DEBIASED CONTRASTIVE LEARNING WITH MULTI-RESOLUTION KOLMOGOROV-ARNOLD NETWORK FOR GRAVITATIONAL WAVE GLITCH DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Time-series gravitational wave glitch detection presents significant challenges for machine learning due to the complexity of the data, limited labeled examples, and data imbalance. To address these issues, we introduce Debiased Contrastive Learning with Multi-Resolution Kolmogorov-Arnold Network(dcMltR-KAN), a novel self-supervised learning (SSL) approach that enhances glitch detection, robustness, explainability, and generalization. dcMltR-KAN consists of three key novel components: Wasserstein Debiased Contrastive Learning (wDCL), a CNN-based encoder, and a Multi-Resolution KAN (MltR-KAN). The wDCL improves the model's sensitivity to data imbalance and geometric structure. The CNN-based encoder eliminates false negatives during training, refines feature representations through similarity-based weighting (SBW), and reduces data complexity within the embedding. Additionally, MltR-KAN enhances explainability, generalization, and efficiency by adaptively learning parameters. Our model outperforms widely used baselines on O1, O2, and O3 data, demonstrating its effectiveness. Extending dcMltR-KAN to other time-series benchmarks underscores its novelty and efficiency, marking it as the first model of its kind and paving the way for future SSL and astrophysics research.

1 INTRODUCTION

Gravitational waves are ripples in spacetime caused by some of the most violent and energetic events in the universe, such as merging black holes, colliding neutron stars, and supernovae. Predicted by Albert Einstein in 1915 as part of his theory of general relativity, these waves carry crucial information about their origins and the nature of gravity itself. Their detection, along with the identification of glitches, offers a new way to observe the cosmos, revealing insights unattainable through traditional methods like light-based telescopes (Bi et al., 2024).

Glitches, short-lived noise transients from environmental or instrumental sources, closely mimic complex gravitational wave signals, making it difficult to distinguish real events from noise. The groundbreaking detection of gravitational waves by LIGO and Virgo in 2015 marked the beginning of a new era in astrophysics, allowing scientists to explore phenomena previously undetectable (Bailes et al., 2021). Since then, identifying and reducing gravitational wave glitches has become increasingly important in both astrophysics and deep learning (Chowdhury, 2024).

However, the unique characteristics of gravitational wave data pose significant challenges for current deep learning models, particularly in terms of data complexity, imbalance, limited labeled data, and explainability, making it difficult to accurately detect meaningful glitches. Gravitational wave raw data is inherently high-dimensional and multi-resolution, with signals captured at high sampling rates across numerous sensors and frequencies (Chua et al., 2019).

This vast, non-stationary complex data, where noise and signal properties change over time, makes it difficult for deep learning models to focus on relevant features across various scales (George & Huerta, 2018; Chowdhury, 2024). The presence of instrumental and environmental noise further complicates the distinction between real gravitational wave signals and glitches. Additionally, data imbalance—with true gravitational wave events being rare compared to the overwhelming volume of noise—leads to model overfitting on the dominant noise class, limiting detection accuracy.

Limited labels pose another obstacle to deep learning models, which typically require large amounts of labeled data for effective training. In gravitational wave glitch detection, labeled data is scarce due to the expert knowledge and resources needed for annotation (Miller & Yunes, 2019). Consequently, models often face generalization issues like underfitting or overfitting when applied to small or imbalanced datasets (Powell et al., 2023; Cuoco et al., 2020).

The lack of explainability in deep learning models is also a major limitation. As black boxes, they offer little insight into their predictions, reducing trust in gravitational wave detection where scientific validation is critical. This lack of transparency hampers collaboration with experts and complicates identifying biases or errors. These challenges highlight the shortcomings of deep learning in effectively detecting and explaining meaningful gravitational wave glitches.

To address these challenges, we introduce Debiased Contrastive Learning with Multi-Resolution Kolmogorov-Arnold Network (dcMltR-KAN), a novel self-supervised learning (SSL) framework designed to learn meaningful representations of gravitational wave data. A key component of this framework is our proposed Multi-Resolution Kolmogorov-Arnold Network (MltR-KAN), which is uniquely suited to capture multi-scale patterns and complexities inherent in gravitational wave data. Inspired by Kolmogorov's superposition theorem, MltR-KAN employs wavelet basis functions and hierarchical structures to enhance explainability, generalization, and efficiency. By leveraging the additivity and learned parameters of KAN, our approach provides interpretable insights into the detection of glitches while improving learning generalization and efficiency.

Problem formulation. Unlike fully-supervised deep learning approach, dcMltR-KAN offers the solution to the following problem: Given a dataset $\mathcal{X} = \{x_i\}_{i=1}^N$ of unlabeled gravitational wave signals, where each $x_i \in \mathbb{R}^d$ is a d-dimensional feature vector extracted from the raw time-series data over the period interval T, the goal is to learn a representation function $f : \mathbb{R}^d \to \mathbb{R}^k$ that maps signals to a k-dimensional feature space where meaningful signal and glitch patterns are captured.

dcMltR-KAN learns meaningful data representations through three novel components: Wasserstein Debiased Contrastive Learning (wDCL), a CNN-based encoder, and a multi-resolution KAN (MltR-KAN). It leverages wDCL to effectively handle both data complexity and the scarcity of labeled data, while increasing sensitivity to the geometric structure of gravitational wave glitch data and managing data imbalance. The CNN-based encoder is employed to conduct false negative elimination (FNE) in training, optimize feature representation through similarity-based weighting (SBW), and reduce data complexity via feature extraction in the embedding. MltR-KAN leverages wavelet basis functions to capture complex, multi-scale patterns in gravitational wave data. Combined with hierarchical learning, multi-resolution FNE and SBW, and learned parameters, this approach enhances glitch detection while improving efficiency, explainability, and generalization.

The proposed model surpasses other supervised deep learning approaches on the benchmark O1, O2, and O3 datasets (Abbott et al., 2019; 2021; 2023) demonstrating its superior performance and advantages. Furthermore, we extend dcMltR-KAN to benchmark audio data, showcasing its superiority. The dcMltR-KAN is the first model to integrate multi-resolution KAN into SSL, and it will inspire future research in SSL and astrophysics.

2 RELATED WORK

Deep learning is widely used in gravitational wave glitch detection for their capabilities in finding complex relationships and handling large-scale data.

Generative Adversarial Networks (GANs): Powell et al. (2023) employed GAN with advanced applications in glitch detection. Dooney et al. (2022) introduced a dual-discriminator approach for time-domain signal generation, improving convergence in gravitational wave detection. However, GANs rely on large and diverse datasets for effective training, and their performance may be limited by the availability of sufficient high-quality labeled data in the gravitational wave domain.

