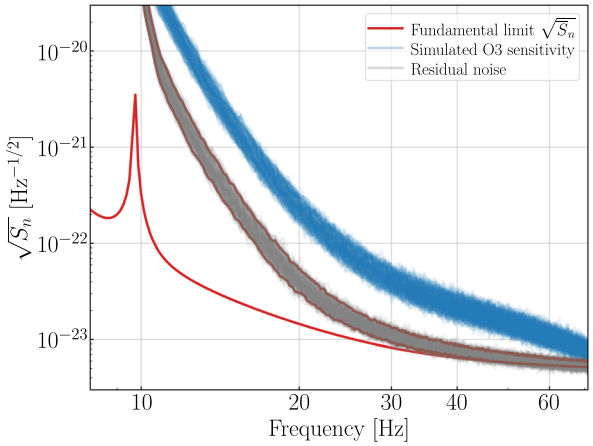

FIG. 6. Residual noise after the cleaning by CNN noise, generated from an additional 9 hours of testing data with 128 different realizations of the spot position rms. Each blue trace corresponds to one realization of our simulated detector noise and the gray ones the residual after noise subtraction. The two brown traces correspond to the 5 and 95 percentiles of the residual. With the current network, we can achieve a factor of 5 broadband subtraction of the noise.

The resultant ASDs of CNN noise are shown in Fig. 6. In the figure, each blue trace is the amplitude spectral density of a realization of the simulated O3 sensitivity. Similar to the real detector noise, it has a nonstationary nature as the rms of $x_{\text{spot}}^{(\text{ac})}$ various with time (and different from realization to realization). The residual after removing the contamination predicted by CNN noise using the 20 auxiliary channels is shown in the gray trace. To generate the plot, we simulate an additional 9 hours of testing data with 128 different realizations of the rms value of the spot position $x_{\text{spot}}^{(\text{ac})}$. Overall, the contamination can be mitigated by a factor of ∼10, which is sufficient to reach the fundamental limit in the >30 Hz band. At lower frequencies, $f < 20$ Hz, even the residual is still an order of magnitude or more above the fundamental limit and it fluctuates as the spot motion rms varies, indicating room for further improvements.

Note that each curve in Fig. 6 is the averaged ASD estimated using Welch's method over 256 second of data in total and 8 second per estimation segment. Therefore the fluctuations in Fig. 6 are due to the long-term variation of the rms of the spot motion. We also show in Fig. 7 directly the time series to compare the original (blue) and the noise-subtracted (gray) series. In the simulated O3 data, the band-limited rms in the [20,60] Hz varies on the timescale of 10 s as indicated by the envelopes of the time series. This is because the spot position on the test masses $x_{\text{spot}}^{(\text{ac})}$ moves due

to the microseismic motion in the 0.1–0.3 Hz band. Such a modulation prohibits the removal of the noise using standard signal processing techniques (such as Wiener filter) assuming a stationary coupling. The "CNN-noise," nevertheless, successfully mitigates the 10-second-timescale nonstationariety in the time series.

Furthermore, as we shown in Fig. 3, with the cleaned sensitivity represented by the gray traces in Figs. 6 and 7, we can get more than 10% of the total SNR 100 seconds prior to the merger. This is sufficient for us to detect nearby BNS events like GW170817 (Sec. V). For future convenience, we also show the 5 and 95 percentiles at each frequency bin of the residual in the two brown traces in Fig. 6. We further define $\tilde{\rho}$ as the SNR computed assuming a stationary noise background whose values are fixed at the 5 percentiles [i.e., using the lower brown trace for $\sqrt{S_n}$ in Eq. (7)]. We will use this as an estimation of the SNR of the signal at a given distance, though one should keep in mind that $\tilde{\rho}$ will in general be greater than the true SNR of each injection.

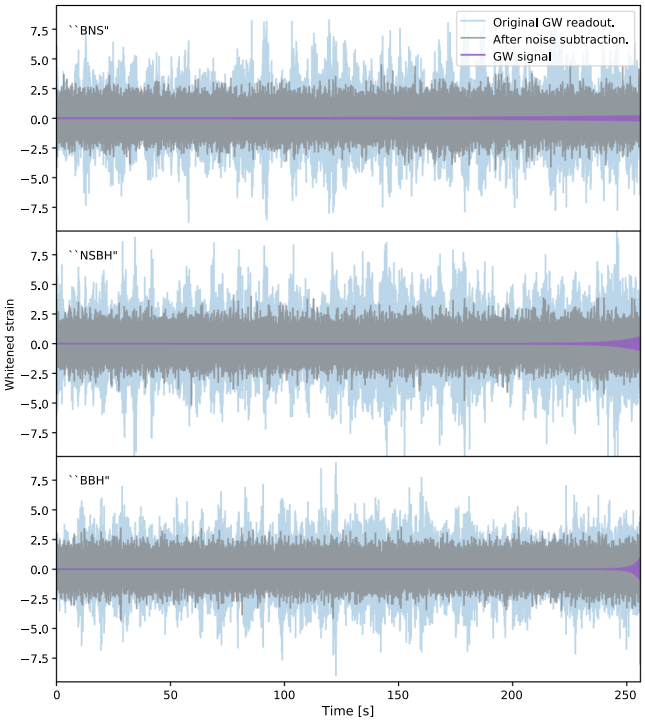

FIG. 7. Sample of the whitened time series. The blue traces are the original GW readout simulated according to O3 sensitivity whose band-limited rms in the [20, 60] Hz band varies on a timescale of 10 s due to modulations caused by the microseismic motion. The gray traces are the GW readout after noise mitigation by CNN noise and they are the inputs to CNN class. The whitened GW signal contained in each realization is highlighted in the purple trace. From top to bottom, they correspond respectively to a typical "BNS," "NSBH," and "BBH." In all the cases we set $\tilde{\rho}(f < f_{\rm cut}) = 16$.

C. Signal detection and classification

Once we have trained the CNN-noise subnetwork, we then inject GW signal onto the cleaned noise background and train the CNN class for signal detection and classification.

Examples of the input time series to CNN class is shown in Fig. 7. It is the sum of a GW signal at most 256 seconds long (or zero for a null event) and a 256-second residual noise background produced by subtracting the prediction of CNN noise and the simulated O3 detector noise (i.e., it corresponds to the gray trace in Fig. 7).

The training target is the label of each sample: we use (0, 1, 2, 3) for ("Null," "BNS," "NSBH," "BBH"), respectively. We further convert the label into the one-hot representation, so that when we use CNN class for prediction, the numerical value at each digit gives the probability of the input time series belonging to the corresponding signal class.

