The Annals of Applied Statistics
2023, Vol. 17, No. 4, 2759–2795
https://doi.org/10.1214/22-AOAS1722

MODEL-INDEPENDENT DETECTION OF NEW PHYSICS SIGNALS USING INTERPRETABLE SEMISUPERVISED CLASSIFIER TESTS

BY PURVASHA CHAKRAVARTI[1,a], MIKAEL KUUSELA[2,b], JING LEI[3,d] AND LARRY WASSERMAN[2,c]

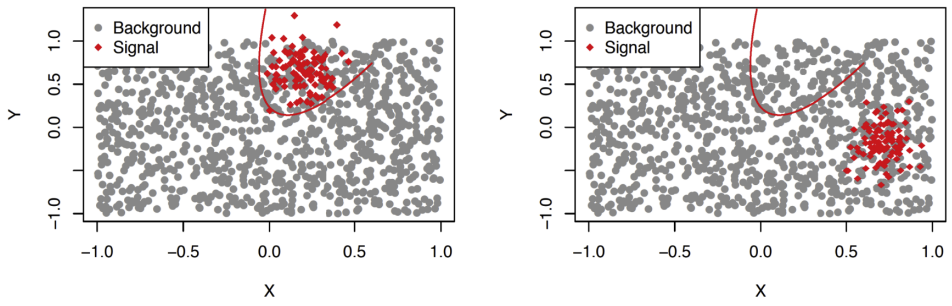
[1]*Department of Statistical Science, University College London,* [a]*p.chakravarti@ucl.ac.uk*
[2]*Department of Statistics & Data Science, and NSF AI Planning Institute for Data-Driven Discovery in Physics, Carnegie Mellon University,* [b]*mkuusela@andrew.cmu.edu,* [c]*larry@stat.cmu.edu*
[3]*Department of Statistics & Data Science, Carnegie Mellon University,* [d]*jinglei@stat.cmu.edu*

A central goal in experimental high energy physics is to detect new physics signals that are not explained by known physics. In this paper we aim to search for new signals that appear as deviations from known Standard Model physics in high-dimensional particle physics data. To do this, we determine whether there is any statistically significant difference between the distribution of Standard Model background samples and the distribution of the experimental observations which are a mixture of the background and a potential new signal. Traditionally, one also assumes access to a sample from a model for the hypothesized signal distribution. Here we instead investigate a model-independent method that does not make any assumptions about the signal and uses a semisupervised classifier to detect the presence of the signal in the experimental data. We construct three test statistics using the classifier: an estimated likelihood ratio test (LRT) statistic, a test based on the area under the ROC curve (AUC), and a test based on the misclassification error (MCE). Additionally, we propose a method for estimating the signal strength parameter and explore active subspace methods to interpret the proposed semisupervised classifier in order to understand the properties of the detected signal. We also propose a score test statistic that can be used in the model-dependent setting. We investigate the performance of the methods on a simulated data set related to the search for the Higgs boson at the Large Hadron Collider at CERN. We demonstrate that the semisupervised tests have power competitive with the classical supervised methods for a well-specified signal but much higher power for an unexpected signal which might be entirely missed by the supervised tests.

1. Introduction. Statistical and machine learning tools have been extensively used over the past few decades to answer fundamental questions in particle physics (Behnke et al. (2013), Bhat (2011)). To answer these questions, one needs to experimentally test the predictions of the Standard Model which describes our current understanding of elementary particles and their interactions. For example, the empirical confirmation of the Higgs boson at CERN in 2012 was an essential step towards its inclusion in the Standard Model (ATLAS Collaboration (2012), CMS Collaboration (2012)).

In this paper we develop statistical tools to address the problem of searching for evidence of new particle physics phenomena in high-dimensional experimental data, which is beyond what is explained by the Standard Model (SM). The goal is to search for a *signal* which represents any anomalous phenomenon that is unexplained by the Standard Model. On the other hand, any event predicted by the Standard Model, including all the previously discovered rich, structured signals predicted by the model, comprise the *background*. For example, for

Received February 2021; revised December 2022.
Key words and phrases. Collective anomaly detection, active subspace, mixture proportion estimation, signal strength estimation, likelihood ratio test, high-dimensional two-sample testing, Large Hadron Collider.



(a) Training Samples (b) Experimental Samples

FIG. 1. *Decision boundary using a supervised classifier to separate the signal* (*red rhombuses*) *from the background* (*grey*): (*a*) *The decision boundary of the classifier when trained on signal generated from the assumed signal model*. (*b*) *When the signal model is misspecified, the classifier completely misses the actual signal data when used on the experimental samples*. *We consider a two-dimensional example here for illustration purposes only. The real data can be of much higher dimensionality. In the experiments considered in this paper, we have a* 16*-dimensional sample*.

more recent searches, background would include Higgs signals which were only recently the focus of attention. The search is performed on the observed unlabelled data, called here the *experimental* data, which are a mixture of the background and a potential new signal.

In experiments conducted with large particle accelerators, such as the Large Hadron Collider (LHC), the searches for new physics signals in the high-dimensional experimental data have been primarily conducted using model-dependent data analysis methods. These searches are generally structured as likelihood ratio tests based on a model assumption for the specific new signal (ATLAS Collaboration and CMS Collaboration (2011), Cowan et al. (2011)). Due to the high dimensionality of the data, tests based on classifiers that are optimized to detect a particular hypothesized signal are preferred over density estimation or mixture model approaches. Specifically, tests based on supervised, multivariate classification algorithms, such as neural networks and boosted decision trees, have proved beneficial. The training samples for the classifier are generated using Monte Carlo (MC) event generators which enable sampling collision events from specific physics models. The classifier output is then used to extract a signal enriched sample which is used to perform likelihood ratio tests for the detection of the signal (ATLAS Collaboration (2012), CMS Collaboration (2012)).

But there are disadvantages to this approach. First, this approach may have trouble finding novel, unexpected deviations from the background model. Second, a search targeting one kind of new physics signal/scenario is not going to be powerful for finding a different new physics signal/scenario. Figure 1 illustrates the problem. If a classifier is trained on training signal data as shown in Figure 1(a), it gives the classification boundary as shown. But what if the signal in the experimental data actually looks like Figure 1(b)? Then the classifier ends up misclassifying the signal as background. So an algorithm trained on a misspecified signal model might completely miss the actual signal. The two-dimensional example considered here is for illustration purposes only. In reality, the data lie in a high-dimensional space which further aggravates the problem.

In this paper we study *model-independent tests*, which do not assume a particular signal model, and compare them to more traditional model-dependent tests for search of new physics signals. We use data from an event generator for background events together with observed experimental data which are a mixture of background and a potential signal. But we do not use data from signal simulations. In other words, we assume access to labeled background data and unlabeled experimental data, where events may either have background or signal labels. (Note that, in practice, systematic uncertainty in the background makes the

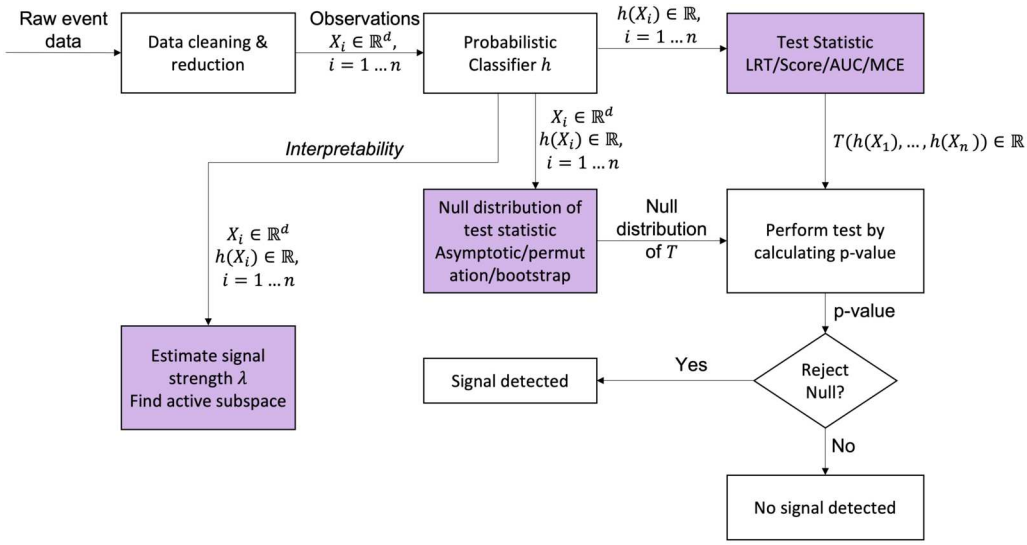


FIG. 2. *Data flow diagram presenting the signal detection approach, including the interpretability of the trained classifier, from raw data to concluding whether there is any signal in the data. The colored boxes represent our contributions presented in this paper. The raw data and data reduction steps are presented, for example, in Adam-Bourdarios et al.* (2015).

situation more complex than this. Please refer to Sections 2 and 7 for further discussion.) We then use a semisupervised approach that trains a classifier to differentiate the background data from the experimental data. Crucially, we do not assume availability of labelled signal data which differentiates this approach from model-dependent or supervised methods.

In particular, we make the following three main contributions:

1. We propose and investigate several variants of classifier-based model-dependent and model-independent tests to detect a new signal in the experimental data. These tests involve four steps:
   (a) *Training a probabilistic classifier* to differentiate between background events and potential new signal events in the model-dependent mode and to differentiate between background events and experimental events in the model-independent mode.
   (b) *Constructing classifier-based test statistics* using the output of the probabilistic classifier. We construct four different test statistics, using two different strategies, and compare them.
   (c) *Estimating the null distributions* of the test statistics using asymptotic, nonparametric bootstrap, and permutation methods, and comparing them.
   (d) *Testing* for the presence of a new signal using the test statistics and their estimated null distributions. We consider each combination of a test statistic and a method for estimating its null distribution as a separate test method.
2. We develop a method for *estimating the signal strength* in the experimental data, which is a challenging problem in model-independent searches, since we do not have information about the signal model to characterize the signal.
3. We develop *active subspace methods* to interpret the signal detected by the semisupervised classifier in the high-dimensional space.

The data flow diagram, presented in Figure 2, illustrates the signal detection approach in four stages: training the probabilistic classifier, computing the test statistics, estimating the null distribution of the test statistics, and finally performing the test. There is also the stage where we use the fitted classifier to estimate the signal strength and interpret the classifier

using active subspace methods. The colored boxes in the diagram represent our contributions in this paper, namely, introducing the different test statistics, introducing methods to estimate the null distribution, estimating the signal strength, and interpretability via active subspace methods. The diagram also demonstrates how the dimensionality of the data reduces from observations (16-D in our case) to the classifier outputs (1-D) which are then used to find the scalar test statistics. Below we detail selected parts of our approach.