Convolutional Neural Networks (CNNs): Razzano & Cuoco (2018) developed a CNN pipeline that efficiently classified detector glitches based on their time-frequency representations, achieving high accuracy, especially in simulated data. Fernandes et al. (2023) improved glitch classification by using transfer learning with advanced CNN architectures like ConvNeXt. Alvarez-Lopez et al. (2023) integrated CNNs within a decision tree framework, showing robustness in diverse noise environments. CNNs, however, often struggle with overfitting and handling diverse noise conditions, limiting real-world generalization (Schäfer et al., 2023).

Variational Autoencoders (VAEs) and SSL: Sakai et al. (2022) applied a VAE with invariant information clustering (IIC) to classify transient noises by learning from 2D image features, aligning well with existing annotations. However, VAEs often produce blurry reconstructions and struggle with non-linear data, resulting in suboptimal feature representation and classification accuracy. Fernandes et al. (2023) combined SSL with CNNs to generate pseudo-labels for the Gravity Spy dataset, showing promise but falling short of supervised methods in accuracy. SSL methods often underperform fully supervised approaches, particularly in detecting subtle glitches in noisy environments.

3 DEBIASED CONTRASTIVE LEARNING WITH MULTI-RESOLUTION KAN

Figure 1 illustrates the proposed novel self-supervised learning (SSL) model, dcMltR-KAN, which enhances debiased contrastive learning using a multi-resolution KAN.

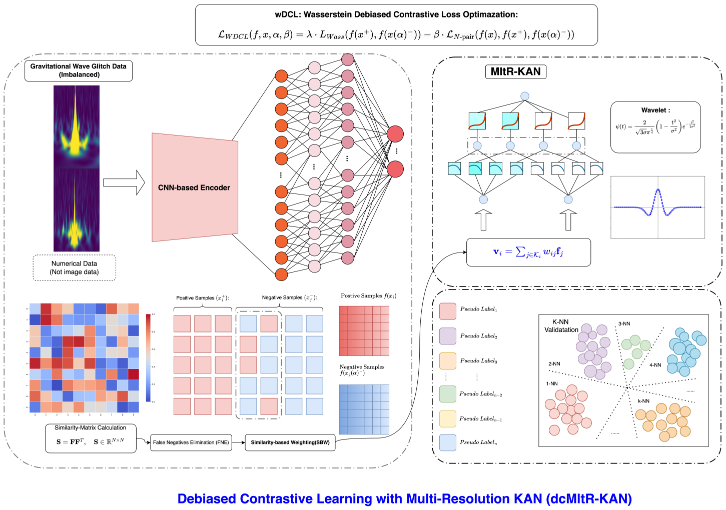
wDCL: Wasserstein Debiased Contrastive Loss Optimazation:
$\mathcal{L}_{WDCL}(f, x, \alpha, \beta) = \lambda \cdot L_{Wass}(f(x^+), f(x(\alpha)^-)) - \beta \cdot \mathcal{L}_{N\text{-pair}}(f(x), f(x^+), f(x(\alpha)^-))$
Gravitational Wave Glitch Data (Imbalanced)
Numerical Data (Not image data)
CNN-based Encoder
MltR-KAN
Wavelet :
$\psi(t) = \frac{2}{\sqrt{3\sigma}\pi^{\frac{1}{4}}}\left(1 - \frac{t^2}{\sigma^2}\right)e^{-\frac{t^2}{2\sigma^2}}$
$\mathbf{v}_i = \sum_{j \in \mathcal{K}_i} w_{ij}\mathbf{f}_j$
Positive Samples (x_i^+):
Negative Samples (x_j^-):
Postive Samples $f(x_i)$
Negative Samples $f(x_j(\alpha)^-)$
Similarity-Matrix Calculation
$\mathbf{S} = \mathbf{F}\mathbf{F}^T, \quad \mathbf{S} \in \mathbb{R}^{N \times N}$
False Negatives Elimination (FNE)
Similarity-based Weighting(SBW)
$Pseudo\ Label_1$
$Pseudo\ Label_2$
$Pseudo\ Label_3$
$Pseudo\ Label_{n-2}$
$Pseudo\ Label_{n-1}$
$Pseudo\ Label_n$
K-NN Validatation
1-NN
2-NN
3-NN
4-NN
k-NN
Debiased Contrastive Learning with Multi-Resolution KAN (dcMltR-KAN)

Figure 1: Debiased Contrastive Learning with Multi-Resolution Kolmogorov-Arnold Network

dcMltR-KAN. The model comprises three novel components. The first is the proposed Wasserstein Debiased Contrastive Learning (wDCL) that optimizes embeddings and addresses data imbalance by capturing the geometric structure of the input data.

The second is a CNN-based encoder that eliminates false negatives during training, refines feature representations in the embedding through similarity-based weighting, uncovers latent features in the input data while removing potential noise, reduces dimensionality, and lowers computational complexity in SSL.

The third component includes multi-resolution KAN (MltR-KAN) layers, which decompose the time-series glitch data into multiple resolutions using wavelets (e.g., Haar) within the KAN architecture to enhance explainability, efficiency, and generalization. Notably, SSL principles are applied across all three components: wDCL, the CNN-based encoder, and MltR-KAN. We describe each component in detail below.

3.1 WASSERSTEIN DEBIASED CONTRASTIVE LEARNING (WDCL)

wDCL extends debiased contrastive learning (DCL) by embedding the Wasserstein distance into the loss function, enhancing the sensitivity to the underlying geometric structure of the data and addressing data imbalance while maintaining the benefits of DCL.

Contrastive Learning and Debiased Contrastive Learning: Contrastive learning aims to minimize the distance between embeddings of positive sample pairs (x, x^+) (similar data points), which are the transformed representations of input data after passing through an encoder, and maximize the distance for negative sample pairs (x, x^-)(dissimilar data points), thereby improving the model's ability to differentiate between classes Chopra et al. (2005). Let $p_{\text{data}}(x)$ represent the data distribution of samples x, and let $p_{\text{pos}}(x, x^+)$ denote the distribution of positive pairs, the objective function for the encoder $f(x)$, which produces an L_2-normalized feature vector, is defined as:$\mathcal{L}_{\text{contrastive}} = \mathbb{E}_{(x,x^+)\sim p_{\text{pos}}}\left[-\log \frac{e^{f(x)^\top f(x^+)/\tau}}{\sum_{i=1}^{M} e^{f(x)^\top f(x_i^-)/\tau}}\right]$, where x_i^- are negative samples drawn from p_{data}, M is the number of negative pairs, and τ denotes the temperature parameter that controls the smoothness or sharpness of the similarity distribution in the softmax function (Wang & Isola, 2022; Chuang et al., 2020).

Contrastive learning assumes that all negative samples x^- are true negatives dissimilar to the anchor sample x. However, in practice, some negative samples might actually be similar to the anchor (false negatives), introducing bias and degrading model performance. Debiased contrastive learning (DCL) addresses this issue by adjusting the loss function to account for the probability that a negative sample might be a false negative. The key idea is to re-weight the contribution of each negative sample based on its similarity to the anchor, thereby reducing the bias introduced by false negatives: $\mathcal{L}_{\text{debiased}} = \mathbb{E}_{(x,x^+)\sim p_{\text{pos}}}\left[-\log \frac{e^{f(x)^\top f(x^+)/\tau}}{e^{f(x)^\top f(x^+)/\tau} + \sum_{i=1}^{M}\left(e^{f(x)^\top f(x_i^-)/\tau} - \gamma e^{2f(x)^\top f(x_i^-)/\tau}\right)}\right]$, where γ is a scaling factor representing the probability of a negative sample being a false negative.