Table III shows the structure of the best performing CNN class we find empirically. It consists of 5 CNN layers, each followed by a pooling layer (for the first one we use average pooling while for the rest maximum pooling is used). While rectified linear unit activation are used in previous studies, we nonetheless find exponential linear unit (ELU) activation gives a better convergence and therefore we use it for all the CNN layers. We include three dense layers with ELU activation afterwards, and lastly, the output is produced by a Dense layer with the Softmax activation. The sparse categorical crossentropy loss is used together with an Adamax optimizer.

To help the convergence of the network, we utilize the "curriculum learning" approach [39,40]. That is, we first train CNN class on very loud GW events with high SNR to guide the CNN to an initial convergence. Then we gradually reduce the SNR of injected GW events in the

TABLE III. Network structure for signal classification. The network includes about 32,000 trainable parameters in total.

Layer	Output dimension	Kernel or pooling size	Activation
Conv1D	64	64	ELU
AveragePooling1D	⋯	4	⋯
Conv1D	16	16	ELU
MaxPooling1D	⋯	4	⋯
Conv1D	32	8	ELU
MaxPooling1D	⋯	8	⋯
Conv1D	16	8	ELU
MaxPooling1D	⋯	8	⋯
Conv1D	8	4	ELU
MaxPooling1D	⋯	8	⋯
Flatten	⋯	⋯	⋯
Dense	16	⋯	ELU
Dense	64	⋯	ELU
Dense	8	⋯	ELU
Dense	4	⋯	Softmax

training set to cover the more realistic SNR space of potential astrophysical events.

Specifically, in the first step, we sample GW events from $\tilde{\rho}(f < f_{\rm cut}^{\rm up}) \in [16, 40)$ and with a probability $\propto [\tilde{\rho}(f < f_{\rm cut}^{\rm up})]^{-2}$, where $\tilde{\rho}(f < f_{\rm cut}^{\rm up})$ is the SNR computed using the 5 percentile noise residual (the lower brown trace in Fig. 6) and integrated to $f_{\rm cut}^{\rm up} = (25, 32, 40)$ Hz for ("BNS," "NSBH," "BBH"). We use the "up" superscript because they correspond to the upper bound of $f_{\rm cut}$ (see Table I). Note $\tilde{\rho}$ in general will be greater than the true SNR of an injected event because both $f_{\rm cut} < f_{\rm cut}^{\rm up}$ for each realization of the GW event and the background noise is typically greater than 5 percentile value. The training set includes ~2000 samples for each signal class, plus ~2000 samples for null events. Additional 64 samples per class are used as validation.

Once the first step converges (both the training and loss plateau), we then reduce the SNR range to $\tilde{\rho}(f < f_{\rm cut}^{\rm up}) \in [10, 40)$ and $\tilde{\rho}(f < f_{\rm cut}^{\rm up}) \in [8, 28)$ in the second and third training steps. In each step, we use ~8000 samples per class. Empirically, we find that the network has a slightly improved performance if we sample events uniformly in $\tilde{\rho}$ during the second and third steps. We set the lower bound of our training set to $\tilde{\rho} = 8$ as systems with $\tilde{\rho} < 8$ (the true SNR ρ will be even smaller) are typically not considered as a detection even when the matched-filter technique is used.

As a comparison, we also construct a network with the same structure as a CNN class but train it on a GW time series with stationary noise background generated according to the fundamental O3 sensitivity (which is similar to the aLIGO design sensitivity for $f \lesssim 40$ Hz of interest; red trace in Fig. 1). This reference network is trained with the same curriculum training steps as CNN class.

D. Combined network

While it is sufficient to train CNN noise and CNN class individually as in Secs. IV B and IV C, we may further optimize the performance by combining the two networks and training globally. This is because CNN noise is trained to achieve a broadband noise reduction so that the residual detector noise could potentially serve as the input for pipelines of various purposes. By combining it with CNN class, the noise subtraction is then optimized specifically for the early detection and classification of GW events.

To achieve so, we utilize the structure shown in Fig. 4 and load the network weights obtained from individual training as the initial condition for the compound network. We generate ~10,000 samples for each class with the SNR $\tilde{\rho}(f < f_{\rm cut}^{\rm up})$ uniformly sampled from [8, 28). Each time series of the main GW channel is input to the compound network together with 20 auxiliary channels to internally mitigate the detector noise. The training target, loss function, and optimizer are the same as described in Sec. IV C.

We find the compound network could achieve an enhanced performance compared to CNN class alone (see the discussions in the following section). We will then use the compound network as our final CNN and examine its performance of BNS early warning.

V. RESULTS

We access the performance of our CNN by examining the receiver operator characteristic (ROC) curves which we construct using the scikit-learn package [73]. This can be obtained by varying the detection threshold of the predicted true probability and compute both the true alarm rate (TAR) and false-alarm rate (FAR) at the given threshold, as demonstrated in Fig. 8.

More conveniently, we can directly consider TAR as a function of FAR for a particular source, as shown in Fig. 9. For the GW event, we consider BNSs with $M_1 = M_2 = 1.4\ M_\odot$, $f_{\rm cut} = 25$ Hz, and a random $\Phi_c \in [0, 2\pi)$, and vary the sources' averaged effective distance from 20 to 100 Mpc (corresponding to traces of different colors). At each distance, we generate 2000 new testing signals and inject them onto different realizations of the detector noises. The solid traces are results using simulated O3 sensitivity as the noise background, including both a stationary Gaussian component and an excess contamination following Eq. (1), with each $x_{\rm spot}^{\rm (ac)}$ having a randomly sampled overall rms $\in [0.21, 0.42]$ mm. The excess low-frequency contamination is witnessed by 20 auxiliary channels as described in Sec. IV A. We input the main GW readout and the auxiliary channels together to the compound CNN (Fig. 4) to simultaneously preform noise mitigation and signal detection/classification. As a comparison, we also show the performance of the reference network in the dotted traces. It has the same structure as CNN class but the noise background for training and

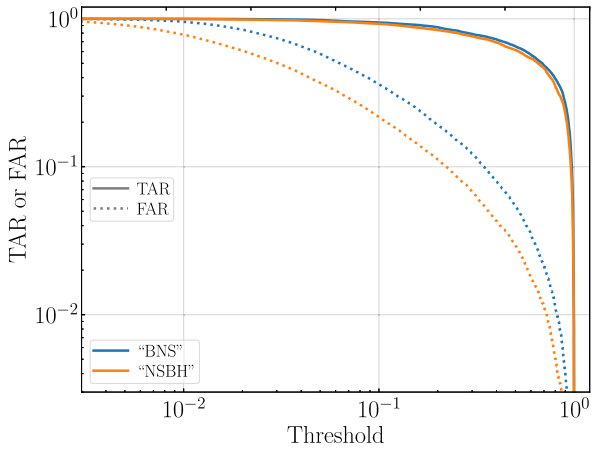

FIG. 8. TAR (solid traces) or FAR (dotted traces) as a function of the threshold above which a detection is claimed. The results are obtained using our compound CNN (Fig. 4) with simulated O3 sensitivity. The blue traces is for a typical BNS event at $d_{\rm eff} = 40$ Mpc (see Figs. 9 and 10) and the orange trace is for a NSBH event at $d_{\rm eff} = 160$ Mpc (see Fig. 11).

