Novelty detection meets collider physics

Jan Hajer,[1,2] Ying-Ying Li,[3,4] Tao Liu,[3] and He Wang[3]

[1]*Institute for Advanced Studies, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong S.A.R, People's Republic of China*

[2]*Centre for Cosmology, Particle Physics and Phenomenology, Université catholique de Louvain, Louvain-la-Neuve B-1348, Belgium*

[3]*Department of Physics, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong S.A.R., People's Republic of China*

[4]*Kavli Institute for Theoretical Physics, University of California Santa Barbara, California 93106-4030, USA*

(Received 11 September 2018; accepted 2 April 2020; published 20 April 2020)

Novelty detection is the machine learning task to recognize data, which belong to an unknown pattern. Complementary to supervised learning, it allows us to analyze data model-independently. We demonstrate the potential role of novelty detection in collider physics, using autoencoder-based deep neural network. Explicitly, we develop a set of density-based novelty evaluators, which are sensitive to the clustering of unknown-pattern testing data or new-physics signal events, for the design of detection algorithms. We also explore the influence of the known-pattern data fluctuations, arising from nonsignal regions, on detection sensitivity. Strategies to address it are proposed. The algorithms are applied to detecting fermionic ditop partner and resonant ditop productions at LHC, and exotic Higgs decays of two specific modes at a future e^+e^- collider. With parton-level analysis, we conclude that potentially the new-physics benchmarks can be recognized with high efficiency.

DOI: 10.1103/PhysRevD.101.076015

I. INTRODUCTION

Since the early developments in the 1950s [1], machine learning (ML) has evolved into a science addressing various *big data* problems. The techniques developed for ML, such as *decision tree learning* [2] and *artificial neural networks* (ANN) [3], allow us to train computers in order to perform specific tasks usually deemed to be complex for handwoven algorithms. For *supervised learning*, the algorithm is first trained on labeled data, and then to classify testing data into the categories defined during training. In contrast, in *semi-supervised* and *unsupervised learning*, where partially labeled or unlabeled data is provided, the algorithm is expected to find the relevant patterns unassisted.

The last decade has seen a rapid progress in ML techniques, in particular the development of deep ANN. A deep ANN is a multilayered network of threshold units [4]. Each unit computes only a simple nonlinear function of its inputs, which allows each layer to represent a certain level of relevant features. Unlike traditional ML techniques (e.g., boosted decision trees) which rely heavily on expert-designed features in order to reduce the dimensionality of the problem, deep ANN automatically extract pertinent features from data, enabling data-mining without prior assumptions. Fueled by vast amounts of big data and the fast development in training techniques and parallel computing architectures, modern deep learning systems have achieved major successes in computer vision [5], speech recognition [6], natural language processing [7], and have recently emerged as a promising tool for scientific research [8–11], where the plethora of experimental data presents a challenge for insightful analysis.

High energy physics (HEP) is a big data science and has a long history of using supervised ML for data analysis. Recently, pioneering works have demonstrated the capability of deep ANN in understanding jet substructure [12–15] and the identification of particles [16] or even whole signal signatures (see e.g., [17], where weakened supervised learning is applied). However, the primary goal of the HEP experiments is to detect predicted or unpredicted physics beyond the Standard Model (BSM) in order to establish the underlying fundamental laws of nature. Despite its significant role in current data analysis, supervised ML techniques suffer from the model dependence introduced during training. This problem can potentially be addressed by the semi-supervised/unsupervised techniques developed for *novelty detection* (for a review see, e.g., [18]). Novelty detection is the ML task to recognize data belonging to an

unknown pattern. If being interpreted as novel signal, BSM physics could be detected without specifying an underlying theory during data analysis. Hence, a combination of novelty detection and supervised ML may lay out a framework for the future HEP data analysis.

Some preliminary and at least partially related efforts have been made at jet [19,20] and event [21–25] level. For novelty detection with given feature representation, its sensitivity depends crucially on the performance of novelty evaluators. Well-designed evaluators will allow to evaluate the data novelty efficiently and precisely. As a matter of fact the design of novelty evaluators or the relevant test statistics defines the frontier of novelty detection [18]. In this paper, we propose a set of density-based novelty evaluators. In contrast to traditional density-based ones, which only quantify isolation of testing data from the known patterns, the new novelty evaluators are sensitive to the clustering of testing data. On this basis, we design algorithms for novelty detection using an autoencoder, which are subsequently applied for detecting several BSM benchmarks at LHC and future e^+e^- colliders.

II. ALGORITHMS

Novelty detection using a deep ANN can be separated into three steps: (1) feature learning, (2) dimensional reduction, (3) novelty evaluation. During the first step the ANN is trained under supervision, using labeled known patterns. The nodes of the trained ANN contain the information gathered for classification and constitute the feature space, which has typically a large dimension. In order to reduce the sparse error and to improve the efficiency of the analysis, one removes the irrelevant features by *dimensional reduction*, which can be implemented using an *autoencoder* [26]. An autoencoder is an ANN with identical number of nodes for input and output layers and fewer nodes for hidden layers. Its loss-function measures the difference between input and output, defined as the reconstruction error $\|x - x'\|^2$. Here x and x' are the vectors of input and output nodes, respectively. Hence the autoencoder learns unsupervised how to reconstruct its input. This allows it to form a submanifold in the full feature space. Afterwards, the novelty of testing data is evaluated, for the final significance analysis. The algorithm is shown in Fig. 1. For the HEP data analysis, the data with known and unknown patterns can be interpreted as SM background and BSM signal, respectively.