Wasserstein-based Debiased Contrastive Learning (wDCL): The DCL loss function overcomes the negative pair selection limitations in constrative learning. However, the $\mathcal{L}_{\text{debiased}}$ can be sensitive to imbalanced data distributions, where negative samples may actually be similar to the anchor (false negatives). Furthermore, it can be hard for $\mathcal{L}_{\text{debiased}}$ to capture the true structure of the data, because of the limitations of Euclidean-based distances.

We propose Wasserstein-based Debiased Contrastive Learning by incorporating the Wasserstein distance into the debiased loss function to handle the challenge, forming the foundation of our dcMltR-KAN model. Unlike Euclidean or similar metrics, the Wasserstein distance captures the true geometric structure of real-world data by measuring the discrepancy between two probability distributions based on the geometry of the data space. This makes it particularly effective in high-dimensional settings. Additionally, its evaluation of entire distributions, rather than isolated points, allows it to handle imbalances between positive and negative samples more robustly. The Wasserstein distance outperforms KL-divergence by handling disjoint supports, avoiding divergence issues. It measures the "cost" of transforming one distribution into another, is robust to outliers, and ensures unbiased, symmetric comparisons, making it ideal for imbalanced datasets in contrastive learning.

The Wasserstein distance between two distributions μ and ν over a metric space $\mathcal{X}$ is defined as:

$$W(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} d(x, y)\, d\gamma(x, y), \tag{1}$$

where $\Gamma(\mu, \nu)$ denotes the set of all joint distributions γ with marginals μ and ν, and $d(x, y)$ represents the metric on space $\mathcal{X}$. Computing the Wasserstein distance is challenging due to optimization over joint distributions. We use the Sinkhorn divergence with entropy regularization for efficiency, preserving sensitivity to distributions (details in Appendix)

Wasserstein contrastive loss $\mathcal{L}_{\text{wdcl}}$: We can build the Wasserstein contrastive loss by integrating Wasserstein distance with $\mathcal{L}_{\text{debiased}}$. However, $\mathcal{L}_{\text{debiased}}$ may not handle multiple pairs of negatives and can not handle multiple similar and dissimilar pairs in high-dimensional spaces, besides more computing in optimization. We tackle this challenge by integrating Wasserstein distance with N-pair contrastive loss: $\mathcal{L}_{N\text{-pair}} = -\log \frac{e^{f(x)^\top f(x^+)}}{e^{f(x)^\top f(x^+)} + \sum_{i=1}^{N-1} e^{f(x)^\top f(x_i^-)}}$ as,

$$\mathcal{L}_{\text{wdcl}}(f, x, \alpha, \beta) = \lambda \cdot \mathcal{L}_{\text{wass}}(f(x^+), f(x(\alpha)^-)) - \beta \cdot \mathcal{L}_{N\text{-pair}}(f(x), f(x^+), f(x(\alpha)^-)) \tag{2}$$

Here $\mathcal{L}_{\text{wass}}(f(x^+), f(x(\alpha)^-))$ is the the Wasserstein loss capturing the geometric structure of positive and negative sample pairs, which is calculated by the Sinkhorn-divergence-based method: $L_{\text{wass}}(f(x^+), f(x^-)) = \text{Sinkhorn-divergence}(f(x^+), f(x^-))$ (Appendix). This explicit L_{wass} term integrates the Wasserstein distance into the contrastive learning objective, ensuring robustness to data imbalance and alignment with the geometric structure of the feature space.

The scaling coefficients λ and β are employed to modulate the Wasserstein and N-pair contrastive objectives, respectively. The term $x(\alpha)^-$ refers to a subset of negative samples determined by the parameter α. These are most likely hard negatives, selected to reduce noise and redundancy, sharpening the embedding space and improving class separability.

The hyperparameter α acts as a discriminative threshold to filter through negative pairs, refining the model's emphasis on important contrasts. The more details about α can be found in the following False Negatives Elimination (FNE) part in section 3.2. It is recommended to select $\lambda = \beta = 1$ by default, but it can be adjusted according to input data.

Positive sample generation. In our SSL setting, positive samples are generated by adding Gaussian noise to the input data, creating slightly varied versions of the same datapoint. This approach preserves key data characteristics while introducing variability, enhancing contrastive learning and enabling the model to better generalize across noisy and imperfect data—an advantage for glitch detection.

Dynamic α adjustment: We recommend to dynamically adjust the α according to input data size, which helps optimizing negative sample selection and avoid the extreme cases where those "easier" negative samples, which are already different from the anchor, are selected if a too large α is picked. For relatively small datasets, a smaller α preserves informative hard-negatives, preventing possible overfitting and enhancing generalization.

For larger datasets, a larger α filters out more negatives, reducing false negatives and focusing on informative hard-negatives, which are negative samples highly similar to the anchor but belong to a different class. For instance, We set from 0.1% to 4% for the O1 dataset (41,717 samples) and from 2% to 10% for larger datasets like O2 and O3 (134,372 and 500,524 samples, respectively in this study), By dynamically adjusting α based on the dataset size, our model adapts its learning strategy to balance easy and hard negatives, enhancing generalization and feature discrimination across task. Theorem 1 proves the robustness of wDCL to data imbalance from a loss function perspective.

Theorem 1 (Robustness of wDCL to imbalance): $\mathcal{L}_{\text{wdcl}}$ is robust to data imbalance than $\mathcal{L}_{\text{debiased}}$.

3.2 CNN-BASED ENCODER

Rationale: After applying wDCL, which uses the Wasserstein distance to optimize embeddings and address data imbalance by capturing the geometric structure of the data, there remains a need to refine false negative samples used in training. Furthermore, how to enhance feature representation quality for learning remains another challenge.

CNN-based encoder. We propose a CNN-based encoder, inspired by (Wu et al., 2018), to address these challenges using False Negatives Elimination (FNE) and Similarity-based Weighting (SBW) techniques. The encoder extracts relevant features, such as transient signal patterns or glitches, while filtering out noise in the embedding. Through convolutional, pooling, and dense layers, it captures both local and global patterns while reducing dimensionality in gravitational wave data.

Feature similarity matrix calculation. FNE in training and following SBW-based data representation optimization both rely on the calculation of a feature similarity matrix $\mathbf{S}$ that evaluate proximity between positive and negative pairs as follows.