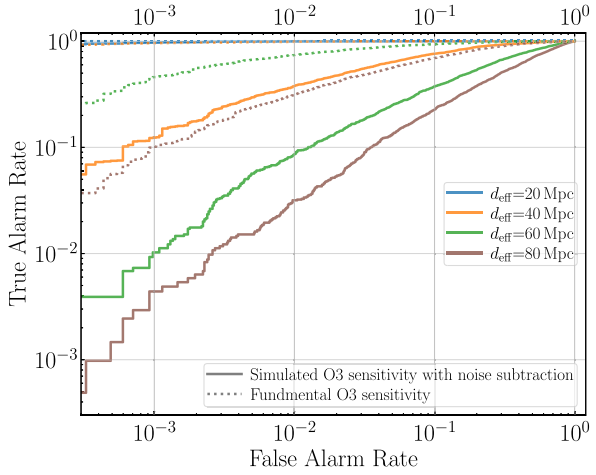

FIG. 9. The ROC curves for typical BNS events at different distances. Here we focus on events with $M_1 = M_2 = 1.4\ M_\odot$ and $f_{\rm cut} = 25$ Hz with varying effective distance (assuming $N_{\rm det} = 3$). The FAR is constructed from "null" events (i.e., detector noise) and FAR $= 0.01$ corresponds to 1 false alarm every 100 256-second samples (approximately 1 in every 7.1 hours). In the solid traces, we use simulated, nonstationary detector noise representing the O3 sensitivity and the GW readout is input to the compound network together with 20 auxiliary witness channels. As a comparison, the dotted traces use stationary detector noise representing the fundamental sensitivity for O3.

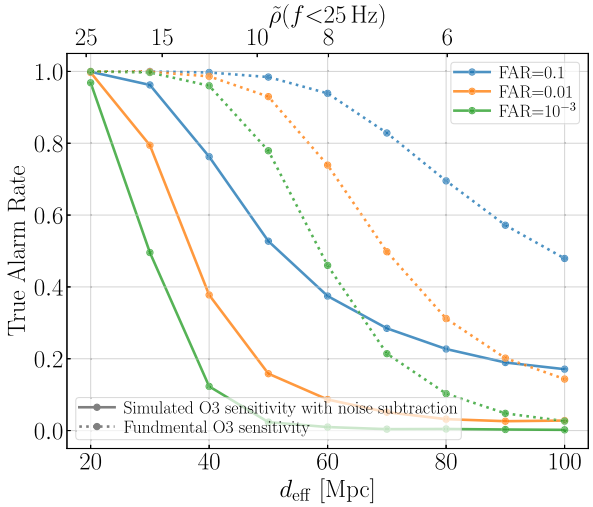

FIG. 10. The CNN's detection efficiency of typical BNS events at fixed values of FAR. As a reference, the top x axis shows the SNR computed using the 5 percentile noise residual (the lower brown trace in Fig. 6; it is typically greater than the true SNR due to both the short and long timescale nonstationarities). A GW170817-like event can be detected about 1.5 minutes prior to the merger if the contamination to the detector's low-frequency sensitivity can be mitigated to the level shown in Figs. 6 and 7. If the design sensitivity is reached, such an early warning would be possible for sources with $d_{\rm eff} \simeq 80$ Mpc.

prediction is generated according to the stationary fundamental O3 sensitivity. Here the FAR is constructed from $\sim 20,000$ realization of detector noises ("null" events; corresponding to two months of data). Note here the rate is measured per 256-second data segment. As a result, FAR $= 0.01$ would correspond to roughly 1 false alarm every 7.1 hr of detector data. Nonetheless, this conversion should only be treated as a crude approximation, as a more careful one should account for effects such as detection by multiple detectors, which we defer to future studies.

Alternatively, we can fix the FAR and examine how the TAR varies as a function of the averaged effective distance $d_{\rm eff}$. The result is shown in Fig. 10. The astrophysical source is still fixed to be BNSs with $M_1 = M_2 = 1.4\ M_\odot$ and $f_{\rm cut} = 25$ Hz and the line styles have the same meaning as in Fig. 9. Different colors now represent different FAR thresholds. We see that if we could mitigate the noise to a level comparable to the gray stripe in Fig. 6, then a GW170817-like event at $d \simeq 40$ Mpc can be detected 1.5 minutes prior to the merger with a decent chance.

Because of the nonstationarity in the background noise, the matched-filter SNR is not a constant even for a fixed effective distance. If we nonetheless treat the noise PSD as being stationary and use the 5 and 95 percentiles in the cleaned spectra (i.e., the two brown traces in Fig. 6), we estimate the SNR to be around 12 to 7.3. On the other hand, if the noise background becomes truly stationary and reaching the designed aLIGO sensitivity, then the early detection can be achieved to $d_{\rm eff} \simeq 80$ Mpc. The corresponding matched-filter SNR is 12. The required SNR for detection of a stationary noise background being similar to the SNR calculated using the 5 percentile of the nonstationary background suggests that our final, global training (Sec. IV D) mitigates the nonstationarity further and improves the CNN's performance than treating the noise subtraction and signal detection as two separate, independent problems.

In addition to BNS mergers, NSBH mergers are another type of sources for multimessenger astronomy, and we access the performance of our CNN for detecting them in Fig. 11. The GW events we inject are $(M_1, M_2) = (8, 1.4) M_\odot$ and $f_{\rm cut} = 32$ Hz ($t_{\rm m} = 12$ s). If we still use the (TAR, FAR) $= (0.4, 0.01)$ as the threshold for detection, we find an NSBH can be detected 12 s before the merger at an averaged effective distance of $d_{\rm eff} \simeq 160$ Mpc using simulated O3 sensitivity with noise subtraction. The matched-filter SNR is estimated to be between 10.5 (5 percentile) and 7.0 (95 percentile). Using the stationary, fundamental O3 sensitivity, we find the detection range to be around $d_{\rm eff} \sim 240$ Mpc. The corresponding SNR is 10.5, again similar to the 5 percentile value when the nonstationary noise is used, suggesting that the nonstationarity is largely removed with the internal noise cleaning.