Training a probabilistic classifier. The approach we take is one introduced by particle physicists in the model-dependent mode, where we first train a probabilistic classifier (h) and then use its output to construct the test statistics. In the model-dependent (MD) mode, the classifier (h) finds differences between the properties of simulated background data and simulated signal data. The classifier (h) is then applied on the experimental data (W_i's) and the output ($h(W_i)$) is used to construct the test statistics. In the model-independent (MI) mode, we do not make assumptions about the signal model, and hence do not have access to simulated signal data. So, the classifier instead is trained to find differences between the simulated background data and the experimental data, and if it is able to differentiate between them, it indicates the presence of a signal component in the experimental data.

Constructing classifier-based test statistics. We propose two different strategies to construct test statistics from the probabilistic classifier. Using each of the two strategies, we then construct two test statistics. The first strategy, creating *density ratio based test statistics*, takes advantage of the fact that, for a probabilistic classifier, the output h is the posterior binary class probability. Now the posterior probability ($h(w)$) can be written as a one-to-one function of the ratio ($\psi(w)$) of the probability densities of the two classes. Hence, the density ratio $\psi(w)$ can be estimated using the classifier output $h(w)$. Using these density ratio estimates, we create two classical test statistics—the likelihood ratio test statistic (LRT) and the score statistic. Note that the two classes are background and signal in the MD mode and background and experimental in the MI mode. This strategy leverages the fact that classifiers have been found to give accurate estimates of density ratios (Cranmer, Pavez and Louppe (2015)).

The second strategy is to construct *classification performance based test statistics*. This strategy, which is only applicable in the MI setting, takes advantage of the fact that, in the MI mode, in the absence of a signal in the experimental data a classifier should not be able to differentiate the experimental data from the background data since they have the same distribution. This is analogous to a high-dimensional two-sample testing problem where we compare the distributions of the background and the experimental data using a classifier (Friedman (2003), Kim, Lee and Lei (2019), Kim et al. (2021)). In this case the performance of the classifier is indicative of whether there is any signal in the data. We measure the performance in two ways, resulting in two test statistics—the area under the ROC curve (AUC) and the misclassification error (MCE). The reason we include these test statistics is because they are properties of the classifier itself, in contrast to the LRT which is being estimated using the classifier.

The four test statistics according to the two strategies are presented in Table 1. In the model-dependent mode, we cannot use classification performance based test statistics since the classifier is trained on background and signal data. In the model-independent mode, we cannot use the score statistic. This is because the score statistic is a function of the density ratio of the *signal* data and the background data, whereas in the model-independent setting, we only have an estimate for the density ratio of the *experimental* data and the background data.

One of our model-independent tests, based on the LRT statistic, is similar to the test proposed by D'Agnolo and Wulzer (2019) and D'Agnolo et al. (2021). The other two model-independent test statistics are the AUC and MCE statistics. To the best of our knowledge, these two are novel in this application. Though the test that uses the LRT is similar to the test

TABLE 1

Test statistics considered for model-dependent (MD) and model-independent (MI) searches. "X" marks cells that do not have a corresponding test statistic

	Test Statistics			
	Density Ratio Based		Classification Performance Based	
	LRT	Score	AUC	MCE
Model-dependent	MD-LRT	MD-Score	X	X
Model-independent	MI-LRT	X	MI-AUC	MI-MCE

proposed by D'Agnolo and Wulzer (2019) and D'Agnolo et al. (2021), unlike them, we use the nonparametric bootstrap as well as permutation methods to estimate the null distribution, comparing the performance of the different re-sampling techniques. We also apply the tests in a higher-dimensional setting than them. Additionally, since our test has a simple null vs. a composite alternative, the LRT is not guaranteed to yield an optimal test in this setting. Hence, it is important to compare the performance of the different test statistics.

Estimating the null distributions and *Testing.* We estimate the null distributions of the test statistics using four different approaches—asymptotic, nonparametric bootstrap, permutation, and slow permutation. The bootstrap and permutation approaches are resampling methods that utilize the fact that, under the null, the experimental data and the background data have the same distribution, and so the data can be resampled from a combination of the two data sets and their labels can be permuted. The bootstrap and the regular permutation method split the data into training and test data sets and resample the test data sets only, whereas the slow permutation method resamples the entire data set and retrains the classifier on each resampled data set which increases its computational complexity. A combination of a test statistic and a null distribution estimation method forms a test. In principle, every test statistic has a true null distribution, which these methods are trying to estimate. So, under the null, these methods should give similar distributions. However, it is important to note that the estimated null distribution is a function of the experimental data, and hence only when there is no signal in the experimental data (when null is true), it estimates the true null distribution. So the power of the tests, when there is signal, is expected to be different for the different null estimation methods. Hence, it is important to compare the performance of the different null estimation methods for the different test statistics.

Estimating the signal strength. Towards our second contribution of estimating the signal strength, the problem is particularly challenging in the MI mode, since the classifier differentiates between background and experimental data and hence cannot directly identify signal events. We solve this by showing that estimating the signal strength is equivalent to estimating a monotone univariate density over the interval $[0, 1]$ at the right boundary at 1. In particular, we need to estimate the density of a statistic built by considering the quantile of a Neyman–Pearson-style likelihood ratio. Similar to the estimator given by Storey (2002), which is a special case of a boundary density estimator using histograms, we use histograms to estimate the density and then use a Poisson regression (Nelder and Wedderburn (1972)) model of the histogram bin counts to estimate the density at the boundary point. The additional use of Poisson regression makes the boundary estimate more stable. We also find confidence intervals for the estimates using GLM regression intervals and three different bootstrap methods.

Active subspace methods. For our third contribution and to interpret the classifier and characterize the signal, we propose using active subspace methods (Constantine (2015), Constantine, Dow and Wang (2014)) to explore which aspects of the covariates are the most

informative for the test. Active subspaces have been used as a form of sensitivity analysis to quantify uncertainty in an output (Constantine et al. (2015)), and they have also been used to analyze the internal structure and vulnerability of deep neural networks (Cui et al. (2020)). Similar to these works, in this paper we propose active subspace methods to interpret the variability in the classifier output in terms of its inputs. Our proposed methods identify the directions that capture the most variability in the gradient of the classifier. These directions are found using an eigen decomposition (PCA) of the gradients of the classifier surface at observed points. These directions give us information about the combinations of the covariates that most influence the classifier in distinguishing the experimental data from the background data, hence giving us an idea of what the signal looks like. Importantly, this process helps us better understand the classifier that detects the signal which otherwise would be a black box.

Alternative variable importance methods have been used to understand the intrinsic predictiveness potential of the covariates (Lei et al. (2018), van der Laan (2006), Williamson et al. (2019)) which identify the individual covariates that play an important role in the classifier. Variable importance has been addressed extensively for random forests (Breiman (2001), Grömping (2009), Ishwaran (2007), Strobl et al. (2008)) and neural networks (Bach et al. (2015), Shrikumar, Greenside and Kundaje (2017), Sundararajan, Taly and Yan (2017)). But contrary to the active subspace methods, these methods concentrate on finding the variables that are individually the most important and potentially miss detecting combinations of covariates that might have more predictive power.

It is worth noting that the methods presented in this paper can also be applied more generally. In this paper we demonstrate an application of the methods to the search of new phenomena in particle physics, but the methods presented here are applicable beyond particle physics. In general, these methods can be applied whenever there is a need to search for, detect, or characterize collective anomalies in a high-dimensional space, where a single data point is not necessarily anomalous but a collection of data points is. For example, these methods could potentially be used to find anomalous weather changes in high-dimensional climate science data sets, where individual weather events might themselves not be anomalous but a collection of certain types of weather events over a period of time might be. Potential further applications range from engineering to medicine and other areas of the physical sciences.

The rest of this paper is structured as follows. In the following section, we describe the role of model-independent methods in the search for new physics phenomena and discuss some existing literature on it. In Section 3 we introduce the problem setup mathematically. We describe the MD methods in Section 3.1 and the proposed MI methods in Section 3.2. We then discuss methods to estimate the signal strength in the experimental data in Section 4. In Section 5 we describe active subspace methods to understand the subspace affecting the classifier the most, leading to an understanding of the detected signal. Finally, in Section 6 we demonstrate the performance of the proposed MI methods and compare them to the MD methods using publicly available simulated data from a machine learning data challenge that includes events corresponding to the SM Higgs boson signal. We include concluding remarks and possible future directions of research in Section 7. We also provide exploratory analysis of the Higgs boson data set that is used to perform the experiments in Section 6 and include some other experimental details in the Supplementary Material (Chakravarti et al. (2023)). We defer the proofs of the theorems and some of the proposed algorithms to the Supplementary Material as well.

2. Overview of model-independent searches of new physics. In particle physics, before any search analysis is performed, the experimental data is typically filtered for specific types of final-state particles. Different choices here specify different search channels. For example, the search could be restricted to a decay channel where four muons are observed.

The data set would then contain kinematic and other features of these particles observed using the particle detector which dictates the dimensionality of the data. The data set is further restricted by selection cuts on these features that are used to remove data points that lie in regions of the phase space that are believed to be signal-free. For example, one might choose only those events where the particles have large transverse momenta (larger than some threshold). These cuts are generally performed to increase the signal-to-background ratio in the final experimental data that is considered in the analysis. Even after the selection cuts, a typical data set that needs to be analyzed may have a sample size ranging from hundreds of thousands of events to million or billions of events. All of the methods presented in this paper can handle such large data sets (as demonstrated in Section 6). In the experiments performed in this paper, we use a 16-dimensional data set which is also typical for the searches, although the dimensionality can also be much larger for complex search channels or if low-level detector information is used.

Most model-independent approaches find the signal in experimental data by comparing it to a reference background dataset which is an essential requirement for these strategies. Hence, it is necessary for these approaches to have a trustable reference dataset. As discussed in D'Agnolo et al. (2022), it is conceptually unavoidable for these model-independent approaches to require a reference background dataset, since they search for "new" phenomena, and hence require a necessary notion of "old" phenomena. The required reference background datasets are usually generated using Monte Carlo event generators which sample collision events consistent with the equations of the Standard Model of particle physics. Some recent approaches, such as ANODE (Nachman and Shih (2020)) and CWoLa (Classification Without Labels) Bump Hunting (Collins, Howe and Nachman (2018), Collins, Howe and Nachman (2019)), estimate the background from the data itself which requires assumptions on the signal region and access to data not in the signal region. Here we do not make any assumptions on the signal region, so we use Monte Carlo simulations from the Standard Model as the reference background dataset. However, it is good to note that model-independent approaches are not necessarily restricted to Monte Carlo backgrounds, as demonstrated by ANODE and CWoLa.