We generate Monte Carlo data using MADGRAPH5_aMC@NLO [27] and rely on Keras [28] (TensorFlow [29]-based) for the ANN construction. For the *supervised classification* of events with n visible-particle four-momenta (which we internally normalize by 200 GeV) and l labeled patterns we use an ANN with $4n$ input nodes, l output nodes, and three hidden layers with 30, 30 and 10 nodes, respectively. We use Nesterov's accelerated gradient descent optimizer [30] with a learning rate of 0.3, a learning momentum of 0.99 and a decay rate of 10^{-4}. The batch size is fixed to be 30 and the loss function is the categorical cross entropy [31,32]. The collection of all nodes constitute the feature space with dimension $4n + 30 + 30 + 10 + l$. This ensures that it contains the nonlinear information learned from classification. We normalize the axes of the feature space to $[-1, 1]$ and use tanh as activation function for the autoencoder. Finally, an autoencoder consisting of five hidden layers with 40, 20, 8, 20, and 40 nodes, respectively, and a learning rate of 2.0 projects this feature space onto an eight-dimensional sub-space. We have checked that the results of all ANNs are stable against variations in the numbers of hidden layers and nodes.

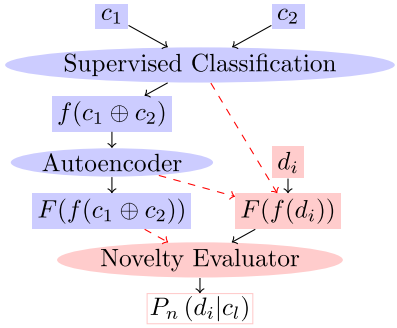


FIG. 1. Novelty detection algorithm. The training and testing phases are marked in blue and red, respectively. Datasets, algorithm and probabilities are indicated by rectangular, elliptic, and plain nodes, respectively. The information gathered during training and used for testing is marked by dashed red arrows. For clarity we have limited the number of labeled known patterns c_l to two. d_i denotes testing data with known and unknown patterns.

III. NOVELTY EVALUATION

Novelty evaluation of testing data is a crucial step for novelty detection. Various approaches have been developed in the past decades [18]. For nontime series data, one of the most popular approaches is density-based [33], in which a local outlier factor (LOF), i.e., the ratio of the local density of a given testing data and the local densities of its neighbors, is proposed as a novelty measure. Explicitly, this traditional measure is [34,35]

$$\Delta_{\text{trad}} = \frac{d_{\text{train}} - \langle d'_{\text{train}} \rangle}{\langle d'^2_{\text{train}} \rangle^{1/2}}, \tag{1}$$

here d_{train} is the mean distance of a testing data to its k nearest neighbors, $\langle d'_{\text{train}} \rangle$ is the average of the mean distances defined for its k nearest neighbors, and $\langle d'^2_{\text{train}} \rangle^{1/2}$ is the standard deviation of the latter. The subscript of "train" indicates that all quantities are defined with respect to the training dataset. We calculate $\langle d'^2_{\text{train}} \rangle^{1/2}$ using the method suggested in [34,35]. The probabilistic novelty evaluator can be defined as the cumulative

distribution function $\mathcal{O}_{\text{trad}} = \frac{1}{2}(1 + \text{erf}\frac{\Delta_{\text{trad}}}{c\sqrt{2}})$. Here c is a normalization factor, defined as the root mean square of the measure values for the testing dataset with known pattern only. This evaluator measures the isolation of testing data from training data. A testing data located away from or at the tail of the training data distribution thus tends to be scored high by $\mathcal{O}_{\text{trad}}$ [33,34].