Interestingly, we note that at a given value of $\tilde{\rho}$, our CNN typically performs better for detecting NSBHs than BNSs.

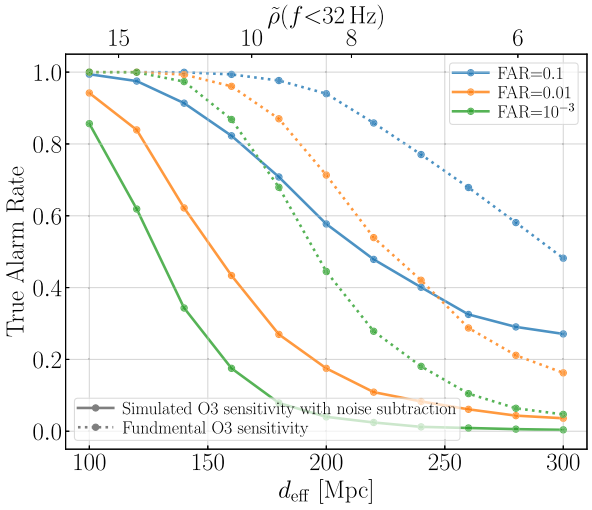

FIG. 11. Similar to Fig. 10 but for an NSBH event with $(M_1, M_2) = (8, 1.4)M_\odot$ and $f_{\mathrm{cut}} = 32$ Hz, with an early warning time of $t_{\mathrm{m}} = 12$ s. At a given $\tilde{\rho}$, our CNN performs better for detecting NSBHs than BNSs in general.

Whereas the CNN's sensitivity starts to drop sharply at $\tilde{\rho} \simeq 15$ and essentially vanishes $\tilde{\rho} \simeq 10$ for the "BNS" signal, for "NSBH" we still have a decent sensitivity at $\tilde{\rho} \simeq 10$. In part, an NSBH has its signal "concentrated" in a shorter duration with a louder time-domain amplitude than a BNS and therefore it is more easily recognized by a CNN (similarly, Ref. [45] also found that a CNN typically performs better for BBHs than BNSs with the same matched-filter SNR). Meanwhile, we have also chosen a higher upper cutoff frequency for NSBHs (32 Hz) than for BNSs (25 Hz), and the fluctuation in the PSD of the background noise is less at higher frequencies between different realizations after the cleaning by CNN noise.

Another quantity of interest is the FCR. Specifically, if there is a BBH event (which typically does not have an EM counterpart) present in the GW readout, we want to address the probability of classifying it as a "BNS" or "NSBH" and falsely triggering subsequent EM followup observations. The result is shown in Fig. 12. The FCR is constructed from 5000 "BBH" injections. The "BBH" events are sampled from a distribution $\propto [\tilde{\rho}(f < 40\ \mathrm{Hz})]^{-2}$ and $\tilde{\rho}(f < 40\ \mathrm{Hz}) \in [8, 40)$. By comparing the top panel of Fig. 12 with Fig. 9, we see that a "BNS" trigger is much less likely to be confused by a true "BBH" event than by the detector noise. The "NSBH" class has slightly more false classifications from the "BBH" class, yet at FCR $= 0.01$, we still have TCR > 0.1 for $\tilde{\rho}(f < 32\ \mathrm{Hz}) > 7$.

Lastly, we point out that our compound CNN not only provides a potential way to achieve real-time noise mitigation and signal detection, it could also serve as an efficient first step to existing match-filter-based pipelines. This is because the computationally expensive part is the training, which would only need to be done once because the noise coupling mechanism [e.g., Eq. (1)] itself is

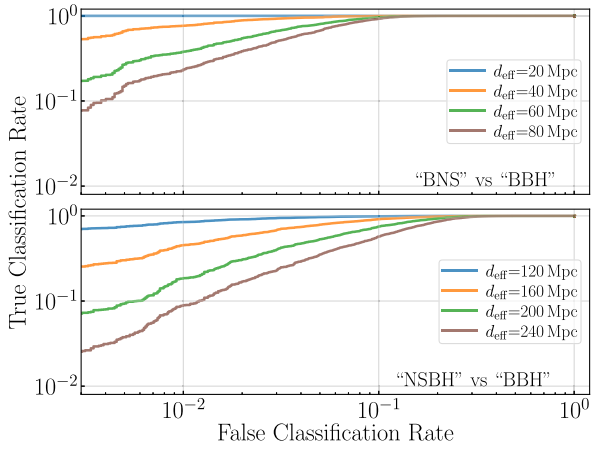

FIG. 12. The true classification rate (TCR) as a function of the false classification rate [FCR, i.e., the rate of falsely classifying a "BBH" event as a "BNS" (top panel) or an "NSBH" (bottom panel) event].

"fixed" and the nonstationarity [e.g., the modulation caused by $x^{(\mathrm{ac})}_{\mathrm{spot}}$] is treated as input to the CNN. Once the network is trained, the subsequent prediction time is typically only 100 ms for doing both noise mitigation and signal classification, or 30 ms for just performing signal classification, as shown in Fig. 13 for the testing data [74]. Indeed, once a signal is detected and classified by the network, subsequent matched-filter analysis would only need to perform searches over a small sub-bank after the classification performed by CNN, potentially enhance the efficiency of the existing pipelines.

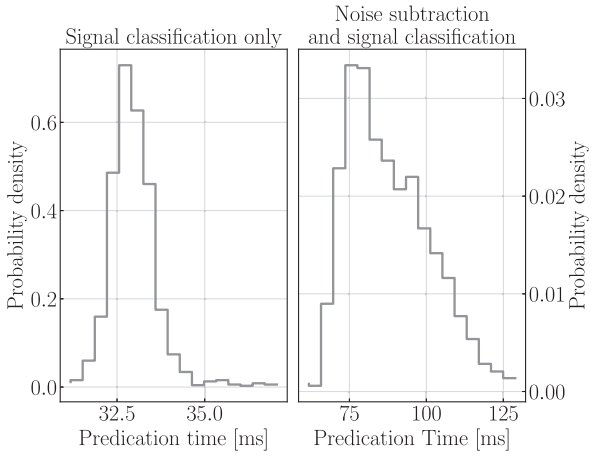

FIG. 13. Histogram of the computational time of our CNNs (using a Tesla P100-PCIE-16 GB GPU). In the left is the time the "CNN class" takes to classify a 256-second time series, and in the right is the time for the compound problem where we input both the GW readout and 20 auxiliary channels. Once the training is complete, the prediction takes only ~30 ms for the CNN to classify a time series; even including real-time noise subtraction, the computation time is still less than 100 ms in most cases.