Even though these model-independent approaches are dependent on the availability of a reliable, trustable background dataset to find new phenomena, the key advantage of model-independent approaches is that they can detect discrepancies between the background data and the experimental data irrespective of the distribution of the signal events. Such capability can be essential for ensuring the maximal reach of the LHC physics program. In the case that a discrepancy is found, it should be investigated further in order to understand if it results from: (a) an inaccurate background MC generator, (b) a particle detector defect or a lack of understanding of the detector, or (c) a previously unknown physics process. The active subspace methods proposed in this paper may provide a preliminary indication in this process. A model-independent search indicating a significant discrepancy can also guide the selection of new model-dependent analyses to further investigate the nature of the discrepancy.

It is important to note here that there is no one "right" approach to finding new physics signals. The different model-dependent and model-independent approaches are complementary and not superior to each other. For example, when a reliable signal model is available, model-dependent approaches should yield greater power since they use the information available about the signal model to optimize the test. It is also good to note that there is no black-and-white distinction between model-independent and model-dependent methods. Rather, there is a continuum of methods that vary in terms of the strength of the assumptions involved. Any "model-independent" method at minimum needs to make assumptions about the background. Furthermore, to perform the model-independent search, one would in practice use some amount of physics knowledge to choose a particular decay channel (say, four muons)

and to impose further cuts within that channel (e.g., to choose events where the particles have large transverse momenta) to narrow down the search to a part of the phase space that is believed to be fertile for new signals and where the background can be modeled sufficiently well. One might then be willing to make further assumptions about the location or shape of the signal region, and so on. At the far end of this spectrum are fully "model-dependent" methods that assume a specific signal model.

Additionally, both model-independent and model-dependent methods are affected by systematic uncertainties in the background (Cranmer (2015), Dorigo and de Castro (2020), Lyons and Wardle (2018)). Background systematics in high-energy physics are typically parameterized using nuisance parameters that affect, among other things, the rate and shape of the background samples. In low-dimensional situations, a standard approach is to handle these nuisance parameters using likelihood profiling (Cranmer (2015)), while recent work has focused on incorporating nuisance parameters into classifier training for high-dimensional data. Training a classifier on simulated data generated using specific values for the nuisance parameters may not be optimal for handling background data generated using other values of the nuisance parameters. Even if the classifiers are trained on data generated using the most likely values of the nuisance parameters and their effect is accounted for in the calibration, the power of the methods to detect new signals may be affected (Dorigo and de Castro (2020)). Some classifier-based model-dependent approaches have been recently suggested to handle the nuisance parameters. For example, Ghosh, Nachman and Whiteson (2021) profile the nuisance parameters by constructing classifiers that are explicitly dependent on the different values of the nuisance parameters. Reviews of model-dependent approaches that deal with nuisance parameters in classifier-based searches can be found in Nachman (2020) and Dorigo and de Castro (2020). Overall, dealing with systematic uncertainties in the background in the model-independent setting is a challenging problem that affects any model-independent method and whose detailed treatment is beyond the scope of this paper. However, we note there is no reason to believe that it would not be possible to extend existing approaches for handling background systematics from the model-dependent setting into the model-independent case. D'Agnolo et al. (2022) is, to the best of our knowledge, a first contribution toward this important goal.

Below, we present a nonexhaustive review of existing literature on model-independent approaches and other related work.

2.1. *Related literature*. Due to their advantages, model-independent approaches have been used for new physics searches (CDF Collaboration (2008), Knuteson (2000), Soha (2008)) at the Tevatron (CDF Collaboration (2009), Choudalakis (2008), D0 Collaboration (2012)), HERA (H1 Collaboration (2004)), and the LHC (ATLAS Collaboration (2019), CMS Collaboration (2017), CMS Collaboration (2020)). These methods typically compare a large set of binned distributions to the prediction from the background Monte Carlo simulation, in search for bins in the experimental data that exhibit a deviation larger than some predefined threshold. For example, the approach of ATLAS Collaboration (2019), employed by the ATLAS experiment, uses a (quasi-)model-independent method that considers some generic features of the potential new physics signals. These approaches have two limitations: (a) they do not consider the multivariate dependency structures between the variables in the data, and (b) they might miss certain signals that do not show a localized excess in one of the studied distributions.

More recent approaches like CWoLa (Classification Without Labels) Bump Hunting (Collins, Howe and Nachman (2018), Collins, Howe and Nachman (2019)), ANODE (Nachman and Shih (2020)), SALAD (Andreassen, Nachman and Shih (2020)) and simulation augmented CWoLa (SA-CWoLa) (Kasieczka et al. (2021)) are also based on searching

for anomalies by assuming that the signal is localized in a single feature. These approaches use this weak assumption about the signal which causes them to have the second limitation mentioned above, namely, they might miss signals that do not show a localized excess along one of the features. Some of the algorithms additionally assume that the signal's distribution for the other features is independent of the selected single feature that is being scanned. Kasieczka et al. (2021) describe a variety of current methods for signal detection along with results on the LHC Olympics 2020 datasets.

A different approach that uses a semisupervised nonparametric clustering algorithm is presented by Casa and Menardi (2018). They assume that, in high energy physics, a new particle manifests itself as a significant peak emerging from the background process. They use nonparametric modal clustering to search for a signal that is expected to emerge as a bump in the background distribution. BuHuLaSpa and Bump Hunter methods, presented in Kasieczka et al. (2021), also use similar bump hunting ideas to find the signal. UCluster, presented in Kasieczka et al. (2021), uses clustering to find the localized signal. These methods suffer from the second problem mentioned above: they can only find localized signals.

Model-independent semisupervised searches were also proposed by Kuusela et al. (2012) and Vatanen et al. (2012) who use multivariate Gaussian mixture models to estimate the densities of the background and the experimental data. They test for the significance of the additional Gaussian components which quantify the anomalous contribution. The drawback of this method is that Gaussian mixture models are very difficult to fit in a high-dimensional setting. Additionally, since the signal strength is typically very small, the quality of the fit influences the power of the test in detecting the signal.

These drawbacks of the mixture modeling methods along with the fact (as mentioned before) that classification algorithms have demonstrated excellent performance in detecting signals in the model-dependent approaches, especially in the high-dimensional setting, motivates us to use classifiers, instead of mixture modeling, to find the deviations of the experimental data from the background.

The problem of comparing the distributions of the background and the experimental data is also analogous to a high-dimensional two-sample testing problem (Kim, Lee and Lei (2019), Kim et al. (2021)), where the signal events appear as a collective anomaly (Chandola, Banerjee and Kumar (2009)) in a cluster close to each other. Hence, we also compare the methods proposed in this paper to nearest-neighbor two-sample tests, introduced in Schilling (1986) and Henze (1988), for a subsample of the data. In the experiments we observe that methods that use a semisupervised classifier have much higher power to detect the signal than the nearest-neighbor two-sample tests.

3. Classifier-based tests for signal detection. In this section we introduce both model-dependent and model-independent approaches to signal detection in the experimental particle physics data. Both approaches use background data from MC event generators for background events, based on the Standard Model, together with experimental data which are a mixture of background and a potential signal. First, we discuss the model-dependent approach that additionally assumes access to signal data that are generated using MC event generators for a hypothesized signal model. Then we describe the model-independent approach that does not assume access to the signal data. We now introduce formal notation to discuss the two different approaches.

The background, signal, and experimental data are samples from Poisson point processes (Cranmer (2015), Reiss (1993)). We condition on the sample sizes of the individual datasets so that the data in all three of the cases may be treated as independent samples from a density (i.e., from a binomial point process). Specifically, we have three datasets,

$$\text{Background}: \quad \mathcal{X} = \{X_1, \ldots, X_{m_b}\}, \qquad X_i \sim p_b,$$

$$\text{Signal}: \quad \mathcal{Y} = \{Y_1, \ldots, Y_{m_s}\}, \qquad Y_i \sim p_s, \tag{1}$$
$$\text{Experimental}: \quad \mathcal{W} = \{W_1, \ldots, W_n\}, \qquad W_i \sim q = (1-\lambda)p_b + \lambda p_s,$$

where p_b, p_s are the densities of the background and signal data, respectively, q is the density of the experimental data, and λ is a scalar parameter representing the signal strength.

The likelihood function for λ, given the experimental data ($\mathcal{W}$) (treating p_b and p_s as known for the moment), is

$$\mathcal{L}(\lambda) = \prod_i \big((1-\lambda)p_b(W_i) + \lambda p_s(W_i)\big). \tag{2}$$

The null hypothesis (H_0) that there is no signal corresponds to $\lambda = 0$. So the goal is to test $H_0 : \lambda = 0$ vs. $H_1 : \lambda > 0$. We additionally have the likelihood ratio.

$$\frac{\mathcal{L}(\lambda)}{\mathcal{L}(0)} = \prod_i \big(1 - \lambda + \lambda\psi(W_i)\big), \tag{3}$$

where $\psi(w) = p_s(w)/p_b(w)$. Note that the function ψ can be seen as an infinite dimensional nuisance parameter.

In the idealized case where p_b and p_s are known, we could use the usual likelihood ratio test (LRT) statistic

$$T = -2\log\big(\mathcal{L}(0)/\mathcal{L}(\hat{\lambda})\big) = 2\sum_i \log\big(1 - \hat{\lambda} + \hat{\lambda}\psi(W_i)\big), \tag{4}$$

where $\hat{\lambda}$ is the maximum likelihood estimator (MLE) of λ based on the experimental data $\mathcal{W}$. Alternatively, we could use the score test statistic

$$\mathcal{T} = \frac{1}{n}\sum_i \big(\psi(W_i) - 1\big). \tag{5}$$

In practice, the densities p_b and p_s are unknown. The two different approaches, described in this section, show how a classifier can be used to estimate the desired statistics directly. We prefer estimating the density ratio required for the desired statistics directly instead of taking the ratio of the estimated densities. This is due to the high-dimensionality of the data which makes estimating the high-dimensional density with limited data very difficult.

3.1. *The model-dependent* (*supervised*) *case*. In this case we make use of the signal data $\mathcal{Y}$, where $Y_1, \ldots, Y_{m_s} \sim p_s$ are generated using a MC event generator for a hypothized signal model. Such approach is standard in most new physics searches at the LHC and serves here as a point of comparison against the model-independent methods. The strategy underlying our implementation of this approach is to use a classifier to estimate $\psi = p_s/p_b$ and then use the LRT or the score test with the estimated ψ.[1] Since ψ is estimated, we cannot rely on standard asymptotics to get the null distribution. Instead, we use permutation or nonparametric bootstrap methods.

Before we train a classifier, we first combine the background and signal data into a single dataset

$$\{Z_1, \ldots, Z_{m_b+m_s}\} = \mathcal{X} \cup \mathcal{Y} = \{X_1, \ldots, X_{m_b}, Y_1, \ldots, Y_{m_s}\},$$

[1] In most LHC analyses, ψ is currently used to extract a signal-enriched subset of the data which is then used in a low-dimensional parametric fit to form a test statistic (Radovic et al. (2018)). Here we use instead the ψ-based high-dimensional LRT or score test to provide an apples-to-apples comparison with the model-independent methods.

and we define $S_i = 1$ if Z_i is from the signal data $\mathcal{Y}$ and $S_i = 0$ otherwise. We treat (Z_i, S_i) as a sample from a density $p(z, s)$ with $\pi_0 := \mathbb{P}(S = 1) = m_s/(m_b + m_s)$, the probability of any sample being from the signal distribution.