The CNN-based encoder f maps each input sample x_i into a feature vector $\mathbf{f}_i$: $\mathbf{f}_i = f(x_i), \quad \mathbf{f}_i \in \mathbb{R}^d, \quad i = 1, 2, \ldots, N$. The feature vectors $\mathbf{f}_i$ are further stacked to form the feature matrix $\mathbf{F}$:

$$\mathbf{F} = \begin{bmatrix} \mathbf{f}_1 & \mathbf{f}_2 & \cdots & \mathbf{f}_N \end{bmatrix}^\top = \begin{bmatrix} f(x_1) & f(x_2) & \cdots & f(x_N) \end{bmatrix}^\top, \quad \mathbf{F} \in \mathbb{R}^{N \times d} \tag{3}$$

We then calculate the matrix $\mathbf{S}$ to evaluate the similarity between positive and negative pairs, enabling the model to optimize the embeddings and assess how well the wDCL's learned feature space represents the underlying structure of the data under the Wasserstein contrastive loss:

$$\mathbf{S} = \mathbf{F}\mathbf{F}^T, \quad \mathbf{S} \in \mathbb{R}^{N \times N} \tag{4}$$

where each element of $\mathbf{S}$ is given by the dot product between feature vectors: $\mathbf{S}_{ij} = \mathbf{f}_i^\top \mathbf{f}_j$, where each feature vector is normalized as $\mathbf{f}_i = \frac{\mathbf{f}_i}{\|\mathbf{f}_i\|_2}$ to remove noise and possible outliers. This normalization ensures that the similarity, which is calculated as a cosine similarity, is based on the shape of the vectors rather than their magnitude, particularly important for robustly comparing noisy samples.

False Negatives Elimination (FNE). To enhance the quality of embeddings and eliminate false negatives during training, we utilize the matrix $\mathbf{S}$ to identify and exclude negative samples that are excessively similar to the anchor sample. For each anchor x_i, rank all other samples x_j ($j \neq i$) by their similarity S_{ij}. Then we eliminate top α fraction by defining $\alpha \in [0,1]$ adaptive to input data size as the elimination ratio (see dynamic α adjustment subsection in 3.1). With $N_{\text{neg}} = N(N-1)$, we remove the top $\alpha \times N_{\text{neg}}$ negatives with the highest similarity scores. By removing false negatives, the refined contrastive loss $\mathcal{L}_{\text{wdcl}}$ converges faster and improves downstream task performance. Hence, FNE leads to a reduction in the overall loss, proving that it enhances SSL training quality. Mathematically, it means the following proposition holds:

Proposition 1: $\mathcal{L}_{\text{wdcl}}$ with FNE is lower than the loss without FNE: $\mathbb{E}_{(x,x(\alpha)^-)}\left[\mathcal{L}_{\text{wdcl}}^{\text{FNE}}\right] < \mathbb{E}_{(x,x^-)}\left[\mathcal{L}_{\text{wdcl}}^{\text{no FNE}}\right]$, where α is the the elimination ratio.

Similarity-Based Weighting (SBW). SBW refines feature representations in the embedding by leveraging the relationships between similar data points. We leverage the similarity matrix $\mathbf{S}$ to get $\mathbf{s}_i$, representing the similarities from sample x_i to the others. We then select top k similar samples by picking indices $\mathcal{K}_i$, which correspond the top $k$highest similarity scores in $\mathbf{s}_i$. Next, we compute weights for selected indices as: $w_{ij} = \exp(s_{ij}), \quad \forall j \in \mathcal{K}_i$ to emphasize the similarity and ensure positive weights. We then aggregate the top k features vectors: $\mathbf{v}_i = \sum_{j \in \mathcal{K}_i} w_{ij}\mathbf{f}_j$. The aggregated feature vector $\mathbf{v}_i$ will replace the original sample x_i to optimize its feature representation. This weighted aggregation improves the quality of the data for subsequent models (e.g., multiR-KAN). As such, SBW refines feature representations, leading to a lower expected contrastive loss by improving the quality of the learned embeddings. Proposition 2 demonstrates the impact of SBW on the expected wDCL loss, suggesting that the aggregated feature vector produced by SBW enhances the quality of data representation, leading to improved model performance.

Proposition 2: The loss $\mathcal{L}_{\text{wdcl}}$ with SBW is lower than the loss without SBW: $\mathbb{E}_{(x,\mathbf{v}_i)}\left[\mathcal{L}_{\text{wdcl}}^{\text{SBW}}\right] < \mathbb{E}_{(x,x^-)}\left[\mathcal{L}_{\text{wdcl}}^{\text{no SBW}}\right]$, where $\mathbf{v}_i$ is the aggregated feature vector obtained from the top k most similar samples through SBW.

3.3 MULTI-RESOLUTION KOLMOGOROV-ARNOLD NETWORK (MLTR-KAN)

Rationale: While FNE and SBW refine the data representation in the embedding, they do not directly address the multi-scale patterns inherent in gravitational wave data. To handle this, we integrate a Multi-Resolution KAN, following the CNN-based encoder, into our SSL framework. MltR-KAN is a two-layer KAN with wavelet basis functions that provide built-in multi-resolution analysis and learnable parameters (Liu et al., 2024). This will enhance the explainability, efficiency, and generalization of the SSL model.

Formulation: Given a feature vector from the CNN-encoder $\mathbf{f}_i = f_{\text{CNN-encoder}}(x_i), \quad \mathbf{f}_i \in \mathbb{R}^n$, MltR-KAN performs the following mapping:

$$\mathbf{F}^{(l)}(\mathbf{f}_i^{(l)}) = \sum_{q=0}^{2n} \chi_q^{(l)} \left(\sum_{p=1}^{n} \psi_{pq}^{(l)}(\mathbf{f}_{ip}^{(l)}) \right), \tag{5}$$

Here, $\mathbf{F}^{(l)}(\mathbf{f}_i^{(l)})$ denotes the transformed feature vector at resolution level l, based on the input $\mathbf{f}_i^{(l)}$, and the number of resolution levels L is set such that $2^L \leq |\mathbf{f}_i|$. $\mathbf{f}_{ip}^{(l)}$ is the p-th component of $\mathbf{f}_i^{(l)}$, n is the total number of components (features) at each resolution level, $\psi_{pq}^{(l)}$ are the wavelet basis functions (e.g., 'db4') applied to the components of the input signal at resolution level l, and $\chi_q^{(l)}$ are the parameters to be learned by the MltR-KAN.

Parameters learning. The parameters χ_q in the mltR-KAN are learned during the SSL training using backpropagation and an optimizer (e.g., SGD). The learning process involves minimizing the Wasserstein-based debiased contrastive loss $\mathcal{L}_{\text{wdcl}}$ with respect to these parameters: $\min_{\{\chi_q\}} \mathcal{L}_{\text{wdcl}}(\mathbf{F}(\mathbf{f}_i))$. Specifically, mltR-KAN's parameters $\chi_q^{(l)}$ are updated to minimize the loss at each resolution level l.