VI. CONCLUSION AND DISCUSSION

We showed that it would be possible to detect BNS (NSBH) signals from the real-time LIGO data series using a ML-based CNN.

To achieve so, it requires improving the LIGO sensitivity in the $\lesssim$60 Hz band, which currently dominated by nonlinear cross-couplings from the auxiliary control loops and/or environmental perturbations. We demonstrated that one potential way to enhance the low-frequency sensitivity is to input the auxiliary channels together with the main GW readout to a CNN and use it to simultaneously perform noise cleaning and signal detection.

With noise mitigation reaching the level shown in Figs. 6 and 7, we can detect BNS (NSBH) ∼100 s (10 s) prior to merger out to $d_{\rm eff} \simeq 40$ Mpc (160 Mpc) with a TAR $\gtrsim 0.4$ and FAR $= 0.01$ (very crudely speaking, this corresponds 1 false alarm every 7.1 hours). If we have a stationary, Gaussian noise background reaching the designed sensitivity, the early warning can be achieved out to $d_{\rm eff} \simeq 80$ and 240 Mpc for BNS and NSBH, respectively. The matched-filter SNR is 12 and 10 for typical BNSs and NSBHs, respectively. Moreover, we find the threshold SNRs for the Gaussian noise background are similar to the SNRs estimated using the 5 percentile of the nonstationary noise (the bottom brown trace in Fig. 6). This indicates that our compound network structure (Fig. 4) largely mitigates complications due to a nonstationary background, and the global training (Sec. IV D) enhances the CNN's performance than treating the noise cleaning and signal detection as two separate problems.

We note that our current CNN has not yet reached a sensitivity comparable to the existing low-latency pipelines. For example, Ref. [29] considered a similar early warning problem using GSTLAL and the designed aLIGO sensitivity. According to the associated data release [75], the authors of Ref. [29] performed 1446 BNS injections with distance from 80 to 100 Mpc in total and they were able to detect 446 (or 31%) out of them at an upper cutoff frequency of 29 Hz and a FAR $= (30\ \text{days})^{-1}$. From the dotted traces in Fig. 10, our CNN can achieve a similar TAR $= 0.3$ only at a FAR that is about 100 times higher than the GSTLAL results. While in part the difference in the performance is due to the fact that we considered a lower cutoff frequency of 25 Hz and the integration time of the signal is thus ≃30 s shorter, it nonetheless indicates that the ML-based CNN still has a large room for future improvement.

Nevertheless, a ML-based CNN has a few advantages over the existing pipelines that warrant future studying. First of all, as multiple authors have pointed out (see, e.g., Refs. [39,40,45]), a CNN is highly efficient in prediction. Indeed, as we showed in Fig. 13, it takes the CNN class only 30 ms to detect and classify a GW signal from a 256-s data segment. In comparison, the typical latency is about 6 s for GSTLAL, indicating the possibility of accelerating the existing pipelines even further.

More importantly, we can input not only the strain readout but also auxiliary channels to the CNN to enhance the detection of GW signal. Here we focused specifically on removing the excess and nonstationary contamination to the low-frequency band. In addition to help the early warning of BNSs and NSBHs, mitigating the nonstationarity could also help to reduce the false triggers of heavy BBHs due to the drift of background PSD [10]. Vetoing and/or mitigating glitches is another task an CNN could help with inputting also auxiliary witnesses [76,77]. In principle, one can combine multiple noise mitigation feed-forwards and data quality checks with a signal detection routine into a single CNN (potentially with a compound structure) that efficiently enhances LIGO's performance.

As a proof of concept, we used simulated data to mimic the O3 LIGO sensitivity and our auxiliary witnesses are designed to emulate realistic channels in LIGO. On the other hand, we note that subtracting the nonlinear, nonstationary noises in the real LIGO interferometers is still significantly more challenging than our simulation (which only captures one of the many noise coupling mechanisms present in LIGO [51]). There is also a public data release containing three hours of outputs from selected LIGO auxiliary channels available at [78]. We encourage interested readers to utilize the CNN structures we proposed in this work or original CNN structures to help the further improvements of the LIGO sensitivity.

ACKNOWLEDGMENTS

We thank Zachary Mark, Katerina Chatziioannou, Erik Katsavounidis, and Deep Chatterjee for useful comments and discussions. H. Y. is supported by the Sherman Fairchild Foundation. R. X. A. is supported by NSF PHY-1764464. S. S. is supported by the Eberly Research Funds of Penn State, The Pennsylvania State University, University Park, Pennsylvania. The authors gratefully acknowledge the computational resources provided by the LIGO Laboratory and supported by NSF Grants No. PHY-0757058 and No. PHY-0823459. LIGO was constructed by the California Institute of Technology and Massachusetts Institute of Technology with funding from the NSF and operates under cooperative agreement PHY-1764464. This paper carries LIGO Document Number LIGO-P2100093.

[1] J. Aasi *et al.* (LIGO Scientific Collaboration), Advanced LIGO, Classical Quant. Grav. **32**, 074001 (2015).

[2] F. Acernese *et al.* (VIRGO Collaboration), Advanced Virgo: A second-generation interferometric gravitational wave detector, Classical Quant. Grav. **32**, 024001 (2015).

[3] Kagra Collaboration *et al.*, KAGRA: 2.5 generation interferometric gravitational wave detector, Nat. Astron. **3**, 35 (2019).

[4] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Observation of Gravitational Waves from a Binary Black Hole Merger, Phys. Rev. Lett. **116**, 061102 (2016).

[5] B. P. Abbott, R. Abbott, T. D. Abbott, S. Abraham, F. Acernese, K. Ackley, C. Adams, R. X. Adhikari *et al.*, GWTC-1: A Gravitational-Wave Transient Catalog of Compact Binary Mergers Observed by LIGO and Virgo during the First and Second Observing Runs, Phys. Rev. X **9**, 031040 (2019).

[6] R. Abbott, T. D. Abbott, S. Abraham, F. Acernese, K. Ackley, A. Adams, C. Adams, R. X. Adhikari *et al.*, GWTC-2: Compact Binary Coalescences Observed by LIGO and Virgo During the First Half of the Third Observing Run, Phys. Rev. X **11**, 021053 (2021).