Let $h_0(z) := \mathbb{P}(S = 1|Z = z)$ denote the probability of a sample being from the signal distribution given $Z = z$. Then using Bayes's rule,

$$h_0(z) = \mathbb{P}(S = 1|Z = z) = \frac{\pi_0 \psi(z)}{\pi_0 \psi(z) + (1 - \pi_0)}. \tag{6}$$

Inverting this,

$$\psi(z) = \left(\frac{1 - \pi_0}{\pi_0}\right)\left(\frac{h_0(z)}{1 - h_0(z)}\right). \tag{7}$$

This leads to our estimate of ψ,

$$\widehat{\psi}(z) = \left(\frac{1 - \pi_0}{\pi_0}\right)\left(\frac{\hat{h}_0(z)}{1 - \hat{h}_0(z)}\right), \tag{8}$$

where $\hat{h}_0(z)$ is a classifier that separates the signal data $\mathcal{Y}$ from the background data $\mathcal{X}$. In this paper we take $\hat{h}_0$ to be a random forest, but in principle, any classifier can be used.

To use $\widehat{\psi}(z)$ to calculate the LRT statistic in (4), we estimate λ using its MLE as follows. We can write the likelihood of the experimental data as

$$\mathcal{L}(\lambda) = \prod_i p_b(W_i) \times \prod_i \big(1 - \lambda + \lambda \psi(W_i)\big). \tag{9}$$

The first term does not involve λ, so we can ignore it. Hence,

$$\mathcal{L}(\lambda) \propto \prod_i \big(1 - \lambda + \lambda \psi(W_i)\big). \tag{10}$$

We then define the estimated likelihood

$$\hat{\mathcal{L}}(\lambda) \propto \prod_i \big(1 - \lambda + \lambda \hat{\psi}(W_i)\big), \tag{11}$$

and we define $\hat{\lambda}$ to be the maximizer of $\hat{\mathcal{L}}(\lambda)$. We can now use $\hat{\lambda}$ and $\hat{\psi}$ to estimate the LRT and the score test statistics, as given in (4) and (5).

To see the effect of maximizing the estimated likelihood, using the estimated ψ instead of maximizing the actual likelihood, let $\ell(\lambda) = \log \mathcal{L}(\lambda)$, let $\hat{\ell}(\lambda) = \log \hat{\mathcal{L}}(\lambda)$, and note that, for small λ,

$$\ell(\lambda) = \lambda \sum_i \big(\psi(W_i) - 1\big) - \frac{\lambda^2}{2} \sum_i \big(\psi(W_i) - 1\big)^2 + o_P\big(\lambda^2\big) + C, \tag{12}$$

where $C = \sum_i \log(p_b(W_i))$ is just a constant and can be ignored. A similar relation holds for $\hat{\ell}(\lambda)$ and $\hat{\psi}$. So

$$\frac{1}{n}\hat{\ell}(\lambda) - \frac{1}{n}\ell(\lambda) = \frac{\lambda}{n} \sum_i \big(\hat{\psi}(W_i) - \psi(W_i)\big) + o_P(\lambda), \tag{13}$$

which shows how the accuracy of the classifier affects the log-likelihood. The maximizer $\tilde{\lambda}$ of $\ell(\lambda)$ and the maximizer $\hat{\lambda}$ of $\hat{\ell}(\lambda)$ are

$$\tilde{\lambda} = \left[\frac{\sum_i (\psi(W_i) - 1)}{\sum_i (\psi(W_i) - 1)^2}\right]_+ + o_P(\lambda), \qquad \hat{\lambda} = \left[\frac{\sum_i (\hat{\psi}(W_i) - 1)}{\sum_i (\hat{\psi}(W_i) - 1)^2}\right]_+ + o_P(\lambda). \tag{14}$$

Hence, $\hat{\lambda} - \tilde{\lambda} = O_P(\frac{1}{n}\sum_i(\hat{\psi}(W_i) - \psi(W_i)))$, emphasizing again the importance of an accurate classifier. Note that, in practice, instead of using the above approximation, we evaluate the MLE of λ by performing a grid search on $[0, 1]$.

As $n \to \infty$, the usual likelihood ratio test statistic (Böhning et al. (1994), Ghosh and Sen (1985)),

$$T = 2\sum_i \log\big((1-\hat{\lambda}) + \hat{\lambda}\psi(W_i)\big) \overset{d}{\rightsquigarrow} \frac{1}{2}\delta_0 + \frac{1}{2}\chi^2_1, \tag{15}$$

where δ_0 is a degenerate distribution at 0. But when $\hat{\psi}$ is substituted for ψ, the asymptotic distribution is unknown, so the null distribution needs to be estimated by simulating from the background model. If the available background data is limited, we can use permutation or nonparametric bootstrap methods; see Section 6.2.

Similar remarks apply to the score statistic $\mathcal{T} = n^{-1}\sum_i(\psi(W_i) - 1)$ and its estimated version $\hat{\mathcal{T}} = n^{-1}\sum_i(\hat{\psi}(W_i) - 1)$. There are conditions under which $\hat{\mathcal{T}}$ has a tractable distribution. Suppose that $\hat{\psi}$ is estimated on part of the data and $\hat{\mathcal{T}}$ on another. Now

$$\sqrt{n}\hat{\mathcal{T}} = \sqrt{n}\mathcal{T} + \sqrt{n}\bigg(\frac{1}{n}\sum_i\big(\hat{\psi}(W_i) - \psi(W_i)\big)\bigg). \tag{16}$$

The first term converges to $N(0, \sigma^2)$, where $\sigma^2 = \mathbb{E}_{p_b}[(\psi(W) - 1)^2]$. If ψ is in a Hölder class of smoothness index β and $\beta > d/2$, then it can be shown that $\frac{1}{n}\sum_i(\hat{\psi}(W_i) - \psi(W_i)) = O_P(n^{-2\beta/(2\beta+d)})$, where d is the dimension of the W_i's. Hence, if $\beta > d/2$, the second term is negligible so that $\sqrt{n}\hat{\mathcal{T}} \rightsquigarrow N(0, \sigma^2)$. However, we have found that the Normal approximation is poor in practice, and instead we use permutation and bootstrap methods to approximate the null distribution. Similar to the LRT, the null distribution can also be estimated by simulating from the background model.

To use nonparametric bootstrap or permutation methods, we randomly split the available background data $\mathcal{X}$ into two sets $\mathcal{X}_1$ and $\mathcal{X}_2$. We use the first set $\mathcal{X}_1$ along with the signal data $\mathcal{Y}$ to train the classifier $\hat{h_0}$. We use the second set $\mathcal{X}_2$ along with the experimental data $\mathcal{W}$ to approximate the null distributions of the LRT and the score statistics. Note that, under the null $H_0 : \lambda = 0$, the experimental data $\mathcal{W}$ and background data $\mathcal{X}_2$ have the same distribution p_b. In the case of nonparametric bootstrap, we approximate the null distributions by repeatedly sampling with replacement from $\mathcal{X}_2 \cup \mathcal{W}$ and computing the test statistics. For the permutation test, we do the same using sampling without replacement.

3.2. *Model-independent (semisupervised) case.* In this case we assume that we do not have access to (or do not completely trust) the signal training sample $\mathcal{Y} = \{Y_1, \ldots, Y_{m_s}\}$. So the data available are $\mathcal{X} = \{X_1, \ldots, X_{m_b}\}$ and $\mathcal{W} = \{W_1, \ldots, W_n\}$, where $X_i \sim p_b$ and $W_i \sim q = (1-\lambda)p_b + \lambda p_s$, with p_b, p_s and λ unknown. We want to test $H_0 : \lambda = 0$ vs. $H_1 : \lambda > 0$ which is equivalent to testing $H_0 : p_b = q$ vs. $H_1 : p_b \neq q$. Hence, we are in a two-sample testing scenario (Kim, Lee and Lei (2019), Kim et al. (2021)).

Again, we want to leverage the fact that classifiers have been found to give accurate estimates of density ratios (Hastie, Tibshirani and Friedman (2009), Ch.14) (Cranmer, Pavez and Louppe (2015), Goodfellow et al. (2014)). One strategy is to use a classifier like before to obtain a likelihood ratio test statistic, but this time we estimate the density ratio $\psi^\dagger = q/p_b$. To do this, we train a classifier to differentiate between the experimental ($\mathcal{W}$) and background events ($\mathcal{X}$), instead of the signal ($\mathcal{Y}$) and background events ($\mathcal{X}$). As mentioned in the Introduction, this strategy is similar to the one taken by D'Agnolo et al. (2021) and D'Agnolo and Wulzer (2019), who use a neural network to estimate the likelihood ratio test statistic.

A second strategy is to use the area under the curve statistic (AUC) (Hanley and McNeil (1982)) or the misclassification error/misclassification rate (MCE) to evaluate the performance of the classifier. The intuition behind the second strategy is that, in the absence of signal under the null, a classifier should not be able to differentiate between the background and the experimental data. We discuss both the strategies below.

As before, let h denote a classifier in the combined (background and experimental) sample, then

$$\psi^{\dagger}(z) = \left(\frac{1-\pi}{\pi}\right)\left(\frac{h(z)}{1-h(z)}\right), \tag{17}$$

where now $\pi = n/(n+m_b)$. This leads to our estimate of $\psi^{\dagger}$

$$\hat{\psi}^{\dagger}(z) = \left(\frac{1-\pi}{\pi}\right)\left(\frac{\hat{h}(z)}{1-\hat{h}(z)}\right), \tag{18}$$

where $\hat{h}(z)$ is the trained classifier.

The likelihood ratio, as given in (3), can be rewritten as

$$\frac{\mathcal{L}(\lambda)}{\mathcal{L}(0)} = \prod_i \psi^{\dagger}(W_i). \tag{19}$$

Hence, the LRT statistic is given by

$$T = 2\sum_i \log \hat{\psi}^{\dagger}(W_i). \tag{20}$$

Similar to the arguments presented in the model-dependent case, since we are estimating $\psi^{\dagger}$ using $\hat{\psi}^{\dagger}$, usual asymptotics do not hold for the LRT statistic. Instead, we propose a conditional asymptotic test. We split the background and the experimental data into training and test data and estimate $\hat{h}$ on the training data. We then use the test experimental data to calculate the test statistic T and use the test background data to approximate the null distribution. Unlike the chi-squared null distribution, considered by D'Agnolo et al. (2021) and D'Agnolo and Wulzer (2019), we instead use an asymptotic Normal distribution to approximate the null distribution. In this paper the split of the data into training and test data sets is performed roughly equally. The reason for this is that if the size of the training data is small, we might not have sufficient signal samples in the training data for the classifier to recognize them, and if the size of the test data is small, there might not be enough signal samples in the test set for the test to give a significant result. We additionally propose nonparametric bootstrap and permutation methods to approximate the null distribution as well.