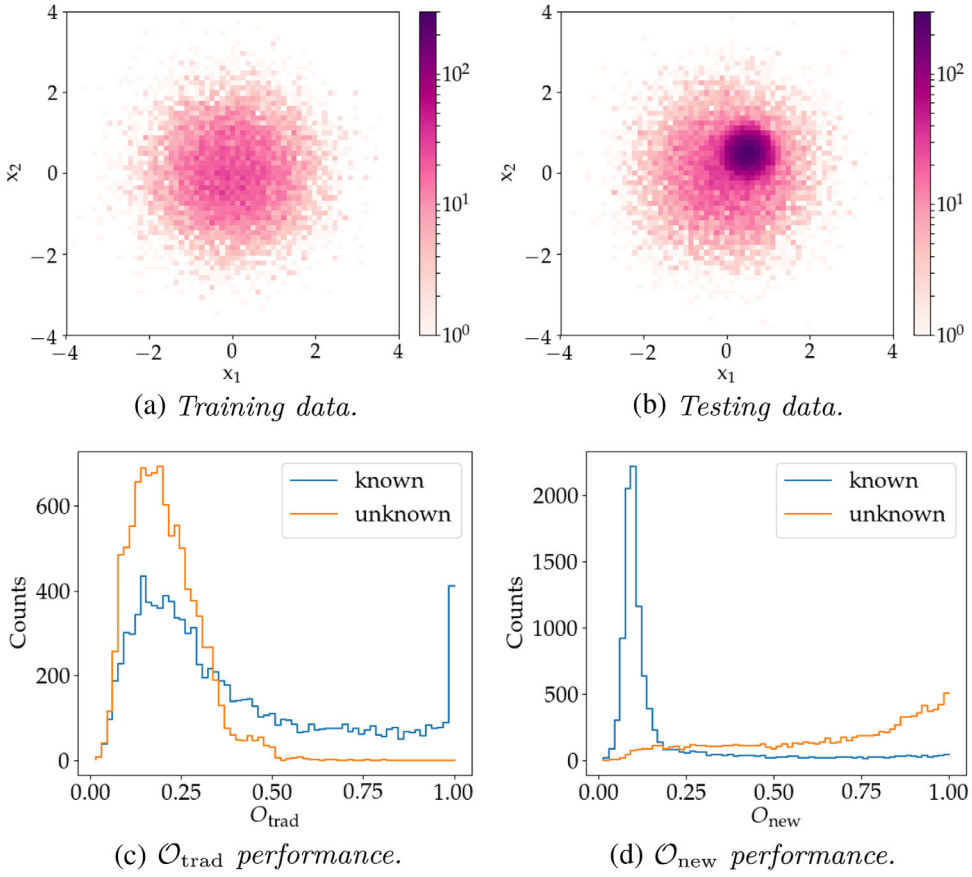
However, $\mathcal{O}_{\text{trad}}$ is blind to the clustering of testing data which generically exists in the BSM datasets and may result in nontrivial structures such as resonance. In order to utilize this feature, we introduce a measure:

$$\Delta_{\text{new}} = \frac{d_{\text{test}}^{-m} - d_{\text{train}}^{-m}}{d_{\text{train}}^{-m/2}}, \tag{2}$$

with m being the dimension of the feature space where the novelty of data is evaluated. In the analysis pursued below, the novelty of data will be evaluated in the eight dimensional latent space of the autoencoder, and hence $m = 8$. d_{test} is the mean distance of the testing data to its k nearest neighbors in the testing dataset, whereas d_{train} is the SM prediction of the same, which can be approximately calculated using the training dataset. To make this measure meaningful in the case where the training and testing datasets are different in size, we require k to scale with the size or the corresponding luminosity of the datasets, respectively. This measure is reminiscent of the test statistic introduced in [36,37], where similar idea is employed for estimating the divergence of data distribution. As Δ_{new} is approximately $\propto \frac{S}{\sqrt{B}}$, with S and B being the numbers of signal and background events in a local bin with unit volume, this measure can be interpreted as the significance of discovery (up to a calibration constant) for this local bin.

$\mathcal{O}_{\text{new}}$ is defined in a similar way as $\mathcal{O}_{\text{trad}}$ does. To compare the performance of $\mathcal{O}_{\text{trad}}$ and $\mathcal{O}_{\text{new}}$ in probing the clustering, we introduce a toy model, where the data resides in a two-dimensional space. The known pattern is a Gaussian distribution centered around the origin $\mathcal{N}(\vec{0}, \mathbf{I})$, while the unknown pattern is an overlapping narrow Gaussian distribution shifted away from the origin $\mathcal{N}((0.5, 0.5)^{\mathsf{T}}, 0.1\mathbf{I})$. The training dataset consists of 10^4 events with known pattern [cf. Fig. 2(a)], while the testing dataset contains from each, known and unknown pattern, 10^4 events [cf. Fig. 2(b)]. As shown in Fig. 2(c) and Fig. 2(d), the clustering of the unknown-pattern data, although being hidden from $\mathcal{O}_{\text{trad}}$, is picked-up by $\mathcal{O}_{\text{new}}$.

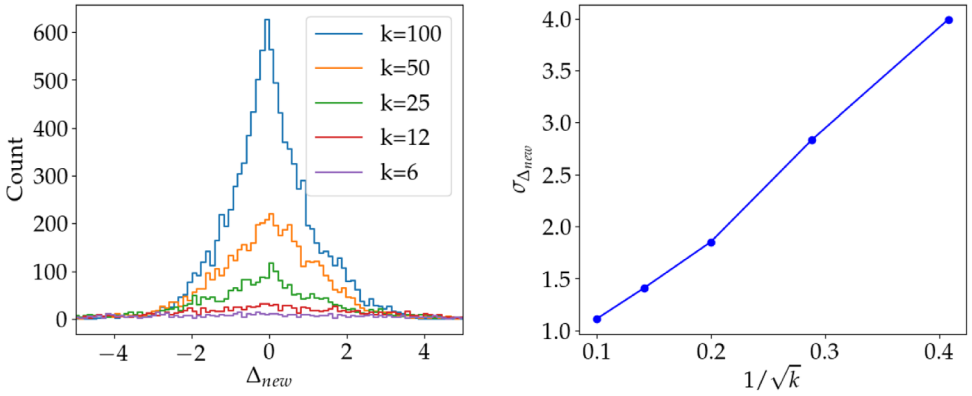
The detection based on Δ_{new} (or $\mathcal{O}_{\text{new}}$) however may suffer from fluctuations of the known-pattern testing data in



(a) *Training data.* (b) *Testing data.*

(c) $\mathcal{O}_{\text{trad}}$ *performance.* (d) $\mathcal{O}_{\text{new}}$ *performance.*

FIG. 2. Comparison between traditional and new novelty evaluators. The toy-data is shown in panels (a) and (b), while the novelty response is given in (c) and (d).