[7] A. H. Nitz, C. Capano, A. B. Nielsen, S. Reyes, R. White, D. A. Brown, and B. Krishnan, 1-OGC: The first open gravitational-wave catalog of binary mergers from analysis of public Advanced LIGO Data, Astrophys. J. **872**, 195 (2019).

[8] A. H. Nitz, T. Dent, G. S. Davies, S. Kumar, C. D. Capano, I. Harry, S. Mozzon, L. Nuttall, A. Lundgren, and M. Tápai, 2-OGC: Open gravitational-wave catalog of binary mergers from analysis of public Advanced LIGO and Virgo data, Astrophys. J. **891**, 123 (2020).

[9] B. Zackay, L. Dai, and T. Venumadhav, Relative binning and fast likelihood evaluation for gravitational wave parameter estimation, arXiv:1806.08792.

[10] B. Zackay, T. Venumadhav, J. Roulet, L. Dai, and M. Zaldarriaga, Detecting gravitational waves in data with non-Gaussian noise, arXiv:1908.05644 [Phys. Rev. D (to be published)].

[11] B. Zackay, T. Venumadhav, L. Dai, J. Roulet, and M. Zaldarriaga, Highly spinning and aligned binary black hole merger in the Advanced LIGO first observing run, Phys. Rev. D **100**, 023007 (2019).

[12] T. Venumadhav, B. Zackay, J. Roulet, L. Dai, and M. Zaldarriaga, New search pipeline for compact binary mergers: Results for binary black holes in the first observing run of Advanced LIGO, Phys. Rev. D **100**, 023011 (2019).

[13] T. Venumadhav, B. Zackay, J. Roulet, L. Dai, and M. Zaldarriaga, New binary black hole mergers in the second observing run of Advanced LIGO and Advanced Virgo, Phys. Rev. D **101**, 083030 (2020).

[14] R. Magee, H. Fong, S. Caudill, C. Messick, K. Cannon, P. Godwin, C. Hanna, S. Kapadia, D. Meacher, S. R. Mohite, D. Mukherjee, A. Pace, S. Sachdev, M. Shikauchi, and L. Singer, Sub-threshold binary neutron star search in Advanced LIGO's first observing run, Astrophys. J. Lett. **878**, L17 (2019).

[15] B. P. Abbott, R. Abbott, T. D. Abbott, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso *et al.*, LIGO Scientific Collaboration, and Virgo Collaboration, GW170817: Observation of Gravitational Waves from a Binary Neutron Star Inspiral, Phys. Rev. Lett. **119**, 161101 (2017).

[16] A. Goldstein, P. Veres, E. Burns, M. S. Briggs, R. Hamburg, D. Kocevski, C. A. Wilson-Hodge, R. D. Preece, S. Poolakkil, O. J. Roberts, C. M. Hui, V. Connaughton, J. Racusin, A. von Kienlin, T. Dal Canton, N. Christensen, T. Littenberg, K. Siellez, L. Blackburn, J. Broida *et al.*, An ordinary short gamma-ray burst with extraordinary implications: Fermi-GBM detection of GRB 170817A, Astrophys. J. Lett. **848**, L14 (2017).

[17] B. P. Abbott, R. Abbott, T. D. Abbott, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso *et al.*, Multi-messenger observations of a binary neutron star merger, Astrophys. J. Lett. **848**, L12 (2017).

[18] B. P. Abbott, R. Abbott, T. D. Abbott, S. Abraham, F. Acernese, K. Ackley, C. Adams, V. B. Adya *et al.*, Model comparison from LIGO-Virgo data on GW170817's binary components and consequences for the merger remnant, Classical Quant. Grav. **37**, 045006 (2020).

[19] N. N. Weinberg *et al.* (LIGO Scientific and Virgo Collaborations, Constraining the p-Mode-g-Mode Tidal Instability with GW170817, Phys. Rev. Lett. **122**, 061104 (2019).

[20] LIGO Scientific and Virgo Collaborations, GW170817: Measurements of Neutron Star Radii and Equation of State, Phys. Rev. Lett. **121**, 161101 (2018).

[21] D. Radice, A. Perego, F. Zappa, and S. Bernuzzi, GW170817: Joint constraint on the neutron star equation of state from multimessenger observations, Astrophys. J. Lett. **852**, L29 (2018).

[22] B. P. Abbott, R. Abbott, T. D. Abbott, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso *et al.*, Tests of General Relativity with GW170817, Phys. Rev. Lett. **123**, 011102 (2019).

[23] D. Tsang, J. S. Read, T. Hinderer, A. L. Piro, and R. Bondarescu, Resonant Shattering of Neutron Star Crusts, Phys. Rev. Lett. **108**, 011102 (2012).

[24] B. D. Metzger and A. L. Piro, Optical and X-ray emission from stable millisecond magnetars formed from the merger of binary neutron stars, Mon. Not. R. Astron. Soc. **439**, 3916 (2014).

[25] B. M. S. Hansen and M. Lyutikov, Radio and X-ray signatures of merging neutron stars, Mon. Not. R. Astron. Soc. **322**, 695 (2001).

[26] E. R. Most and A. A. Philippov, Electromagnetic precursors to gravitational-wave events: Numerical simulations of flaring in pre-merger binary neutron star magnetospheres, Astrophys. J. Lett. **893**, L6 (2020).

[27] D. Thornton, B. Stappers, M. Bailes, B. Barsdell, S. Bates, N. D. R. Bhat, M. Burgay, S. Burke-Spolaor, D. J. Champion, P. Coster, N. D'Amico, A. Jameson, S. Johnston, M. Keith, M. Kramer, L. Levin, S. Milia, C. Ng, A. Possenti, and W. van Straten, A population of fast radio bursts at cosmological distances, Science **341**, 53 (2013).

[28] T. Totani, Cosmological fast radio bursts from binary neutron star mergers, Publ. Astron. Soc. Jpn. **65**, L12 (2013).

[29] S. Sachdev, R. Magee, C. Hanna, K. Cannon, L. Singer, J. R. SK, D. Mukherjee, S. Caudill, C. Chan, J. D. E. Creighton, B. Ewing, H. Fong, P. Godwin, R. Huxford,

S. Kapadia, A. K. Y. Li, R. K. Lok Lo, D. Meacher, C. Messick, S. R. Mohite *et al.*, An early-warning system for electromagnetic follow-up of gravitational-wave events, Astrophys. J. Lett. **905**, L25 (2020).

[30] 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, Toward early-warning detection of gravitational waves from compact binary coalescence, Astrophys. J. **748**, 136 (2012).