For the second strategy that uses the performance of the classifier ($\hat{h}$) in separating the background data from the experimental data, we consider two statistics—AUC (area under the curve) statistic and MCE (misclassification error) statistic.

A conventional summary of the ROC curve is the area under the curve, or AUC. The ROC curve (Hanley et al. (1989), Metz (1978)) demonstrates the performance of a classifier by plotting the true positive rate (TPR) vs. the false positive rate (FPR) at various threshold settings of the classifier output. For example, for our classifier $\hat{h}(z)$, given a threshold parameter t, an instance Z is classified as experimental ("positive") if $\hat{h}(Z) > t$ and background ("negative") otherwise. Now $Z \sim q$ if Z actually belongs to the experimental class, and $Z \sim p_b$ if Z actually belongs to the background class. Therefore, the TPR is given by $\mathrm{TPR}(t) = \mathbb{P}_q(\hat{h}(Z) > t)$, and the FPR is given by $\mathrm{FPR}(t) = \mathbb{P}_{p_b}(\hat{h}(Z) > t)$, where $\mathbb{P}_q$ is the probability when $Z \sim q$ and $\mathbb{P}_{p_b}$ is the probability when $Z \sim p_b$. Since the ROC curve plots $\mathrm{TPR}(t)$ (y-axis) vs. $\mathrm{FPR}(t)$ (x-axis) with varying t, the AUC θ is given by

$$\theta = \int_0^1 \mathrm{TPR}\big(\mathrm{FPR}^{-1}(x)\big)\,dx = \mathbb{P}\big(\hat{h}(W) > \hat{h}(X)\big), \tag{21}$$

by standard derivation. So the AUC can also be interpreted as the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. Hence, we can estimate it using

$$\hat{\theta} = \frac{1}{m_b n} \sum_i \sum_j \mathbb{I}\{\hat{h}(W_j) > \hat{h}(X_i)\}. \tag{22}$$

Under the null $H_0 : \lambda = 0$, we have that $q = p_b$, that is, X and W have the same distribution. So under H_0, $\theta = 0.5$ and an AUC that is significantly greater than 0.5 provides evidence of $q \neq p_b$. In other words, testing $H_0 : \lambda = 0$ vs. $H_1 : \lambda > 0$ is equivalent to testing $H_0 : \theta = 0.5$ vs. $H_1 : \theta > 0.5$.

We can also use the MCE (misclassification error/classification error rate) to measure the performance of the classifier (Kim et al. (2021)), where we define the MCE as

$$M = 0.5\mathbb{P}\big(\hat{h}(X) > \pi\big) + 0.5\mathbb{P}\big(\hat{h}(W) < \pi\big). \tag{23}$$

Note that this is the average of the false positive rate (first term) and the false negative rate (second term) for threshold $\pi = n/(n + m_b)$. This can be estimated using

$$\hat{M} = \frac{1}{2}\left[\frac{1}{m_b} \sum_i \mathbb{I}\{\hat{h}(X_i) > \pi\} + \frac{1}{n} \sum_j \mathbb{I}\{\hat{h}(W_j) < \pi\}\right]. \tag{24}$$

Under the null, $H_0 : \lambda = 0$, and as a result, $q = p_b$; that is, $X_i \overset{d}{=} W_j$. Hence, $M = 0.5$ under the null. The classifier will have a true accuracy significantly above half (and hence M below half), only if $q \neq p$. Hence, we can use $\hat{M}$ as a test statistic and test $H_0 : M = 0.5$ vs. $H_1 : M < 0.5$.

As with the LRT statistic, the tests using the AUC and the MCE statistics can also be performed using asymptotic, bootstrap and permutation methods. The asymptotic AUC method, unlike the asymptotic LRT and MCE methods, does not use a conditional test. For the AUC we derive the asymptotic distribution of the statistic using results presented in Newcombe (2006). These methods also use data splitting, where we use the training data to fit the classifier ($\hat{h}$) and use the test data to perform the test. We detail the nonparametric bootstrap and one of the permutation methods in Method 3.1 below. We provide algorithms for the other methods, including the asymptotic methods using the three statistics LRT, AUC, and MCE, and the slower in-sample permutation method in the Supplementary Material (Chakravarti et al. (2023)).

REMARK.

1. All the model-independent methods presented in this section, except the in-sample permutation method, use data splitting of the background and the experimental data into training and test data, to approximate the null distribution and to perform the test. This has the benefit that the classifier can be kept fixed when estimating the null distribution. Here splitting a sample means randomly splitting the sample into two disjoint subsamples. The training data is used to train the classifier $\hat{h}$, and the test data is used to perform the signal detection test. The in-sample permutation method, on the other hand, retrains the classifier on permuted background and experimental data when approximating the null. This has the advantage of using all the data to train the classifier, but the computing time is much longer due to the classifier retraining required when obtaining the null distribution.

2. It is also important to note that here we used the random forest (RF) classifier to predict posterior probabilities. However, the "vote tallies" produced by RF classifiers are not posterior probabilities from a generative model and very often are uncalibrated when interpreted

as such. To calibrate the random forest classifier outputs, one can use methods to improve the calibration by using Platt scaling or other more sophisticated methods (Boström (2008), Niculescu-Mizil and Caruana (2005)) on the random forest outputs before performing the tests. We performed several experiments for the model-independent asymptotic tests using Platt scaling on the random forest outputs for calibration and the results were similar to the uncalibrated random forests used here. So we leave the decision of whether to further calibrate the random forest outputs before using the tests to the discretion of the user.

METHOD 3.1. Bootstrap and permutation methods—faster than the in-sample permutation method, but slower than the asymptotic methods:

1. Split background data $\mathcal{X} = \{X_1, \ldots, X_{m_b}\}$ into $\mathcal{X}_1$ and $\mathcal{X}_2$ of sizes m_1 and m_2, respectively.
2. Split experimental data $\mathcal{W} = \{W_1, \ldots, W_n\}$ into $\mathcal{W}_1$ and $\mathcal{W}_2$ of sizes n_1 and n_2, respectively, with $n_2 = m_2$
3. Train the classifier $\hat{h}$ on $\mathcal{X}_1$ and $\mathcal{W}_1$.
4. Evaluate the LRT statistic T on $\mathcal{W}_2$ using (18) and (20) as

$$\tilde{T} = \frac{T}{2n_2} = \log\Big(\frac{1-\pi}{\pi}\Big) + \frac{1}{n_2}\sum_{W_i \in \mathcal{W}_2} \log\Big(\frac{\hat{h}(W_i)}{1-\hat{h}(W_i)}\Big), \tag{25}$$

where $\pi = n_1/(m_1+n_1)$. Similarly, evaluate the AUC statistic $\hat{\theta}$, as defined in equation (22), and the MCE statistic $\hat{M}$, as defined in equation (24), on $\mathcal{X}_2$ and $\mathcal{W}_2$ as

$$\hat{\theta} = \frac{1}{m_2 n_2}\sum_{X_i \in \mathcal{X}_2}\sum_{W_j \in \mathcal{W}_2} \mathbb{I}\{\hat{h}(W_j) > \hat{h}(X_i)\}, \tag{26}$$

$$\hat{M} = \frac{1}{2}\Big[\frac{1}{m_2}\sum_{X_i \in \mathcal{X}_2} \mathbb{I}\{\hat{h}(X_i) > \pi\} + \frac{1}{n_2}\sum_{W_j \in \mathcal{W}_2} \mathbb{I}\{\hat{h}(W_j) < \pi\}\Big], \tag{27}$$

respectively.

5. Estimate the null distribution of $\tilde{T}$, $\hat{\theta}$ and $\hat{M}$, by repeatedly drawing $m_2 + n_2$ random observations from $\mathcal{X}_2 \cup \mathcal{W}_2$ (with replacement for bootstrap and without replacement for permutation) and randomly labelling m_2 of them as X's and n_2 of them as W's before computing $\tilde{T}$, $\hat{\theta}$ and $\hat{M}$ on them. (Note that, under the null, $q = p_b$, implying that the X's and the W's have the same distribution.)
6. Calculate the p-values based on the estimated null distributions.

4. Estimating λ in the model independent scenario. Here we discuss the problem of estimating the signal strength λ in the semisupervised setting. As we have seen in Section 3.2, the ratio of the densities $\psi^\dagger = q/p_b$ can be written as

$$\psi^\dagger(z) = \Big(\frac{1-\pi}{\pi}\Big)\Big(\frac{h(z)}{1-h(z)}\Big),$$

where $\pi = n/(n+m_b)$ and h is the classifier differentiating the experimental data from the background data. Since

$$\psi^\dagger(z) = \frac{q(z)}{p_b(z)} = \frac{(1-\lambda)p_b(z) + \lambda p_s(z)}{p_b(z)} = 1 - \lambda + \lambda\psi(z), \tag{28}$$

we see that, for any z such that $\psi(z) \neq 1$, we have

$$\lambda = \frac{1 - (\frac{1-\pi}{\pi})(\frac{h(z)}{1-h(z)})}{1-\psi(z)}. \tag{29}$$

Hence, if we can find any z for which $p_s(z) = 0$ (no signal), then we can estimate λ by

$$\hat{\lambda} = 1 - \left(\frac{1-\pi}{\pi}\right)\left(\frac{\hat{h}(z)}{1-\hat{h}(z)}\right), \tag{30}$$

where $\hat{h}(z)$ is the trained semisupervised classifier. The problem is that the search for such a z may not be obvious.

Instead, we take a different approach. To ensure identifiability, we assume $\inf_z p_s(z)/p_b(z) = 0$. We believe this is true in most high-energy physics problems. To simplify the discussion, we also assume $p_b, q > 0$ everywhere.

Next, we show that the problem of estimating λ can be transformed into a problem of estimating $g_q(1)$, where g_q is the density of a univariate random variable supported on $[0, 1]$. We define for any $t \geq 0$, $C_t = \{z \in \mathbb{R}^d : p_s(z) \geq t p_b(z)\}$. Then for any $z \in \mathbb{R}^d$, we define the Neyman–Pearson quantile transform of z as

$$\rho(z) = \mathbb{P}_{X\sim p_b}\big(\psi^\dagger(X) \geq \psi^\dagger(z)\big) = \mathbb{P}_{X\sim p_b}\big(h(X) \geq h(z)\big). \tag{31}$$

Let g_0, g_1, and g_q be the density functions of $\rho(Z)$ when $Z \sim p_b$, $Z \sim p_s$, and $Z \sim q$, respectively. Then we can show, as stated in Theorem 4.1 below, that $g_q(1) = 1 - \lambda$, and, therefore, $\lambda = 1 - g_q(1)$. So, we can estimate λ using

$$\hat{\lambda} = 1 - \hat{g_q}(1), \tag{32}$$

where $\hat{g_q}$ is a density estimate based on the $\hat{\rho}(W_i)$'s defined as

$$\hat{\rho}(W_i) = \frac{1}{m_b}\sum_j \mathbb{I}\big\{\hat{h}(X_j) \geq \hat{h}(W_i)\big\}. \tag{33}$$

THEOREM 4.1. *Assuming $\rho(Z)$ has continuous density under either $Z \sim p_b$ or $Z \sim p_s$, then the following hold*:

1. $g_0(u) = 1$ *for all* $u \in [0, 1]$. *That is*, $\rho(Z) \sim \text{Unif}(0, 1)$ *if* $Z \sim p_b$.
2. $g_1(u) = t_u$ *where* t_u *satisfies* $P_{X\sim p_b}(X \in C_{t_u}) = u$. *In particular*, $g_1(1) = 0$.
3. $g_q(1) = 1 - \lambda$.