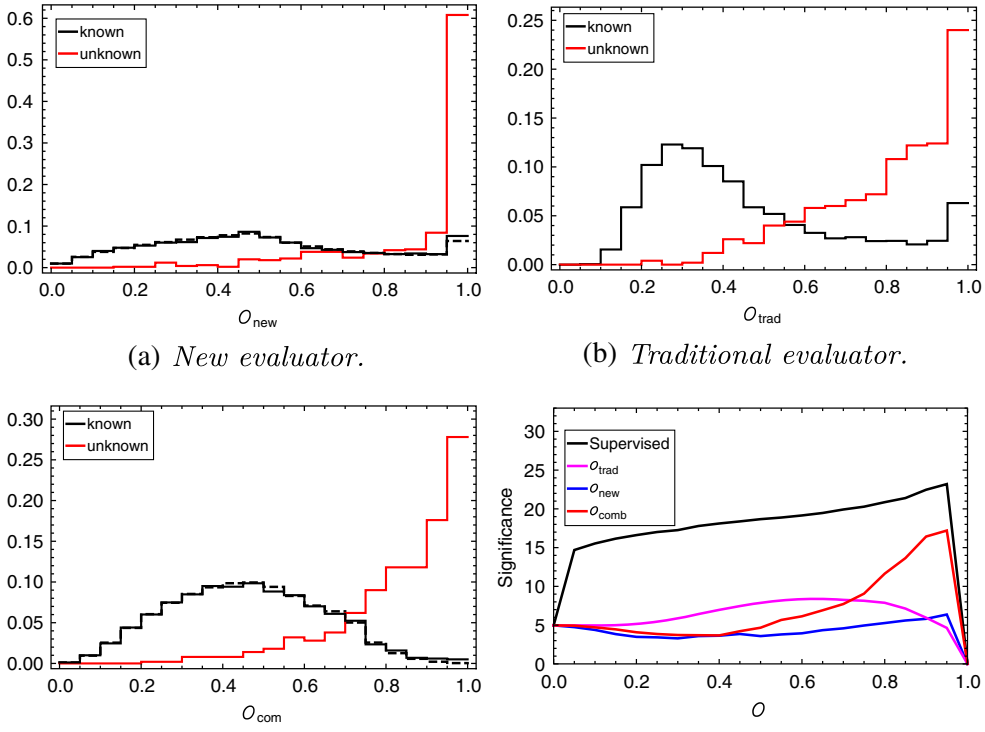


(a) Δ_{new} *response.* (b) *CLT.*

FIG. 3. Dependence of the Δ_{new} response on k, for the testing data with known patterns only. While the training dataset is composed of 50000 points, the testing dataset consists of 10000, 5000, 2500, 1250 and 625 points, with k scaling linearly as 100, 50, 25, 12 and 6, respectively. Both datasets are Gaussian. Panel (a) shows the Δ_{new} response in all cases. Panel (b) shows that its standard deviation $\sigma_{\Delta_{\text{new}}}$ scales linearly with $1/\sqrt{k}$ or $1/\sqrt{L}$, as predicted by the CLT.

the nonsignal regions, via the $1/d^m_{\text{test}}$ term in Eq. (2). While Δ_{new} is expected to be zero if the data only consists of events with known patterns, the fluctuations result in nonzero values, since the measure picks up local data excess. This in essence is a kind of look elsewhere effect (LEE). The fluctuations in $1/d^m_{\text{train}}$ on the other hand can be neglected, as long as the training dataset used for calculating $1/d^m_{\text{train}}$ is much larger than the testing one, with k being properly scaled.

The influence of fluctuations on detection sensitivity can be compensated for as the luminosity L increases, if k scales with L. In this case more and more data are used to calculate $1/d^m_{\text{test}}$ in the local bin which is barely changed. This compensation is approximately predicted by the



(a) *New evaluator.* (b) *Traditional evaluator.*

(c) *Combined evaluator.* (d) *Significance.*

FIG. 4. Normalized data responses to the novelty evaluators $\mathcal{O}_{\text{new}}$ (a), $\mathcal{O}_{\text{trad}}$ (b), and $\mathcal{O}_{\text{comb}}$ (c), and significance performance of these evaluators (d). The dashed black curves in (a) and (c) represent the response of the testing data with known-pattern only. They serve as the reference distribution of novelty response for the significance calculation.

central limit theorem (CLT), which states in this context that the standard deviation of the Δ_{new} response scales with $1/\sqrt{k}$ or $1/\sqrt{L}$, for the testing data with known patterns only. We show this in Fig 3, using the known-pattern Gaussian datasets defined before. Indeed, as the number of testing data increases, Δ_{new} becomes less and less sensitive to the fluctuations [see Fig. 3(a)].