Wavelet selection. For effective glitch detection, wavelets should be orthogonal for precise signal reconstruction, smooth with compact support to localize transient glitches, and computationally efficient for handling large datasets. We recommend 'db4', 'sym4', or similar. The 'db4' wavelet is defined by scaling coefficients h_n and wavelet coefficients $g_n = (-1)^n h_{3-n}$, with scaling function $\phi(t) = \sum_{n=0}^{3} h_n \phi(2t-n)$ and wavelet function $\psi(t) = \sum_{n=0}^{3} g_n \phi(2t-n)$. If computational speed is critical, the Haar ('db1') wavelet is orthogonal and compact but lacks smoothness. Its simplicity makes it ideal for detecting short, abrupt glitches like "blips" in gravitational wave data (Robson & Cornish, 2019).

Enhance explainability and efficiency. MltR-KAN leverages hierarchical learning, feature weighting, and multi-resolution FNE. MltR-KAN achieves hierarchical learning by decomposing feature representation into at multiple resolution levels l. It brings a hierarchical loss structure: $\mathcal{L}_{\text{wdcl}} = \sum_{l=1}^{L} \mathcal{L}^{(l)}$, where $\mathcal{L}^{(l)}$ represents the loss contribution from resolution level l. This decomposition enhances explainability by providing insights into how features at different resolutions contribute to the final prediction. Figure S1 in the Appendix illustrates the explainability enhancement process within the MltR-KAN model for the SNR feature extracted from the O1 data.

Moreover, feature weighting is incorporated to further refine the model's performance: $\mathcal{L}_{\text{weighted}} = \sum_{l=1}^{L} \omega^{(l)} \mathcal{L}^{(l)}$, where the learned weights $\omega^{(l)}$ adjust the contribution of each resolution level to the overall loss. These weights, distinct from the core MltR-KAN parameters (e.g., $\chi_q^{(l)}$), are optimized during training to balance multi-scale features and improve the model's efficiency and performance.

MltR-KAN reduces false negatives by leveraging multi-resolution features that capture fine details missed by single-scale models. Its hierarchical loss structure, $\mathcal{L}_{\text{FNE}} = \sum_{l=1}^{L} \mathcal{L}_{\text{FNE}}^{(l)}$, minimizes false negatives at each resolution. By focusing on fine-grained levels, the model captures subtle patterns across scales, effectively lowering the false negative rate. Similarly, we have the following result indicating that applying SBW before MltR-KAN reduces the overall loss. Figures S2 and S3 in the Appendix demonstrate a simulated training scenario under hierarchical loss where SBW or FNE is integrated within MltR-KAN. Furthermore, Proposition 3 demonstrates that applying SBW before MltR-KAN across all resolution levels reduces the overall expected wDCL loss compared to not applying SBW.

Proposition 3: Hierarchical feature representation with Multi-Resolution SBW. Let the overall loss be $\mathcal{L}_{\text{wdcl}} = \sum_{l=1}^{L} \mathcal{L}^{(l)}$, where $\mathcal{L}^{(l)}$ is the contribution from resolution level l, and $\omega^{(l)}$ are learned weights. With multi-resolution SBW applied before MltR-KAN, the refined feature representation $\mathbf{v}_i$ at each scale l leads to:

$$\mathbb{E}_{x_i \sim p_{\text{data}}} \left[\mathcal{L}_{\text{wdcl}}^{\text{SBW}} \right] < \mathbb{E}_{x_i \sim p_{\text{data}}} \left[\mathcal{L}_{\text{wdcl}}^{\text{no SBW}} \right]. \tag{6}$$

MltR-KAN has a lower norm-based Rademacher complexity than KAN using B-spline basis functions and MLP, suggesting better generalization capability. Theorem 2 further supports that MltR-KAN offers better generalization than the standard KAN using B-spline basis functions and MLP, specifically in terms of norm-based Rademacher complexity.

Definition: Norm-based Rademacher Complexity of a hypothesis class $\mathcal{H}$ over a sample $S = \{x_1, \ldots, x_n\}$ is defined as:$\hat{\mathcal{R}}_n(\mathcal{H}) = \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}, \|h\| \le C} \frac{1}{n} \sum_i \sigma_i h(x_i) \right]$, where $\sigma_i \in \{-1, 1\}$ are independent Rademacher variables and $\|h\| \le C$ constrains the function norm. It quantifies the capacity of $\mathcal{H}$ to fit random noise, with lower values indicating better generalization.

Theorem 2: Let $\mathcal{F}_{\text{KAN-W}}$, $\mathcal{F}_{\text{KAN-S}}$, and $\mathcal{F}_{\text{MLP}}$ represent the hypothesis classes of KAN with wavelet basis, B-spline basis, and MLP, respectively. The norm-based Rademacher complexity of these classes satisfies:

$$\mathcal{R}_n(\mathcal{F}_{\text{KAN-W}}) \prec \mathcal{R}_n(\mathcal{F}_{\text{KAN-S}}) \prec \mathcal{R}_n(\mathcal{F}_{\text{MLP}}), \tag{7}$$

where $\prec$ denotes strict inequality.

dcMltR-KAN generalization. We have proved that the dcMltR-KAN model with wavelet basis functions exhibits a lower upper bound on the generalization error compared to dcMltR-KAN with spline basis functions or the dc-MLP model, where MltR-KAN is replaced by MLP in the proposed SSL model, as established in Theorem 2.

Theorem 3: Let $\mathcal{F}_{\text{dcMltR-KAN-W}}$, $\mathcal{F}_{\text{dcMltR-KAN-S}}$, and $\mathcal{F}_{\text{dc-MLP}}$ represent the hypothesis classes of dcMltR-KAN with wavelet basis, B-spline basis, and dc-MLP model, respectively. The upper-bound on the generalization error for these models satisfies:

$$\mathcal{E}_{\text{gen}}(\mathcal{F}_{\text{dcMltR-KAN-W}}) \prec \mathcal{E}_{\text{gen}}(\mathcal{F}_{\text{dcMltR-KAN-S}}) \prec \mathcal{E}_{\text{gen}}(\mathcal{F}_{\text{dc-MLP}}), \tag{8}$$

where $\mathcal{E}_{\text{gen}}(\cdot)$ denotes the generalization error, and $\prec$ signifies strict inequality.

4 RESULTS

We evaluate our proposed dcMltR-KAN on benchmark gravitional wave dataset (O1,O2 and O3 (Abbott et al., 2019; 2021; 2023)), besides extending it to other time-series data.

Data and preprocessing. We employed the benchmark O1, O2, and O3 data preprocessed from the Gravity Spy project (Glanzer et al., 2021). The preprocessed data are typically not full time-series but rather a condensed form containing key extracted features. However, challenges inherent to the original data—such as data complexity, noise, class imbalance, and the potential loss of certain temporal dynamics—can still persist Bahaadini et al. (2018).

This preprocessing derived 33 meaningful features (such as trigger timing, peak frequency, signal-to-noise ratio (SNR), amplitude, and bandwidth) from the original high-dimensional gravitational wave data, resulting in O1 with 41,717 samples and 22 glitch types, O2 with 134,372 samples and 22 types, and O3 with 500,524 samples and 24 types. Figure 2 illustrates the glitch types and their distributions for each dataset, showing that different datasets have different dominant types, with the percentage of the smallest groups reaching as low as 0.02% (e.g., Chirp in O2 and O3).