[31] C. Messick, K. Blackburn, P. Brady, P. Brockill, K. Cannon, R. Cariou, S. Caudill, S. J. Chamberlin, J. D. E. Creighton, R. Everett, C. Hanna, D. Keppel, R. N. Lang, T. G. F. Li, D. Meacher, A. Nielsen, C. Pankow, S. Privitera, H. Qi, S. Sachdev *et al.*, Analysis framework for the prompt discovery of compact binary mergers in gravitational-wave data, Phys. Rev. D **95**, 042001 (2017).

[32] S. Sachdev, S. Caudill, H. Fong, R. K. L. Lo, C. Messick, D. Mukherjee, R. Magee, L. Tsukada, K. Blackburn, P. Brady, P. Brockill, K. Cannon, S. J. Chamberlin, D. Chatterjee, J. D. E. Creighton, P. Godwin, A. Gupta, C. Hanna, S. Kapadia, R. N. Lang *et al.*, The GstLAL search analysis methods for compact binary mergers in Advanced LIGO's Second and Advanced Virgo's First Observing Runs, arXiv:1901.08580.

[33] A. H. Nitz, T. Dal Canton, D. Davis, and S. Reyes, Rapid detection of gravitational waves from compact binary mergers with PyCBC Live, Phys. Rev. D **98**, 024050 (2018).

[34] T. Adams, D. Buskulic, V. Germain, G. M. Guidi, F. Marion, M. Montani, B. Mours, F. Piergiovanni, and G. Wang, Low-latency analysis pipeline for compact binary coalescences in the advanced gravitational wave detector era, Classical Quant. Grav. **33**, 175012 (2016).

[35] J. Luan, S. Hooper, L. Wen, and Y. Chen, Towards low-latency real-time detection of gravitational waves from compact binary coalescences in the era of advanced detectors, Phys. Rev. D **85**, 102002 (2012).

[36] Q. Chu, M. Kovalam, L. Wen, T. Slaven-Blair, J. Bosveld, Y. Chen, P. Clearwater, A. Codoreanu, Z. Du, X. Guo, X. Guo, K. Kim, T. G. F. Li, V. Oloworaran, F. Panther, J. Powell, A. S. Sengupta, K. Wette, and X. Zhu, The SPIIR online coherent pipeline to search for gravitational waves from compact binary coalescences, arXiv:2011.06787.

[37] B. P. Abbott, R. Abbott, T. D. Abbott, S. Abraham, F. Acernese, K. Ackley, C. Adams, R. X. Adhikari *et al.*, Low-latency gravitational-wave alerts for multimessenger astronomy during the second Advanced LIGO and Virgo observing Run, Astrophys. J. **875**, 161 (2019).

[38] R. Magee, D. Chatterjee, L. P. Singer, S. Sachdev, M. Kovalam, G. Mo, S. Anderson, P. Brady, P. Brockill, K. Cannon, T. Dal Canton, Q. Chu, P. Clearwater, A. Codoreanu, M. Drago, P. Godwin, S. Ghosh, G. Greco, C. Hanna, S. J. Kapadia *et al.*, First demonstration of early warning gravitational wave alerts, Astrophys. J. Lett. **910**, L21 (2021).

[39] D. George and E. A. Huerta, Deep neural networks to enable real-time multimessenger astrophysics, Phys. Rev. D **97**, 044039 (2018).

[40] D. George and E. A. Huerta, Deep learning for real-time gravitational wave detection and parameter estimation: Results with Advanced LIGO data, Phys. Lett. B **778**, 64 (2018).

[41] H. Gabbard, M. Williams, F. Hayes, and C. Messenger, Matching Matched Filtering with Deep Networks for Gravitational-Wave Astronomy, Phys. Rev. Lett. **120**, 141103 (2018).

[42] E. A. Huerta, A. Khan, X. Huang, M. Tian, M. Levental, R. Chard, W. Wei, M. Heflin, D. S. Katz, V. Kindratenko, D. Mu, B. Blaiszik, and I. Foster, Accelerated, scalable and reproducible AI-driven gravitational wave detection, Nat. Astron. (2021), https://doi.org/10.1038/s41550-021-01405-0.

[43] W. Wei, A. Khan, E. A. Huerta, X. Huang, and M. Tian, Deep learning ensemble for real-time gravitational wave detection of spinning binary black hole mergers, Phys. Lett. B **812**, 136029 (2021).

[44] B.-J. Lin, X.-R. Li, and W.-L. Yu, Binary neutron stars gravitational wave detection based on wavelet packet analysis and convolutional neural networks, Front. Phys. **15**, 24602 (2020).

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

[46] G. Baltus, J. Janquart, M. Lopez, A. Reza, S. Caudill, and J.-R. Cudell, Convolutional neural networks for the detection of the early inspiral of a gravitational-wave signal, Phys. Rev. D **103**, 102003 (2021).

[47] W. Wei and E. A. Huerta, Deep learning for gravitational wave forecasting of neutron star mergers, Phys. Lett. B **816**, 136185 (2021).

[48] W. Wei, E. A. Huerta, M. Yun, N. Loutrel, R. Haas, and V. Kindratenko, Deep learning with quantized neural networks for gravitational wave forecasting of eccentric compact binary coalescence, arXiv:2012.03963.

[49] One exception is Ref. [40], yet the authors focused on BBH signals whose duration is typically short ($\lesssim$1 s) compared to BNS signals, and the nonstationarity is less critical. Refs. [47,48] also uses the real aLIGO data. However, Refs. [47,48] input the spectrograms where the phase information is discarded.

[50] D. V. Martynov, E. D. Hall, B. P. Abbott, R. Abbott, T. D. Abbott, C. Adams, R. X. Adhikari, R. A. Anderson *et al.*, Sensitivity of the Advanced LIGO detectors at the beginning of gravitational wave astronomy, Phys. Rev. D **93**, 112004 (2016).

[51] A. Buikema, C. Cahillane, G. L. Mansell, C. D. Blair, R. Abbott, C. Adams, R. X. Adhikari, A. Ananyeva *et al.*, Sensitivity and performance of the Advanced LIGO detectors in the third observing run, Phys. Rev. D **102**, 062003 (2020).

[52] B. Allen, W. G. Anderson, P. R. Brady, D. A. Brown, and J. D. E. Creighton, FINDCHIRP: An algorithm for detection of gravitational waves from inspiraling compact binaries, Phys. Rev. D **85**, 122006 (2012).