If the fluctuations are not fully compensated for by luminosity, the known-pattern testing data could still be scored high by Δ_{new}, and hence diminish the detection sensitivity. This is often true if $S_{\text{tot}}/B_{\text{tot}}$ is small, as typically occurs in the analyses at LHC. To address this potential problem, we propose one more evaluator

$$\mathcal{O}_{\text{comb}} = \sqrt{\mathcal{O}_{\text{trad}}\mathcal{O}_{\text{new}}}. \tag{3}$$

This evaluator utilizes the fact that the known-pattern testing data with high $\mathcal{O}_{\text{new}}$ scores pretty often come from the high-density regions in the feature space, whereas such data are typically scored low by $\mathcal{O}_{\text{trad}}$. As indicated in Fig. 4, $\mathcal{O}_{\text{comb}}$ performs very well in a typical case where the known and unknown-pattern data distributions are partially overlapped, and many of the known-pattern data, especially the ones in the central region, are scored high by $\mathcal{O}_{\text{new}}$ due to the fluctuations. The known-pattern datasets used here are the same as before, containing 10^4 events. The unknown pattern is defined as $\mathcal{N}((1.5, 1.5)^{\mathsf{T}}, 0.1\mathbf{I})$, with $S_{\text{tot}}/B_{\text{tot}} = 1/20$. Indeed, many high-scoring data of known pattern in Fig. 4(a) are pushed to the low-scoring end in Fig. 4(c), due to the compensation of $\mathcal{O}_{\text{trad}}$. This effect results in ~100% improvement, compared to the sensitivities based on $\mathcal{O}_{\text{trad}}$ or $\mathcal{O}_{\text{new}}$ only, as is shown in Fig. 4(d). As a reference, a significance based on ANN with supervised learning is also presented in the same panel. Here (and similarly below) the significance is calculated against the known-pattern only hypothesis for the testing data. Only the events are counted which pass a threshold with respect to the novelty response to the evaluator $\mathcal{O}$. At $\mathcal{O} = 0$ the significance is calculated with no cut being applied. In the $\mathcal{O}_{\text{trad}}$-based analysis, a Poisson-probability-based test statistic [38] has been applied. In the $\mathcal{O}_{\text{new}}$- and $\mathcal{O}_{\text{comb}}$-based analyses, the independence of events is lost to some extent when their novelty response is evaluated. So, we generated 1000 known-pattern datasets, calculated the standard deviation of the 1000 event numbers in the signal bin (based on the novelty responses of the 1000 datasets to $\mathcal{O}_{\text{new}}$ and $\mathcal{O}_{\text{comb}}$, respectively), and applied this information to approximately calculating the significance of event excess (in each benchmark analysis below, 100 background datasets are generated for this purpose.).

IV. STUDY ON BENCHMARK SCENARIOS

In order to illustrate their performance, we apply the algorithms designed above to two parton-level analyses, with two BSM benchmarks defined for each. Though being unrealistic, it is sufficient for proof of concept.

In the first analysis, we simulate the final state $\bar{b}bl^+l^-E_T^{\text{miss}}$ at the 14 TeV LHC, with a luminosity of 3 ab^{-1}. We require exactly two bottom quarks with $p_T >$ 20 GeV and two charged leptons ($e^{\pm}$ and $\mu^{\pm}$) with $p_T > 10$ GeV. The SM backgrounds (i.e., the labeled known patterns which are used for training the supervised classifier shown in Fig. 1) mainly include

(i) $pp \to \bar{t}_l t_l$, $\sigma = 11.5$ fb,
(ii) $pp \to t_l \bar{b} W_l^{\pm}$, $\sigma = 0.365$ fb,
(iii) $pp \to Z_b Z_l$, $\sigma = 0.0765$ fb,

Here the physical cross sections have been universally suppressed by a factor 2000 for effectively testing the applicability of novelty evaluators. The signal could arise from multiple BSM scenarios in this analysis. Here we consider:

X_1 $pp \to \bar{T}T \to W_l^+ W_{\bar{l}}^- \bar{b}b$ where $\bar{T}$ and T are fermionic top partners,
X_2 $pp \to Z' \to \bar{t}t$ where Z' is a new gauge boson.

For simulating the detection sensitivities, we generate the training and testing datasets with a luminosity of 15 ab^{-1} and 3 ab^{-1}, respectively.

In the second analysis, we simulate unpolarized $e^+e^- \to Zh$ production with the final state $\bar{b}bl^+l^-E_T^{\text{miss}}$ at $\sqrt{s} = 240$ GeV, with a luminosity of 5 ab^{-1}. We require exactly two bottom quarks with $p_T > 10$ GeV and two charged leptons ($e^{\pm}$ and $\mu^{\pm}$) with $p_T > 5$ GeV. The SM background arises mainly from

(i) $e^+e^- \to hZ \to Z_{\text{inv}}^* Z_{\bar{b}b} l^+l^-$, $\sigma = 0.00686$ fb,
(ii) $e^+e^- \to hZ \to Z_{\bar{b}b}^* Z_{\text{inv}} l^+l^-$, $\sigma = 0.00259$ fb,

For BSM scenarios, we consider two specific modes of exotic Higgs decay [39]:

Y_1 $h \to \tilde{\chi}_1\tilde{\chi}_2 \to \tilde{\chi}_1\tilde{\chi}_1 a$. This decay topology can arise from the nearly Peccei-Quinn symmetric limit in the NMSSM [40,41], where $\tilde{\chi}_2$ and $\tilde{\chi}_1$ are bino- and singlinolike neutralinos, respectively, and a is a light CP-odd scalar.
Y_2 $h \to Za$ in the 2HDM and the NMSSM [39].