We detail the proof of the theorem in Section 1 of the Supplementary Material (Chakravarti et al. (2023)).

Now the problem of estimating λ reduces to estimating a monotone density at a boundary point. We can estimate the density $g_q(1)$, based on the $\hat{\rho}(W_i)$'s, using a simple histogram based estimator,

$$\hat{g_q}(1) = \frac{1}{nb}\sum_i \mathbb{I}\big\{\hat{\rho}(W_i) \in (1-b, 1]\big\}, \tag{34}$$

where b is the bin-width of the histogram estimator. But, since the density is a monotonically decreasing function, the density estimates at points close to 1 could also be indicative of the estimate at 1.

We, therefore, propose using a Poisson regression on bins close to 1 in order to estimate the density at 1. We fit a Poisson regression $\hat{f}(t) = \exp(\beta_0 + \beta_1 t)$, with $\beta_1 \leq 0$ to the histogram estimates

$$H_t = \sum_i \mathbb{I}\big\{\hat{\rho}(W_i) \in (t-b, t]\big\}, \quad T \leq t - b \leq t \leq 1, \tag{35}$$

where b is the bin-width of the histogram estimator and T determines the neighborhood of 1 that is used to estimate the density at 1. Then the estimated density at 1 is given by

$$\hat{g_q}(1) = \hat{f}(1), \tag{36}$$

the estimate given by the Poisson regression at 1.

REMARK.

1. Note that the bins $(t-b, t]$ for $T \le t-b \le t \le 1$ form a partition of $[T, 1]$, and we regress on the bin end points for the Poisson regression model.
2. We constrain $\beta_1 \le 0$, since we know that g_q is a monotonically decreasing function. In practice, we implement this condition by setting $\widehat{\beta_1} = 0$, when the Poisson regression models estimates $\widehat{\beta_1} > 0$ and fit $\hat{f}(t) = \exp(\beta_0)$.
3. In practice, we additionally perform data-splitting in order to get out-of-sample estimates of λ. It's important to consider out-of-sample estimates since the Poisson regression model is conditioned on the trained classifier and computing the signal strength estimates on the same data that is used to train the classifier could give biased estimators.
4. Instead of fitting a Poisson regression model on the histogram estimates H_t, we additionally tried using just the mean of the density estimates, given by the histogram above a threshold T, to estimate $\hat{f}(1)$. We also tried using a simple linear regression on $\log(H_t)$. These methods experimentally gave poorer estimates compared to the Poisson model.

We compute the out-of-sample estimate of λ, as mentioned in Method 4.1 below. We can additionally use nonparametric bootstrap to understand the stability of the signal strength estimates to perturbations in the data. We detail the bootstrap process below in Method 4.2. The method provides standard error estimates as well as bootstrapped confidence intervals that characterize the stability of the estimated signal strength λ.

METHOD 4.1. Estimating the signal strength λ:

1. Split background data $\mathcal{X} = \{X_1, \ldots, X_{m_b}\}$ into $\mathcal{X}_1$ and $\mathcal{X}_2$ of sizes m_1 and m_2, respectively.
2. Split experimental data $\mathcal{W} = \{W_1, \ldots, W_n\}$ into $\mathcal{W}_1$ and $\mathcal{W}_2$ of sizes n_1 and $n_2 = m_2$, respectively.
3. Train the classifier $\hat{h}$ on $\mathcal{X}_1$ and $\mathcal{W}_1$.
4. Compute $\hat{\rho}(W_i)$ for all $W_i \in \mathcal{W}_2$, as defined in (33), based on $\hat{h}$ and $\mathcal{X}_2$ as

$$\hat{\rho}(W_i) = \frac{1}{m_2} \sum_{X_j \in \mathcal{X}_2} \mathbb{I}\{\hat{h}(X_j) \ge \hat{h}(W_i)\}, \quad W_i \in \mathcal{W}_2. \tag{37}$$

5. Get histogram estimates H_t of $\hat{\rho}(W_i)$'s for bins larger than T with bin-width b using equation (35).
6. Use Poisson regression to estimate the density at 1 as $\hat{g_q}(1)$. Then $\hat{\lambda} = 1 - \hat{g_q}(1)$.

METHOD 4.2. Bootstrapped uncertainty intervals for $\hat{\lambda}$:

1. Repeatedly draw with replacement m_b background samples from $\mathcal{X}$ and n experimental samples from $\mathcal{W}$, and split them into $\mathcal{X}_1^*$, $\mathcal{X}_2^*$, $\mathcal{W}_1^*$ and $\mathcal{W}_2^*$ of sizes m_1, m_2, n_1, and n_2 at random ensuring no overlap between the training and test data sets. That is, $\mathcal{X}_1^* \cap \mathcal{X}_2^* = \phi$ and $\mathcal{W}_1^* \cap \mathcal{W}_2^* = \phi$. m_1, n_1, m_2 and n_2 are same as ones used in Method 4.1 above.
2. Find $\hat{\lambda}^*$ in each case using steps 3–6 in Method 4.1. Note that the classifier is retrained on $\mathcal{X}_1^* \cup \mathcal{W}_1^*$ for every random sample.

3. We can then use the empirical standard deviation or quantiles of the $\hat{\lambda}^*$'s to create bootstrap confidence intervals that will give estimates of the stability of the estimator $\hat{\lambda}$ to perturbations in the data.

REMARK. Since the data are split to consider an out-of-sample estimate of λ, it is also necessary in the bootstrapped samples for the training set to be disjoint of the test set to avoid problems caused by considering an in-sample estimate.

5. Interpreting the classifier. The signal detection test relies on the classifier; here we use random forest. So to understand, characterize and interpret the new physics signal detected by the test, it is useful and necessary to find a way to interpret the fitted classifier. In this section we propose two methods to help interpret the random forest classifier.

The methods are based on understanding how the gradient of the classifier surface ($\hat{h}(z)$) varies. The underlying idea is that directions in which the surface is sloped or changes a lot contain useful information for the classification, while directions in which the surface is flat do not. For the first method, we look at the density of the gradients of the classifier marginally along each of the variables. For the second method, we find multivariate dependencies that jointly affect the gradient of the classifier via finding the active subspace (Constantine (2015)) of the classifier surface. The active subspace is found by looking at the leading eigenvectors of the standardized gradients of the classifier surface by performing PCA on the standardized gradients.

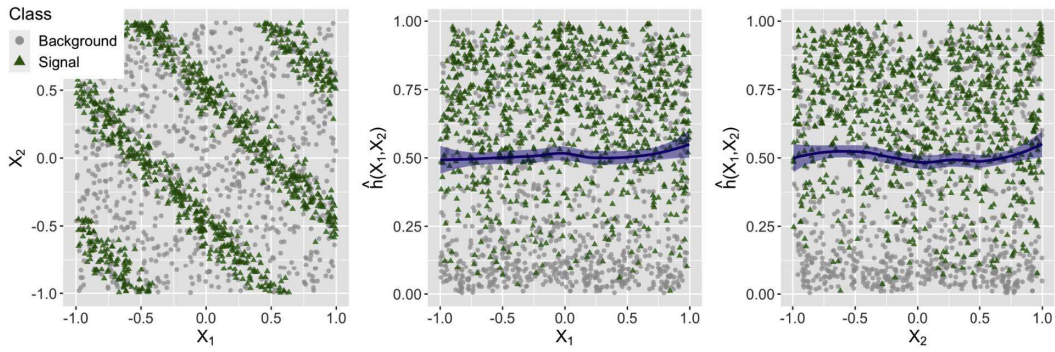
Figure 3 demonstrates the active subspace method on a two-dimensional example for simplicity, where the data $(X_1, X_2) \in [-1, 1] \times [-1, 1]$ and the signal lies around lines parallel to the line $X_1 + X_2 = 0$. We then train a classifier $\hat{h}(X_1, X_2)$ to separate the signal from the background. We notice that the classifier output does not appear to have any relationship with either X_1 or X_2 marginally, as shown in Figure 3a. But the smoothed classifier surface detects the signal around lines parallel to the line $X_1 + X_2 = 0$, as can be seen by the ridges along those lines in Figure 3b. So looking at the standardized gradients of the classifier surface gives us information about the direction in which the surface changes. As seen in Figure 3c, the first eigenvector of the standardized gradients reflects the direction in which the classifier surface changes the most which is along the $X_1 - X_2 = 0$ line. Identifying this subspace helps us understand the directions in the feature space that separate the signal from the background. Note that we consider a two-dimensional example here for illustration purposes only. The real data can be of much higher dimensionality.

In what follows, instead of directly looking at the classifier $\hat{h}(z)$, we look at $\operatorname{logit}(\hat{h}(z)) = \log(\hat{h}(z)/(1 - \hat{h}(z)))$. In practice for computational purposes, we add a small noise (e.g., 10^{-10}) to $\hat{h}(z)$ when $\hat{h}(z) \in \{0, 1\}$ to ensure that $\operatorname{logit}(\hat{h}(z))$ stays finite. We notice that taking the logit transformation provides more stable estimates of the gradients of the surface. Henceforth, we will instead look at the gradients of $H(z) := \operatorname{logit}(\hat{h}(z))$.