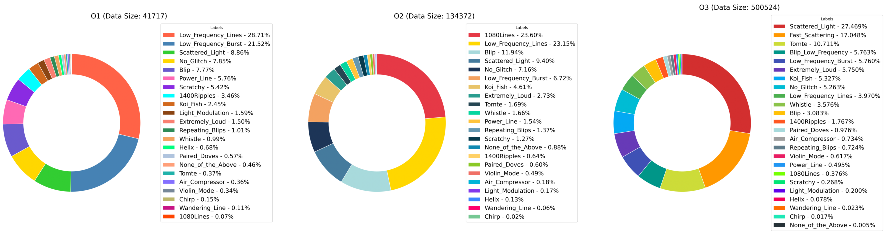

Figure 2: The imbalanced Glitch type distribution across the preprocessed O1, O2, and O3 datatsets

Baselines: We compare dcMltR-KAN with widely-used fully-supervised baselines (CNN, GRU, ResNet, GAN-DNN, Transformer) and three SOTA SSL models: CPC (Contrastive Predictive Coding), TS-TCC (Time-Series Representation Learning via Temporal and Contextual Contrasting), and SimCLR (Simple Contrastive Learning of Representations) (van den Oord et al., 2018; Eldele et al., 2021; Chen et al., 2020). Supervised models were trained with an 80/20 train-test split, with details in the Appendix. While CPC and TS-TCC are tailored for time-series data, SimCLR, though originally designed for other domains, has been adapted for such tasks (Zhang et al., 2022). Like its SSL peers, dcMltR-KAN is evaluated using top-1 results from a k-NN classifier on learned representations. Our implementation features a CNN-based encoder with two convolutional layers, a max-pooling layer, a dense layer, and a two-layer MltR-KAN, optimized using SGD.

D-index. To assess performance, we use accuracy and the D-index (Diagnostic Index) proposed by (Han et al., 2023). While accuracy can be biased in imbalanced data scenarios, the D-index effectively detects subtle performance differences and accounts for data imbalances. As an interpretable measure ranging within $(0, 2]$, the D-index measures performance by calculating the expected value of local index values across all classes:$d = \frac{1}{K}\sum_{i=1}^{K}\left(\log_2(1+\alpha_i)+\log_2\left(1+\frac{s_i+p_i}{2}\right)\right)$, where α_i is accuracy, s_i is sensitivity, and p_i is specificity for class i among K classes and $i \in K$. A higher D-index indicates better learning performance.

Superiority of dcMltR-KAN: Figure 3 compares dcMltR-KAN (with Haar wavelets) against the baseline models on the O1, O2, and O3 datasets. dcMltR-KAN consistently outperforms all the other models in terms of accuracy and D-index. For O1 (41,717 samples), dcMltR-KAN achieved an accuracy of 0.9817 ± 0.0017 and a D-index of 1.9936 ± 0.0009, surpassing the Transformer's accuracy of 0.9402 and D-index of 1.9110. This trend continues for larger datasets, such as O3 (500,524 samples), where dcMltR-KAN achieved an accuracy of 0.9009 and a D-index of 1.9377, outperforming the Transformer's accuracy of 0.8418 and D-index of 1.8389. These results demonstrate that dcMltR-KAN provides superior performance, robust generalization, and effectiveness in gravitational wave glitch detection, even compared to fully-supervised baselines.

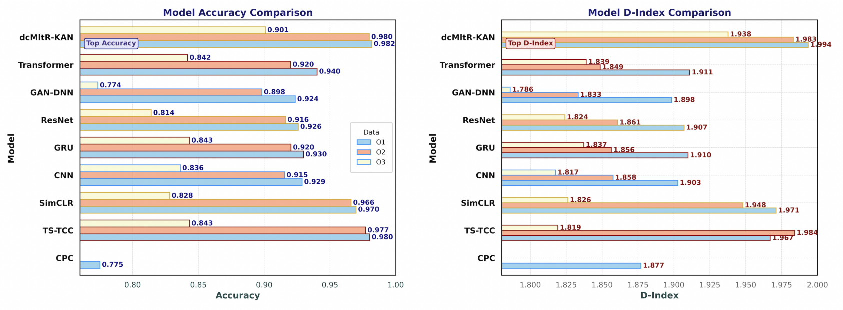

Figure 3: Comparisons of dcMltR-KAN with peer methods on O1, O2, and O3 datatsets

Although CPC showed poor performance, SimCLR and TS-TCC performed slightly worse than dcMltR-KAN for the O1 and O2 datasets in terms of accuracy and d-index. However, their performance significantly lagged behind dcMltR-KAN for the larger O3 data with 500,524 samples. This disparity underscores the superior scalability of dcMltR-KAN, which maintains robust performance even with large-scale data, thanks to its feature weighting (SBW), hierarchical learning, False Negative Elimination (FNE), and multi-resolution KAN mechanisms.

Baseline overfitting on data imbalance. We find that the CNN shows higher accuracy but a lower D-Index than ResNet on both the O1 and O3 datasets, suggesting it may overfit to majority classes while underperforming on minority ones. Despite CNN's higher overall accuracy (0.9288 on O1 and 0.8363 on O3), its lower D-Index reflects weaker performance on minority classes compared to ResNet (accuracy: 0.9259 on O1, 0.8141 on O3; D-Index: 1.9071 and 1.8240. In other words, while the CNN has high accuracy, it is biased toward the majority classes, meaning it overfits to the dominant patterns without effectively learning from the minority instances. Similar trends are observed for the Transformer on O2 compared to ResNet, as well as for GRU compared to the Transformer on O3. Similar trends are observed for the Transformer on O2 compared to ResNet, as well as for GRU compared to the Transformer on O3. Additionally, TS-TCC achieves 97.7% accuracy on O1 data, slightly lower than its 98.0% accuracy on O2 data. However, its D-Index on O1 is 1.984, which is higher than its D-Index on O2 (1.967), suggesting overfitting to the majority groups.

Table 1: Ablation study of dcMltR-KAN on O1, O2, and O3 data.