For simulating the detection sensitivities, we generate the training and testing datasets with a luminosity of 3000 ab^{-1} and 5 ab^{-1}, respectively. The parameter values and cross sections for the four benchmark scenarios are summarized in Table I.

TABLE I. Parameter values and cross sections (after preselection) in the benchmark scenarios of BSM physics.

	Parameter values	$\sigma(fb)$
$X1$	$m_T = m_{\bar{T}}$1.2 TeV, BR$(T \to W_l^+ b) = 50\%$	0.152
$X2$	$m_{Z'} = 3$ TeV, $g_{Z'} = g_Z$, BR$(Z' \to \bar{t}t) = 16.7\%$	1.55
$Y1$	$m_{N_1} = \frac{m_{N_2}}{9} = \frac{m_a}{4} = 10$ GeV, BR$(h \to \bar{b}bE_T^{\text{miss}}) = 1\%$	0.108
$Y2$	$m_a = 25$ GeV, BR$(h \to \bar{b}bE_T^{\text{miss}}) = 1\%$	0.053

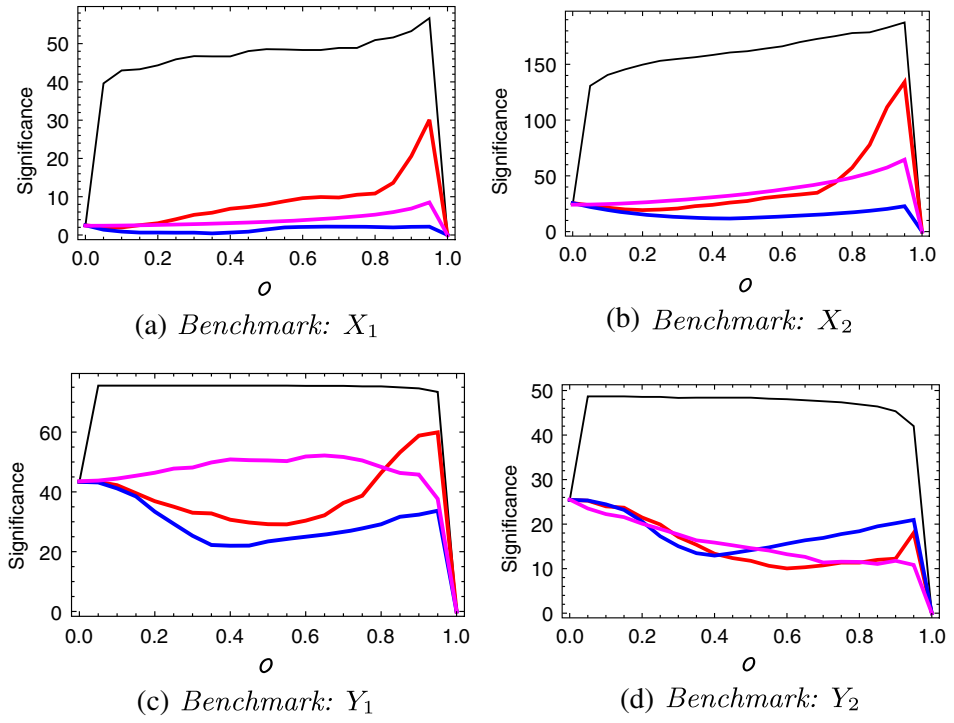


FIG. 5. Significance performance of the novelty-detection evaluators.

The significance performance of the three novelty evaluators is shown in Fig. 5. Again a significance based on ANN with supervised learning is presented as a reference. By comparing this figure and Fig. 4(d), one can see that the 2D Gaussian toy model discussed above displays very good representativeness. In the first analysis, the BSM signal and the SM background are partially overlapped in the feature space for both benchmarks of X_1 and X_2. Many of the background events in the nonsignal region have a strong response to $\mathcal{O}_{\text{new}}$, due to statistical fluctuations, and hence diminish the detection sensitivity. However, with the $\mathcal{O}_{\text{trad}}$ compensation, significant improvement in sensitivity is achieved. As is shown in Fig. 5(a) and Fig. 5(b), the sensitivities are at least doubled using $\mathcal{O}_{\text{comb}}$, compared to the ones using $\mathcal{O}_{\text{new}}$ or $\mathcal{O}_{\text{trad}}$ only. In the second analysis, the fluctuation effect on $\mathcal{O}_{\text{new}}$ is relatively small, due to $S_{\text{tot}}/B_{\text{tot}} > 1$ (typical for the analyses at e^+e^- collider). This limits the performance of $\mathcal{O}_{\text{comb}}$ to a large extent, resulting in a significance with 10–20% improvement in Y_1 and with null improvement in Y_2, compared to the best sensitivities which can be achieved using $\mathcal{O}_{\text{trad}}$ and $\mathcal{O}_{\text{new}}$.