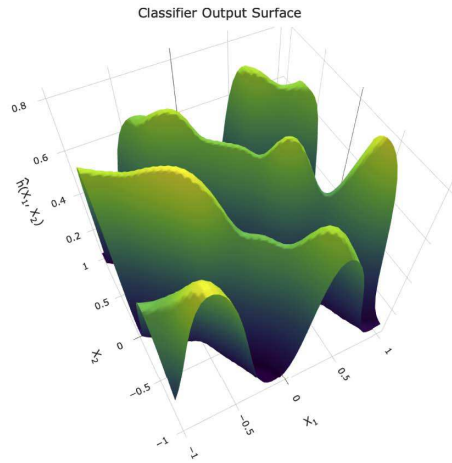
To find the gradient $\nabla_z H(z)$, we fit a local linear smoother to the logit of the random forest output. That is, we fit the logit of the classifier outputs $H(Z_i) = \operatorname{logit}(\hat{h}(Z_i))$ locally around z^* using

$$\hat{H}(z) = \alpha(z^*) + \beta(z^*)^T (z - z^*) + o(\|z - z^*\|_2^2). \tag{38}$$

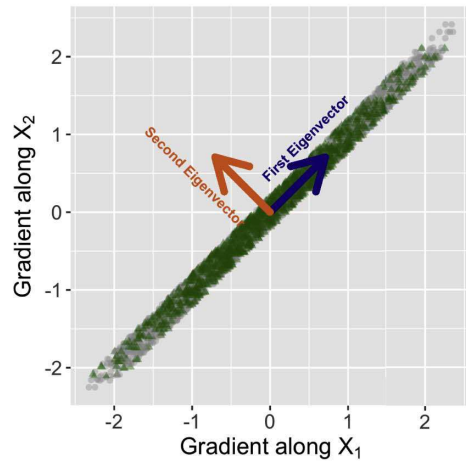
Then $\hat{\beta}(Z_i)$ provides estimates of the gradient of the logit classifier output on the data. We furthermore replace the gradient estimates $\hat{\beta}_j(Z_i)$ by their standardized versions $\hat{\beta}_j(Z_i)/sd(\hat{\beta}_j(Z_i))$ at every data point Z_i to get more stabilized values. So henceforth, $\hat{\beta}(Z_i)$ will be used to indicate the standardized gradient estimates for notational simplicity.



(a) X_1 vs. X_2, $\widehat{h}(X_1, X_2)$ vs. X_1 and $\widehat{h}(X_1, X_2)$ vs. X_2



(b) Smoothed Classifier Surface



(c) PCA of the Standardized Gradients

FIG. 3. *We demonstrate the active subspace method on a two-dimensional example, as shown in* (*a*), *where grey circles denote the uniformly distributed background and the green triangles denote the signal. The signal lies around lines parallel to the line* $X_1 + X_2 = 0$. *Additionally,* (*a*) *also shows the output of the classifier that separates the signal from the background,* $\hat{h}(X_1, X_2)$, *as a function of* X_1 *and* X_2 *individually. Neither of these variables marginally has separation power for the signal.* (*b*) *shows the smoothed classifier surface as a function of both* X_1 *and* X_2. *The active subspace method then performs PCA on the standardized gradients of the smoothed classifier output in* (*b*). (*c*) *shows a scatter plot of the standardized gradients of the smoothed classifier output as well as the two eigenvectors. We see that, as expected, the first eigenvector picks the direction in which the classifier output varies the most, namely,* $X_1 - X_2 = 0$. *We consider a two-dimensional example here for illustration purposes only. The real data can be of much higher dimensionality.*

REMARK. The data used for estimating the gradient can either be the background data or the experimental data or a combination of both. We use a combination of both, since experimentally we see that using the combination captures the classifier surface better.

We can then look at:

(i) The density of $\hat{\beta}_j(Z)$, the estimated standardized gradient of $H(Z)$ along the jth variable.

(ii) The active subspace found using the estimated standardized gradients $\hat{\beta}(Z)$, as detailed in Section 5 below.

The density of $\hat{\beta}_j(Z)$ explains how the classifier behaves marginally along each variable, whereas the active subspace identifies relationships between multiple variables that cause the most change in the classifier. We now describe the active subspace method in detail below.

5.1. *Active subspace of the classifier.* In this section we find the active subspace of the classifier $\hat{h}$, which is assumed to be fixed, and all the expectations in this section are with respect to the inputs of the fixed classifier. Let us consider the mean and the covariance matrix of $\nabla_z H(z)$. We define

$$C = \mathbb{E}_f\big[(\nabla_z H - \mathbb{E}[\nabla_z H])(\nabla_z H - \mathbb{E}[\nabla_z H])^T\big], \tag{39}$$

where $\mathbb{E}_f$ is the expectation with respect to the density $f = cp_b + (1-c)q$, where $c = m_2/(m_2+n_2)$ gives the proportion of background data in the combined sample of background data $\mathcal{X}_2$ of size m_2 and experimental data $\mathcal{W}_2$ of size n_2 and $q = (1-\lambda)p_b + \lambda p_s$. We consider expectations with respect to f since, as explained in the remark above, we use a combination of the experimental data and the background data to find the active subspace of the classifier.

The mean $\mathbb{E}_f[\nabla_z H]$ gives the expected slope of the surface of $H(z)$. For example, a positive mean gradient along the jth variable indicates an increasing trend in the classifier output with that variable, meaning that increasing that variable increases the probability of an experimental data point and decreases the probability of a background data point. The covariance matrix C, on the other hand, encodes the variable relationships that cause variability in the classifier surface.

Consider the real eigenvalue decomposition of C,

$$C = M\Lambda M^T, \quad \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_d), \lambda_1 \geq \cdots \geq \lambda_d \geq 0, \tag{40}$$

where M has columns $\{M_{\cdot 1}, \ldots, M_{\cdot d}\}$, the normalized eigenvectors of C. Then the vector $M_{\cdot 1}$, corresponding to λ_1, gives the association between the variables (i.e., the direction in the input space) that best captures the changes in $H(z)$ around the expected slope, followed by $M_{\cdot 2}$ and so on. Therefore, the eigenvectors corresponding to the leading eigenvalues $\lambda_1, \lambda_2, \ldots$, give us an idea about the directions along which the classifier output changes the most. These are directions that contain meaningful information for separating the experimental data from the background data and, therefore, enable us to characterize how the experimental data differs from the background data. Towards this end, we propose Method 5.1 that uses the gradient estimates $\hat{\beta}(Z)$ derived from a local linear smoother.

METHOD 5.1. Active subspace for the classifier $\hat{h}$:

1. Split background data into $\mathcal{X}_1$ and $\mathcal{X}_2$ of sizes m_1 and m_2, respectively. Split experimental data into $\mathcal{W}_1$ and $\mathcal{W}_2$ of sizes n_1 and $n_2 = m_2$, respectively. Train the classifier $\hat{h}$ on $\mathcal{X}_1$ and $\mathcal{W}_1$.

2. Fit a local linear smoother that estimates $H(Z_i) = \operatorname{logit}(\hat{h}(Z_i))$ for $Z_i \in \mathcal{X}_2 \cup \mathcal{W}_2$ using the model given in (38).

3. Consider the coefficients of the local linear smoother $\hat{\beta}(Z_i)$ for $Z_i \in \mathcal{X}_2 \cup \mathcal{W}_2$. They provide an estimate of the gradient $\nabla_z H$ at Z_i, that is, $\widehat{\nabla_z H(Z_i)} = \hat{\beta}(Z_i)$ for $Z_i \in \mathcal{X}_2 \cup \mathcal{W}_2$.

4. Standardize the estimated gradients $\hat{\beta}_j(Z_i)$ to $\hat{\beta}_j(Z_i)/\widehat{sd(\hat{\beta}_j(Z_i))}$ by using the estimated $\widehat{sd(\hat{\beta}_j(Z_i))}$ given by the local linear smoother.

5. Then, instead of C, we can use the estimate

$$\widehat{C} = \frac{1}{N} \sum_{Z_j \in \mathcal{X}_2 \cup \mathcal{W}_2} (\hat{\beta}(Z_j) - \overline{\hat{\beta}(Z)})(\hat{\beta}(Z_j) - \overline{\hat{\beta}(Z)})^T, \tag{41}$$

where $\overline{\hat{\beta}(Z)} = \sum_{Z_j \in \mathcal{X}_2 \cup \mathcal{W}_2} \hat{\beta}(Z_j)/N$ and $N = m_2 + n_2$.

6. Find the eigenvalue decomposition of $\widehat{C}$ as

$$\widehat{C} = \widehat{M}\widehat{\Lambda}\widehat{M}^T,$$

which gives the estimates $\widehat{M}$ and $\widehat{\Lambda}$ of M and Λ, as defined in (40), respectively. We find the eigenvectors by performing PCA on the standardized gradients.

7. Then $\overline{\hat{\beta}(Z)}$ and the estimated eigenvectors $\hat{M}_{\cdot 1}, \hat{M}_{\cdot 2}, \ldots$,given by the columns of $\widehat{M}$, best capture the slope and variations in $H(\cdot)$ and hence the classifier surface, respectively.

We can additionally construct bootstrapped uncertainty intervals by repeatedly drawing with replacement m_b background samples from $\mathcal{X}$ and n experimental samples from $\mathcal{W}$ and split them into $\mathcal{X}_1^*$, $\mathcal{X}_2^*$, $\mathcal{W}_1^*$ and $\mathcal{W}_2^*$ of sizes m_1, m_2, n_1 and n_2 at random, ensuring no overlap between the training and test data sets. That is, $\mathcal{X}_1^* \cap \mathcal{X}_2^* = \phi$ and $\mathcal{W}_1^* \cap \mathcal{W}_2^* = \phi$. In each case we first retrain the classifier on the new training data $\mathcal{X}_1^* \cup \mathcal{W}_1^*$ and then find $\overline{\hat{\beta}(Z)}$ and the estimated eigen values $\hat{M}_{\cdot 1}, \hat{M}_{\cdot 2}, \ldots$ on the new test data $\mathcal{X}_2^* \cup \mathcal{W}_2^*$. Here m_1, m_2, n_1, and n_2 are the same as in Method 5.1. We can then use the standard deviation or quantiles of the bootstrapped estimates to create bootstrap confidence intervals that will give estimates of the stability of the estimators, given by the algorithm, to perturbations in the data.

6. Experiments: Search for the Higgs boson. We demonstrate the performance of the proposed semisupervised classifier tests on the Higgs boson machine learning challenge data set available on the CERN Open Data Portal at http://opendata.cern.ch/record/328 (ATLAS Collaboration (2014)). The data set consists of simulated data provided by the ATLAS experiment at CERN's Large Hadron Collider to optimize the search for the Higgs boson.

Our goal is to demonstrate the performance of the proposed tests in identifying the presence of the Higgs boson particle, without assuming an a priori ansatz of the signal, and demonstrate their applicability to model-independent searches of new physics signals in experimental particle physics.

6.1. *Data description*. The Higgs boson has many different ways through which it can decay in an experiment and produce other particles. This particular challenge, from which our data set originates, focuses on the collision events where the Higgs boson decays into two tau particles (Adam-Bourdarios et al. (2015)). The data provided for the challenge consist of collision events labelled as background and signal. The signal class is comprised of events in which a Higgs boson (with a fixed mass of 125 GeV) was produced and then decayed into two taus. The events are simulated using the official ATLAS full detector simulator. The simulator yields simulated events with properties that mimic the statistical properties of the real events of the signal type as well as several important backgrounds. For the sake of simplicity, background events generated from only three different background processes were included in the challenge data set. Our objective is to show that semisupervised classifier tests are able to identify the Higgs signal without any prior knowledge.