Dataset	Components	Accuracy (mean ± std)	D-Index (mean ± std)
O1	*w/o wDCL*	0.9219 ± 0.0014	1.9187 ± 0.0015
	w/o mltR-KAN	0.9254 ± 0.0069	1.9174 ± 0.0025
O2	*w/o wDCL*	0.8887 ± 0.0015	1.8154 ± 0.0014
	w/o mltR-KAN	0.8850 ± 0.0059	1.9272 ± 0.0035
O3	*w/o wDCL*	0.8888 ± 0.0008	1.9293 ± 0.0004
	w/o mltR-KAN	0.8639 ± 0.0018	1.8830 ± 0.0012

Abalation studies. dcMltR-KAN consists of wDCL, a CNN-based encoder, and mltR-KAN. Since the CNN-based encoder serves as the backbone of this SSL model, we focus our ablation study on evaluating the contributions of wDCL and mltR-KAN individually. Table 1 highlights the essential roles each component plays in enhancing the model's performance across three datasets: O1, O2, and O3, where mltR-KAN with Harr. The results reveal that removing wDCL leads to a noticeable decrease in the D-Index across all datasets. Specifically, the D-Index drops to 1.9187, 1.8154, and 1.9293 on O1, O2, and O3 respectively, down from the original values of 1.9936, 1.9832, and 1.9377. Similarly, excluding mltR-KAN also results in diminished performance: on O1, the D-Index decreases to 1.9174; on O2, it decreases to 1.9272; and on O3, it decreases to 1.8830. These findings underscore the essential roles of both wDCL and mltR-KAN in maintaining and enhancing the model's performance across all evaluated datasets. The similar results can be found on other wavelets (Appendix).

Impact of dcMltR-KAN on Data Representation We employ UMAP to visualize gravitional wave data before and after dcMltR-KAN to examine this SSL model's impacts on data representation. Figure 4 presents UMAP visualizations of O1, O2, and O3 data, showcasing dcMltR-KAN's effectiveness in enhancing feature separability and uncovering latent structures within large-scale gravitational wave data. This is further supported by silhouette analysis, which consistently validates these findings by showing a significant increase in silhouette scores for data representation after applying dcMltR-KAN (Appendix). We use UMAP instead of t-SNE because it preserves global and local structures, handles large datasets efficiently, and produces stable embeddings.

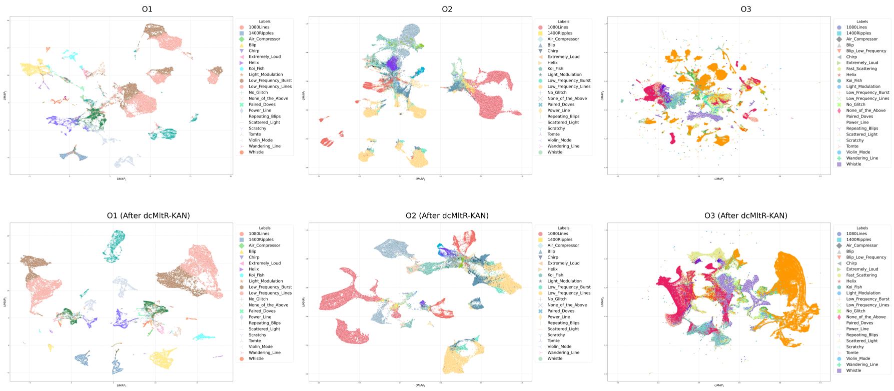

Figure 4: UMAP visualization before and after dcMltR-KAN on O1, O2 and O3 data

Extending dcMltR-KAN to other time-series data. We further extend dcMltR-KAN to other time-series data, demonstrating its applicability to audio tasks. For this, we use the benchmark EMODB dataset, a widely used resource for speech emotion recognition. The dataset contains 535 audio samples across seven imbalanced emotion categories: Anger (127), Boredom (81), Disgust (46), Fear (69), Happiness (71), Sadness (62), and Neutral (79). After preprocessing and feature extraction, 54 features are retained for analysis (see Appendix). dcMltR-KAN demonstrates its superiority on this dataset. Table S5 in Appendix shows the Top-1 results of dcMltR-KAN on the EMODB dataset and in the ablation study. Our model achieved 93.26% accuracy with the Mexican-hat wavelet and 88.86% accuracy with the Haar wavelet. These results outperform almost all previous fully supervised and SSL models. The ablation study further highlights the contribution of key components, showing a performance drop when wDCL or mltR-KAN is excluded. For SOTA comparison, Baek & Lee (2023) reported 90.4% weighted accuracy (WA) and 91.3% unweighted accuracy (UA) with their CNN-BiLSTM model, while Wang et al. (2023) reported 86.31% using Fairtune with a self-supervised wav2vec 2.0 model.

5 DISCUSSION AND CONCLUSION

While dcMltR-KAN shows strong performance in glitch detection and extends effectively to audio data, it has some weaknesses. 1) The high computational complexity of dcMltR-KAN ($\mathcal{O}(N^2+m^2)$) for large datasets like O3 ($N = 5 \times 10^6$, $m = 100$) arises from similarity matrix calculations in FNE and SBW, where N and m are the number of observations and training batch size during training. This challenge can be mitigated through GPU acceleration, sparsification (Liu & Liu, 2019), mini-batching (Recht et al., 2011), subsampling (Coates et al., 2011), and efficient computation libraries like cuML (Rapp et al., 2021). 2) Dynamically adjusting α to balance easy and hard negatives is challenging and may not generalize across datasets. Fine-tuning λ and β also remains complex and requires further exploration. 3) While dcMltR-KAN excels on gravitational wave and speech emotion datasets, its performance on diverse data types (e.g., audio, image) needs further evaluation.

We plan to enhance dcMltR-KAN by replacing the existing CNN-based encoder with MltR-KAN, aiming to achieve greater complexity advantages through optimized FNE and SBW directly under the MltR-KAN encoder. This will address current weaknesses and extend its applications to more data domains besides extend our model to handle raw gravitational wave data. Additionally, we intend to conduct further theoretical investigations to extend MltR-KAN in other SSL-related topics

As the first model to introduce multi-resolution KAN into SSL, dcMltR-KAN brings novelty, efficiency, and explainability to gravitational wave glitch detection and SSL, while inspiring future research in AI and astrophysics.

REFERENCES

B. P. Abbott 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(3): 031040, 2019. doi: 10.1103/PhysRevX.9.031040.

R. Abbott 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. doi: 10.1103/PhysRevX.11.021053.

R. Abbott et al. GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo during the Second Part of the Third Observing Run. *Phys. Rev. X*, 13(4):041039, 2023. doi: 10.1103/PhysRevX.13.041039.

Sofia Alvarez-Lopez, Annudesh Liyanage, Julian Ding, Raymond Ng, and Jess McIver. Gspynettree: A signal-vs-glitch classifier for gravitational-wave event candidates, 2023.

Ji-Young Baek and Seok-Pil Lee. Enhanced speech emotion recognition using dcgan-based data augmentation. *Electronics*, 12(18):3966, 2023. doi: 10.3390/electronics12183966. URL `https://doi.org/10.3390/electronics12183966`. (This article belongs to the Special Issue Theories and Technologies of Network, Data and Information Security).

S. Bahaadini, V. Noroozi, N. Rohani, S. Coughlin, M. Zevin, J. R. Smith, V. Kalogera, and A. Katsaggelos. Machine learning for gravity spy: Glitch classification and dataset. *Information Sciences*, 444:172–186, 2018. doi: 10.1016/j.ins.2018.02.068. URL `https://doi.org/10.1016/j.ins.2018.02.068`.