The analyses of these benchmark scenarios, together with the study on the 2D Gaussian toy model, show that both $\mathcal{O}_{\text{new}}$ and $\mathcal{O}_{\text{comb}}$ can effectively pick up the clustering of signal events in the feature space, even if the signal and background events overlap with each other. A large LEE will weaken the efficiency of $\mathcal{O}_{\text{new}}$, but this effect can be significantly suppressed using $\mathcal{O}_{\text{comb}}$. In the case with relatively small LEE (often characterized by relatively big $S_{\text{tot}}/B_{\text{tot}}$), the performance of $\mathcal{O}_{\text{comb}}$ tends to be weakened. A better sensitivity could be achieved either by $\mathcal{O}_{\text{trad}}$ or $\mathcal{O}_{\text{new}}$, or by pursuing event counting with no cut being applied (i.e., at $\mathcal{O} = 0$).

V. SUMMARY AND DISCUSSION

In this paper, we proposed a set of density-based novelty evaluators, $\mathcal{O}_{\text{new}}$ and $\mathcal{O}_{\text{comb}}$, which are sensitive to the clustering of the unknown-pattern testing data, for novelty detection in the HEP data analysis. These evaluators allow to design the algorithms with broad applications in detecting BSM physics. They can be also applied to measuring the SM processes yet to be discovered, if we interpret them as "novel" events. As these algorithms are designed using only general assumptions their application could be extended to other big-data domains as well.

This study could be generalized in multiple directions. We have focused on developing the algorithms for novelty detection in HEP, using parton-level analysis to demonstrate their sensitivity performance. To fill up the gap between the concept and its application to real data analysis, hadron-level analysis is definitely needed. In addition, the algorithms could be improved in several aspects. First, the feature selection in the ANN training process might be not yet fully optimized. The features learned from classification of data with labeled known patterns are likely to be sub-optimal for enhancing the isolation or clustering of the unknown-pattern data. Nevertheless, we may introduce dynamical ML or some

feedback mechanisms using the testing dataset, to reinforce the learning of the unknown-pattern features. Second, the distance definition of data depends on the geometry of the feature space. We adopted the Euclidean geometry for simplicity, but it is worthwhile to explore the other possibilities. Third, the amount of memory and time needed to implement O_{trad} increases rapidly with the data size and dimension, which renders O_{trad} not very efficient for large dataset. Ways of accelerating the calculation might be needed. More than that, we would extend the performance analysis of the algorithms to other BSM scenarios, e.g., the ones with interference between the known and unknown patterns, or nontrivial data clusters such as a dip [42]. Although it is beyond the scope of this study, at last we mention that, a full analysis of the systematic and theoretical uncertainties is absent (for recent effort partially addressing this see [43]). We leave these topics to a future study.

ACKNOWLEDGMENTS

We would like to thank the anonymous referee greatly for his/her constructive suggestions for improving the benchmark analyses, with the novelty evaluators proposed in this paper. We would greatly thank Prof. Michael Wong, our colleague at the HKUST, for highly valuable discussions on these novelty evaluators and the relevant ANN algorithms. T. Liu would thank Huai-Ke Guo for discussions on CLT in this context during the MITP (the Mainz Institute for Theoretical Physics) workshop "Probing Baryogenesis via LHC and Gravitational Wave Signatures". We would thank the experimental colleagues Aurelio Juste, Kirill Prokofiev and Junjie Zhu for reading the manuscript and raising valuable comments. We would also thank Lian-Tao Wang and Zhen Liu for general discussions on this idea at an early stage. J. Hajer is partly supported by the General Research Fund (GRF) under Grant No. 16304315. Y. Y. Li would thank the Kavli Institute for Theoretical Physics, where most of her work was done, for the award of the graduate fellowship which was provided by Simons Foundation under Grant No. 216179 and Gordon and Betty Moore Foundation under Grant No. 4310. This research was also supported in part by the National Science Foundation under Grant No. PHY-1748958. T. Liu is jointly supported by the GRF under Grants No. 16312716 and No. 16302117. The GRF is issued by the Research Grants Council of Hong Kong S. A. R. He would also thank the MITP for its hospitality, where part of his work was done.

Note added.—While this paper was being finalized, [44] appeared. Both the novelty evaluators proposed here and the test statistic defined in [44] (as well as the one developed in [25] recently) are able to measure the clustering of testing data with unknown pattern. We would like to stress that we developed this project and the relevant ideas independently. Especially, two significant differences exist between them. First, unlike the test statistic in [25,44] which measures the divergence of the testing dataset from the training dataset, the evaluators proposed in this work quantify the novelty of individual testing data. Such a design enables the evaluators to probe the fine/differential structure of the clustering such as peak-dip (a famous BSM example can be found in [42]) efficiently. Second, as the LEE could be a severe problem for novelty detection at Hadron colliders, we explored how to diminish its influences on detection sensitivity (in relation to this, $\mathcal{O}_{\text{comb}}$ was designed). This was not developed in [25,44].