[53] H. Yu, D. Martynov, S. Vitale, M. Evans, D. Shoemaker, B. Barr, G. Hammond, S. Hild *et al.*, Prospects for Detecting Gravitational Waves at 5 Hz with Ground-Based Detectors, Phys. Rev. Lett. **120**, 141102 (2018).

[54] J. C. Driggers, S. Vitale, A. P. Lundgren, M. Evans, K. Kawabe, S. E. Dwyer, K. Izumi, R. M. S. Schofield,

A. Effler, D. Sigg, P. Fritschel, M. Drago, A. Nitz, and LIGO Scientific Collaboration Instrument Science Authors, Improving astrophysical parameter estimation via offline noise subtraction for Advanced LIGO, Phys. Rev. D **99**, 042001 (2019).

[55] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE **86**, 2278 (1998).

[56] C. F. Da Silva Costa, C. Billman, A. Effler, S. Klimenko, and H. P. Cheng, Regression of non-linear coupling of noise in LIGO detectors, Classical Quant. Grav. **35**, 055008 (2018).

[57] N. Mukund, J. Lough, C. Affeldt, F. Bergamin, A. Bisht, M. Brinkmann, V. Kringel, H. Lück, S. Nadji, M. Weinert, and K. Danzmann, Bilinear noise subtraction at the GEO 600 observatory, Phys. Rev. D **101**, 102006 (2020).

[58] G. Vajente, Y. Huang, M. Isi, J. C. Driggers, J. S. Kissel, M. J. Szczepańczyk, and S. Vitale, Machine-learning non-stationary noise out of gravitational-wave detectors, Phys. Rev. D **101**, 042003 (2020).

[59] R. Ormiston, T. Nguyen, M. Coughlin, R. X. Adhikari, and E. Katsavounidis, Noise reduction in gravitational-wave data via deep learning, Phys. Rev. Research **2**, 033066 (2020).

[60] M. Maggiore, *Gravitational Waves: Volume 1: Theory and Experiments*, Gravitational Waves (OUP Oxford, 2008).

[61] L. S. Finn and D. F. Chernoff, Observing binary inspiral in gravitational radiation: One interferometer, Phys. Rev. D **47**, 2198 (1993).

[62] H. T. Cromartie, E. Fonseca, S. M. Ransom, P. B. Demorest, Z. Arzoumanian, H. Blumer, P. R. Brook, M. E. DeCesar, T. Dolch, J. A. Ellis, R. D. Ferdman, E. C. Ferrara, N. Garver-Daniels, P. A. Gentile, M. L. Jones, M. T. Lam, D. R. Lorimer, R. S. Lynch, M. A. McLaughlin, C. Ng *et al.*, Relativistic Shapiro delay measurements of an extremely massive millisecond pulsar, Nat. Astron. **4**, 72 (2020).

[63] B. P. Abbott, R. Abbott, T. D. Abbott, S. Abraham, F. Acernese, K. Ackley, C. Adams, R. X. Adhikari *et al.*, GW190425: Observation of a Compact Binary Coalescence with Total Mass $\sim 3.4\ M_{\odot}$, Astrophys. J. Lett. **892**, L3 (2020).

[64] T. Tsutsui, A. Nishizawa, and S. Morisaki, Early warning of precessing compact binary merger with third-generation gravitational-wave detectors, arXiv:2011.06130 [Phys. Rev. D (to be published)].

[65] F. Chollet *et al.*, Keras, https://keras.io (2015).

[66] 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 *et al.*, TensorFlow: Large-scale machine learning on heterogeneous systems (2015), software available from tensorflow.org.

[67] B. Kiziltan, A. Kottas, M. De Yoreo, and S. E. Thorsett, The neutron star mass distribution, Astrophys. J. **778**, 66 (2013).

[68] Because of the finite computation time, the sliding of the detection window is not exactly continuous. However, as we will see later in Sec. V and in Fig. 13, the prediction time on each 256-s segment is only 100 ms, suggesting that the sliding can be done in time steps of 0.25 ms or 0.5 s, much shorter compared to the local merger time [Eq. (4)].

[69] There are four instead of eight independent degrees of $\theta(t)$ because only a specific linear combination of the input and end test masses, known as the "hard mode" [70], requires a high control bandwidth ($\sim$3 Hz) and thus injects a significant amount of sensing noise at $\gtrsim$10 Hz Buikema:20. The orthogonal combination, or the "soft mode," only requires to be controlled with a bandwidth of $\sim$0.5 Hz and has negligible contribution to the noise.

[70] J. A. Sidles and D. Sigg, Optical torques in suspended Fabry Perot interferometers, Phys. Lett. A **354**, 167 (2006).

[71] More specifically, we simulate the fast channels $(\theta(t))$ with similar characteristics as LIGO auxiliary channels like `H1:ASC-CHARD _P _OUT _DQ`. The slow channels are designed to mimic auxiliary channels like `H1:ASC-ADS _PIT4 _DOF _OUT _DQ` for the dithering based spot position sensors, and `H1:SUS-ITMX _L3 _OPLEV _PIT _OUT _DQ` for the optical lever outputs.

[72] Note the angular motion sensed by the optical levers are due to low-frequency seismic motion. It is the source for spot-position motion $x_{\mathrm{spot}}(t)$ and should be distinguished from the high-frequency $(>10$ Hz$)$ angular motion $\theta(t)$ due to sensing noises in the control feedback.

[73] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. **12**, 2825 (2011), https://www.jmlr.org/papers/v12/pedregosa11a.html.

[74] For the current implementation, we use only a single CPU (no parallelization). When running the full prediction code (noise subtraction and signal classification), it takes about 15 GB of GPU memory and 2 GB of CPU memory. More dedicated optimization on the implementation of the routine is deferred to future works.

[75] https://gstlal.docs.ligo.org/ewgw-data-release/data-full.html.

[76] R. Essick, P. Godwin, C. Hanna, L. Blackburn, and E. Katsavounidis, iDQ: Statistical inference of non-gaussian noise with auxiliary degrees of freedom in gravitational-wave detectors, arXiv:2005.12761.

[77] E. Cuoco, J. Powell, M. Cavaglià, K. Ackley, M. Bejger, C. Chatterjee, M. Coughlin, S. Coughlin, P. Easter, R. Essick, H. Gabbard, T. Gebhard, S. Ghosh, L. Haegel, A. Iess, D. Keitel, Z. Marka, S. Marka, F. Morawski, T. Nguyen *et al.*, Enhancing gravitational-wave science with machine learning, Mach. Learn. Sci. Tech. **2**, 011002 (2021).

[78] https://www.gw-openscience.org/auxiliary/GW170814/.