M. Bailes, B. K. Berger, P. R. Brady, M. Branchesi, K. Danzmann, M. Evans, K. Holley-Bockelmann, B. R. Iyer, T. Kajita, S. Katsanevas, M. Kramer, A. Lazzarini, L. Lehner, G. Losurdo, H. Luck, D. E. McClelland, M. A. McLaughlin, M. Punturo, S. Ransom, S. Raychaudhury, D. H. Reitze, F. Ricci, S. Rowan, Y. Saito, G. H. Sanders, B. S. Sathyaprakash, B. F. Schutz, A. Sesana, H. Shinkai, X. Siemens, D. H. Shoemaker, J. Thorpe, J. F. J. van den Brand, and S. Vitale. Gravitational-wave physics and astronomy in the 2020s and 2030s. *Nature reviews physics*, 3(5):23, 2021. ISSN 2522-5820. doi: 10.1038/s42254-021-00303-8.

Yan-Chen Bi, Yu-Mei Wu, Zu-Cheng Chen, and Qing-Guo Huang. Constraints on the velocity of gravitational waves from the nanograv 15-year data set. *Physical Review D*, 109(6):L061101, 2024.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*. PMLR, 2020.

S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pp. 539–546, 2005. doi: 10.1109/CVPR.2005.202.

Mohammad Abu Thaher Chowdhury. Advancements in Glitch Subtraction Systems for Enhancing Gravitational Wave Data Analysis: A Brief Review. 6 2024.

Alvin JK Chua, Chad R Galley, and Michele Vallisneri. Reduced-order modeling with artificial neurons for gravitational-wave inference. *Physical review letters*, 122(21):211101, 2019.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 8765–8775. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/63c3ddcc7b23daa1e42dc41f9a44a873-Paper.pdf`.

Adam Coates, Andrew Y Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pp. 215–223. JMLR, 2011. URL `https://proceedings.mlr.press/`.

Elena Cuoco, Jade Powell, Marco Cavaglià, Kendall Ackley, Michał Bejger, Chayan Chatterjee, Michael Coughlin, Scott Coughlin, Paul Easter, Reed Essick, et al. Enhancing gravitational-wave science with machine learning. *Machine Learning: Science and Technology*, 2(1):011002, 2020.

Tom Dooney, Stefano Bromuri, and Lyana Curier. Dvgan: Stabilize wasserstein gan training for time-domain gravitational wave physics, 2022.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI-21)*, 2021.

Tiago Fernandes, Samuel Vieira, Antonio Onofre, Juan Calderón Bustillo, Alejandro Torres-Forné, and José A Font. Convolutional neural networks for the classification of glitches in gravitational-wave data streams. *Classical and Quantum Gravity*, 40(19):195018, September 2023. ISSN 1361-6382. doi: 10.1088/1361-6382/acf26c. URL `http://dx.doi.org/10.1088/1361-6382/acf26c`.

Daniel George and Eliu Antonio Huerta. Deep learning for real-time gravitational wave detection and parameter estimation: Results with advanced ligo data. *Physics Letters B*, 778:64–70, 2018.

Jane Glanzer, Sharan Banagari, Scott Coughlin, Michael Zevin, Sara Bahaadini, Neda Rohani, Sara Allen, Christopher Berry, Kevin Crowston, Mabi Harandi, Corey Jackson, Vicky Kalogera, Aggelos Katsaggelos, Vahid Noroozi, Carsten Osterlund, Oli Patane, Joshua Smith, Siddharth Soni, and Laura Trouille. Gravity spy machine learning classifications of ligo glitches from observing runs o1, o2, o3a and o3b, 2021. URL `https://doi.org/10.5281/zenodo.5649212`.

Henry Han, Yi Wu, Jiacun Wang, and Ashley Han. Interpretable machine learning assessment. *Neurocomputing*, 561:126891, 2023. doi: 10.1016/j.neucom.2023.126891. URL `https://www.sciencedirect.com/science/article/pii/S0925231223010147`.

Liang Liu and Heng Liu. Sparse learning and optimization. *Journal of Machine Learning Research*, 20(1):1–45, 2019. URL `https://www.jmlr.org/`.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*, 2024.

M Coleman Miller and Nicolás Yunes. The new frontier of gravitational waves. *Nature*, 568(7753): 469–476, 2019.

Jade Powell, Ling Sun, Katinka Gereb, Paul D Lasky, and Markus Dollmann. Generating transient noise artefacts in gravitational-wave detector data with generative adversarial networks. *Classical and Quantum Gravity*, 40(3):035006, January 2023. ISSN 1361-6382. doi: 10.1088/1361-6382/acb038. URL `http://dx.doi.org/10.1088/1361-6382/acb038`.

Elias Rapp, Ajay Pandey, Scott Zink, et al. cuml: A machine learning library for gpus. *ACM Transactions on Mathematical Software*, 2021. URL `https://developer.nvidia.com/cuml`.

Massimiliano Razzano and Elena Cuoco. Image-based deep learning for classification of noise transients in gravitational wave detectors. *Classical and Quantum Gravity*, 35(9):095016, April 2018. ISSN 1361-6382. doi: 10.1088/1361-6382/aab793. URL `http://dx.doi.org/10.1088/1361-6382/aab793`.

Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems*, volume 24, 2011. URL `https://proceedings.neurips.cc/`.

Travis Robson and Neil J Cornish. Detecting gravitational wave bursts with lisa in the presence of instrumental glitches. *Physical Review D*, 99(2):024019, 2019.

Yusuke Sakai, Yousuke Itoh, Piljong Jung, Keiko Kokeyama, Chihiro Kozakai, Katsuko T. Nakahira, Shoichi Oshino, Yutaka Shikano, Hirotaka Takahashi, Takashi Uchiyama, Gen Ueshima, Tatsuki Washimi, Takahiro Yamamoto, and Takaaki Yokozawa. Unsupervised learning architecture for classifying the transient noise of interferometric gravitational-wave detectors. *Scientific Reports*, 12(1), June 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-13329-4. URL `http://dx.doi.org/10.1038/s41598-022-13329-4`.

Marlin B Schäfer, Ondřej Zelenka, Alexander H Nitz, He Wang, Shichao Wu, Zong-Kuan Guo, Zhoujian Cao, Zhixiang Ren, Paraskevi Nousi, Nikolaos Stergioulas, et al. First machine learning gravitational-wave search mock data challenge. *Physical Review D*, 107(2):023021, 2023.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *arXiv preprint arXiv:2005.10242*, 2022.

Yang Wang, Qibin Liang, Chenghao Xiao, Yizhi Li, Noura Al Moubayed, and Chenghua Lin. Audio contrastive based fine-tuning, 2023. URL `https://arxiv.org/abs/2309.11895`.

Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination, 2018. arXiv:1805.01978 [cs.CV].

Yue Zhang, Weiqi Li, Jinglin Zhang, Liang Chen, Jinguo Gao, Lingfei Wu, and Yu Jiang. Ts2vec: Towards universal representation of time series. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(6):7306–7314, 2022. URL `https://arxiv.org/abs/2110.14782`.