[1] A. L. Samuel, IBM J. Res. Dev. **3**, 210 (1959).
[2] J. R. Quinlan, Mach. Learn. **1**, 81 (1986).
[3] C. Peterson, T. Rögnvaldsson, and L. Lönnblad, Comput. Phys. Commun. **81**, 185 (1994).
[4] Y. LeCun, Y. Bengio, and G. Hinton, Nature (London) **521**, 436 (2015).
[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, in *Advances in Neural Information Processing Systems 25*, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Curran Associates, Inc., 2012), pp. 1097–1105.
[6] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, IEEE Signal Process. Mag. **29**, 82 (2012).
[7] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, in *Advances in Neural Information Processing Systems 26*, edited by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Curran Associates, Inc., 2013), pp. 3111–3119.
[8] L. Zdeborová, Nat. Phys. **13**, 420 (2017).
[9] G. Carleo and M. Troyer, Science **355**, 602 (2017).
[10] J. Carrasquilla and R. G. Melko, Nat. Phys. **13**, 431 (2017).
[11] P. Zhang, H. Shen, and H. Zhai, Phys. Rev. Lett. **120**, 066401 (2018).
[12] P. Baldi, K. Bauer, C. Eng, P. Sadowski, and D. Whiteson, Phys. Rev. D **93**, 094034 (2016).
[13] A. Butter, G. Kasieczka, T. Plehn, and M. Russell, SciPost Phys. **5**, 028 (2018).
[14] A. J. Larkoski, I. Moult, and B. Nachman, Phys. Rep. **841**, 1 (2020).
[15] S. Macaluso and D. Shih, J. High Energy Phys. 10 (2018) 121.

[16] P. Baldi, P. Sadowski, and D. Whiteson, Nat. Commun. **5**, 4308 (2014).
[17] T. Cohen, M. Freytsis, and B. Ostdiek, J. High Energy Phys. 02 (2018) 034.
[18] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, Signal Process. **99**, 215 (2014).
[19] E. M. Metodiev, B. Nachman, and J. Thaler, J. High Energy Phys. 10 (2017) 174.
[20] A. Andreassen, I. Feige, C. Frye, and M. D. Schwartz, Eur. Phys. J. C **79**, 102 (2019).
[21] T. Aaltonen *et al.* (CDF Collaboration), Phys. Rev. D **78**, 012002 (2008).
[22] MUSIC—An automated scan for deviations between data and Monte Carlo aimulation (2008).
[23] M. Kuusela, T. Vatanen, E. Malmi, T. Raiko, T. Aaltonen, and Y. Nagai, J. Phys. Conf. Ser. **368**, 012032 (2012).
[24] J. H. Collins, K. Howe, and B. Nachman, Phys. Rev. Lett. **121**, 241803 (2018).
[25] R. T. D'Agnolo and A. Wulzer, Phys. Rev. D **99**, 015014 (2019).
[26] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, in *Proceedings of the 25th International Conference on Machine Learning*, ICML '08 (ACM, New York, 2008), pp. 1096–1103.
[27] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, J. High Energy Phys. 07 (2014) 079.
[28] F. Chollet *et al.*, Keras: The python deep learning library, https://keras.io (2015).
[29] M. Abadi *et al.*, TensorFlow: Large-scale machine learning on heterogeneous systems, https://tensorflow.org (2015).
[30] Y. Nesterov, in *Doklady AN USSR* (1983), Vol. 269, pp. 543–547.
[31] R. Rubinstein, Methodol. Comput. Appl. Probab **1**, 127 (1999).
[32] R. Y. Rubinstein, in *Stochastic Optimization: Algorithms and Applications* (Springer, New York, 2001), pp. 303–363.
[33] M. Breunig, H. Kriegel, R. Ng, and J. Sander, in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, SIGMOD '00 (ACM, New York, 2000), Vol. 29, pp. 93–104.
[34] H. Kriegel, P. Kroger, E. Schubert, and A. Zimek, in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09 (ACM, New York, 2009), pp. 1649–1652.
[35] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, in *Advances in Neural Information Processing Systems 26*, edited by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Curran Associates, Inc., 2013). pp. 935–943.
[36] Q. Wang, S. R. Kulkarni, and S. Verdu, *2006 IEEE International Symposium on Information Theory* (2006), https://doi.org/10.1109/ISIT.2006.261842.
[37] T. Dasu, S. Krishnan, S. Venkatasubramanian, and K. Yi, in *Proceedings of Symposium on the Interface of Statistics, Computing Science, and Applications* (2006).
[38] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, Eur. Phys. J. C **71**, 1554 (2011); **73**, 2501(E) (2013).
[39] D. Curtin *et al.*, Phys. Rev. D **90**, 075004 (2014).
[40] P. Draper, T. Liu, C. E. M. Wagner, L.-T. Wang, and H. Zhang, Phys. Rev. Lett. **106**, 121805 (2011).
[41] J. Huang, T. Liu, L.-T. Wang, and F. Yu, Phys. Rev. Lett. **112**, 221803 (2014).
[42] D. Dicus, A. Stange, and S. Willenbrock, Phys. Lett. B **333**, 126 (1994).
[43] C. Englert, P. Galler, P. Harris, and M. Spannowsky, Eur. Phys. J. C **79**, 4 (2019).
[44] A. De Simone and T. Jacques, Eur. Phys. J. C **79**, 289 (2019).