The data set has 818,238 observations, where each observation is a simulated proton-proton collision event in the detector. Each of these collision events produces clustered showers of hadrons which originate from a quark or a gluon produced during the collision. These showers are called *jets*. The data contain information on the measured properties of the jets as well as the other particles produced during the collision. There are $d = 35$ features whose individual details can be found on CERN's Open Data Portal or in Appendix B of Adam-Bourdarios et al. (2015). Here we give some insight into the most important characteristics of the features.

The features whose names start with `PRI` are primitive variables that record the raw quantities, as measured by the detector. The features, whose names start with `DER`, are derived variables which are evaluated as functions of the primitive variables. Since all collision events do not produce the same number of jets, the number of jets produced in the collisions, denoted by `PRI_jet_num`, ranges from zero to three (events with more than $three$ jets are capped at $three$). Note that it is possible for the collisons to not produce any jets (`PRI_jet_num`$= 0$),

TABLE 2

Descriptions of the variables used in the analysis of the Higgs boson machine learning challenge data set (Adam-Bourdarios et al. (2015)). We drop the prefix `PRI` *from the variable names and further shorten some of the variable names intuitively for convenience*

Variable	Description
`tau_pt`	The transverse momentum of the hadronic tau.
`tau_eta`	The pseudorapidity η of the hadronic tau.
`tau_phi`	The azimuth angle ϕ of the hadronic tau.
`lep_pt`	The transverse momentum of the lepton (electron or muon).
`lep_eta`	The pseudorapidity η of the lepton.
`lep_phi`	The azimuth angle ϕ of the lepton.
`met`	The missing transverse energy.
`met_phi`	The azimuth angle ϕ of the missing transverse energy.
`met_sumet`	The total transverse energy in the detector.
`lead_pt`	The transverse momentum of the leading jet, that is, the jet with the largest transverse momentum (undefined if `PRI_jet_num` $= 0$).
`lead_eta`	The pseudorapidity η of the leading jet (undefined if `PRI_jet_num` $= 0$).
`lead_phi`	The azimuth angle ϕ of the leading jet (undefined if `PRI_jet_num` $= 0$).
`sublead_pt`	The transverse momentum of the subleading jet, that is, the jet with the second largest transverse momentum (undefined if `PRI_jet_num` ≤ 1).
`sublead_eta`	The pseudorapidity η of the subleading jet (undefined if `PRI_jet_num` ≤ 1).
`sublead_phi`	The azimuth angle ϕ of the subleading jet (undefined if `PRI_jet_num` ≤ 1).
`all_pt`	The scalar sum of the transverse momentum of all the jets in the event.
Weight	The event weight.
Label	The event label (string) (s for signal, b for background).

and hence there are structurally absent missing values in the data that relate to the jets produced in the collisions. To avoid these missing values in the data, we only consider events that have two jets (`PRI_jet_num`$= 2$). This results in 165,027 events, 80,806 background events, and 84,221 signal events. Since the derived quantities are functions of the primitives, we use just the primitive variables ($d = 16$) for our analysis. Since we only use the primitive variables, we drop the prefix `PRI` from the variable names and further shorten some of the variable names intuitively for convenience. Descriptions of the primitive variables used are provided in Table 2.

Among the primitive features, five of them provide the azimuth angle ϕ of the particles generated in the event (variables ending with `_phi`). These features are rotation invariant in the sense that the event does not change if all of them are rotated together by the same angle. Hence, to interpret these variables more easily using the active subspace method, we remove the invariance of the azimuth angle variables by rotating all the ϕ's so that the azimuth angle of the leading jet at 0 (`lead_phi`$= 0$). `lead_phi` can then be removed from the analysis, leading to a 15-dimensional feature space.

Additionally, we take logarithmic transformations of the variables that give the transverse momentum of the particles produced (variables ending with `_pt`), the missing transverse energy (`met`), and the total transverse energy in the detector (`met_sumet`). Exploratory analysis of the data as well as details and justifications for the transformations considered above can be found in the Supplementary Material (Chakravarti et al. (2023)).

Experimental setting: In all the following experiments on the Higgs boson data set, we randomly sample without replacement background and signal events from the original data set to form background data with $m_b = 40{,}403$ events, signal data with $m_s = 20{,}403$ events, and experimental data with $n = 40{,}403$ events. In this manner we generate 50 replicates of each of these data sets for use in our power studies.

Recall that the experimental data is a mixture of background and signal data, where the proportion of signal data is given by λ, the signal strength. So the number of signal events in the experimental data is decided according to λ by randomly generating from $\text{Bin}(n, \lambda)$. For methods that require data-splitting, the data-splitting is done by randomly splitting the background data into $m_1 = 20{,}403$ training and $m_2 = 20{,}000$ test background samples and by randomly splitting the experimental data into $n_1 = 20{,}403$ training and $n_2 = 20{,}000$ test experimental samples. We construct the splits so the number of signal events in the training and test experimental samples are randomly generated from $\text{Bin}(n_1, \lambda)$ and $\text{Bin}(n_2, \lambda)$, respectively. Additionally, when randomly generating the experimental data, events are selected in a weighted fashion using the Weight variable provided in the data set, as described in Table 2. Note that the classifiers for the model-dependent approaches are trained on $m_1 = 20{,}403$ background and $m_s = 20{,}403$ signal training samples, and the classifiers for the model-independent approaches are trained on $m_1 = 20{,}403$ background and $n_1 = 20{,}403$ experimental training samples. Hence, all the classifiers are trained on balanced datasets. For each of the bootstrap and permutation methods, we consider 1000 bootstrap and permutation cycles respectively.

In the following sections, we first explore the power of the classifier tests, described in Section 3, to detect the presence of the Higgs boson signal in the experimental data. We then estimate the signal strength (λ) using the methods introduced in Section 4 and then use the active subspace method introduced in Section 5 to characterize the signal. All the code used for this section is available at https://github.com/purvashac/MIDetectionClassifierTests.

6.2. *Anomaly detection using the classifier tests.* We compare the power of the model-dependent supervised methods and the model-independent semisupervised methods, introduced in Section 3 in detecting the Higgs boson signal, by varying the signal strength from $\lambda = 0.15$ to $\lambda = 0.01$. We also check that the tests have the right error control under the null case ($\lambda = 0$). We demonstrate the performance of the different methods—asymptotic, bootstrap, permutation (using out-of-sample statistics), and slow permutation (using in-sample statistics), used along with the different test statistics—Likelihood Ratio Test statistic (LRT), Score statistic, Area Under the Curve (AUC), and Misclassification Error (MCE).

For the model-dependent methods, since we have only finitely many samples from the background available, and not the background generator itself, we split the available background data into training and test sets, as described above. We then use the training set for fitting the classifier and the test set for estimating the null distribution of the test statistic by bootstrapping or permuting as described in Section 3. In real life, since the background MC generator is known, we should, in many cases, be able to generate more training background samples for estimating the null distribution of the test statistic.

6.2.1. *Anomaly detection when data has a correctly specified signal.* We first compare the methods when the signal model is correctly specified. The tests are run on 50 random samplings of the data, and the percentage of times each of the tests rejects the null that there is no signal is given in Table 3. "Permutation" indicates the faster permutation method from Section 3.2, that uses out-of sample test statistics for testing, and "Slow Perm" indicates the slower permutation method from Section 3.2 that uses in-sample test statistics for testing and retrains the classifier in every permutation cycle. The significance level for all the tests is considered at $\alpha = 0.05$.

As seen in Table 3, among the supervised methods the permutation tests outperform all the other supervised methods in terms of power. The permutation test with the score statistic has the most power for smaller values of λ. However, it is worth noting that it might be anti-conservative for $\lambda = 0$ and hence might not have the right significance level. Conversely, as

TABLE 3

Power of detecting the signal in the Higgs boson data for each method in percentages. We consider 50 *random samplings of the data and* 1000 *bootstrap and permutation cycles. We perform each test at* $\alpha = 5\%$ *significance level. The last column* ($\lambda = 0$) *represents the Type I error of the methods*

		Signal Strength (λ)						
Model	Method	0.15	0.1	0.07	0.05	0.03	0.01	0
Supervised LRT	Asymptotic	100	100	96	62	18	18	6
	Bootstrap	100	96	78	58	6	0	0
	Permutation	100	98	98	86	28	6	0
Supervised Score	Bootstrap	64	66	74	50	18	0	0
	Permutation	94	92	100	92	80	24	12
Semisupervised LRT	Asymptotic	100	98	74	38	16	6	2
	Bootstrap	100	98	48	10	2	2	0
	Permutation	100	98	72	38	16	6	2
	Slow Perm	82	8	0	4	2	0	4
Semisupervised AUC	Asymptotic	100	96	78	32	14	4	2
	Bootstrap	100	98	70	32	20	6	2
	Permutation	100	98	68	32	20	4	2
	Slow Perm	100	100	94	56	20	8	4
Semisupervised MCE	Asymptotic	100	92	60	28	14	2	2
	Bootstrap	100	96	52	28	16	6	4
	Permutation	100	96	52	30	14	6	6
	Slow Perm	100	98	86	58	16	6	2

Table 3 and Figure 4 show, the supervised tests that use the LRT statistic appear to be overly conservative for smaller values of λ, because the LRT statistic uses an estimate of the density ratio p_s/p_b using the classifier. This problem mainly appears when two conditions are both in effect: (1) the classifier for background vs. signal overfits the class probabilities, by overfitting to the data, and (2) when $\lambda = 0$, that is, when the experimental data has no signal. Under this situation and with probability higher than expected, almost the entire experimental data is classified as background with $\hat{h}(Z_i) = 0$, and $\widehat{p_s/p_b} = 0$. This makes $\hat{\lambda}_{\mathrm{MLE}}$ the maximum likelihood estimator of λ, also zero, making the LRT statistic zero as well. This is also observed when we use the permutation or bootstrap methods to estimate the null distribution. The estimated distributions have a point mass at zero that is much higher than 0.5 (as given by the asymptotic distribution in equation (15) when true p_s/p_b is used). Note that we additionally leave out the asymptotic score test both in Table 3 and Figure 4 since we tried the test for a subsample of the data and decided to not consider that test for the larger data set due to the poor quality of the asymptotic approximation.

Among the semisupervised approaches, the AUC and MCE slow permutation methods (slow perm) have power that is only slightly worse than that of the supervised methods, as shown in Table 3. Importantly, the semisupervised methods achieve this without having access to the labeled signal sample during training. Among the asymptotic and the faster permutation methods, using LRT or MCE gives similar performance to AUC in the semisupervised approaches. We also observe that the slow permutation method using the LRT statistic has very low power. This is due to bias in estimating the LRT using an in-sample estimate which is influenced by over-fitting.

Figure 5 shows the empirical distributions of the p-values produced by the semisupervised tests for different λ values. All the tests appear to have correct (or at least almost correct) type I error control, as indicated by the close to uniform CDFs in the $\lambda = 0$ case. We also notice that none of the tests have much power to detect signals that are less than 3% of the

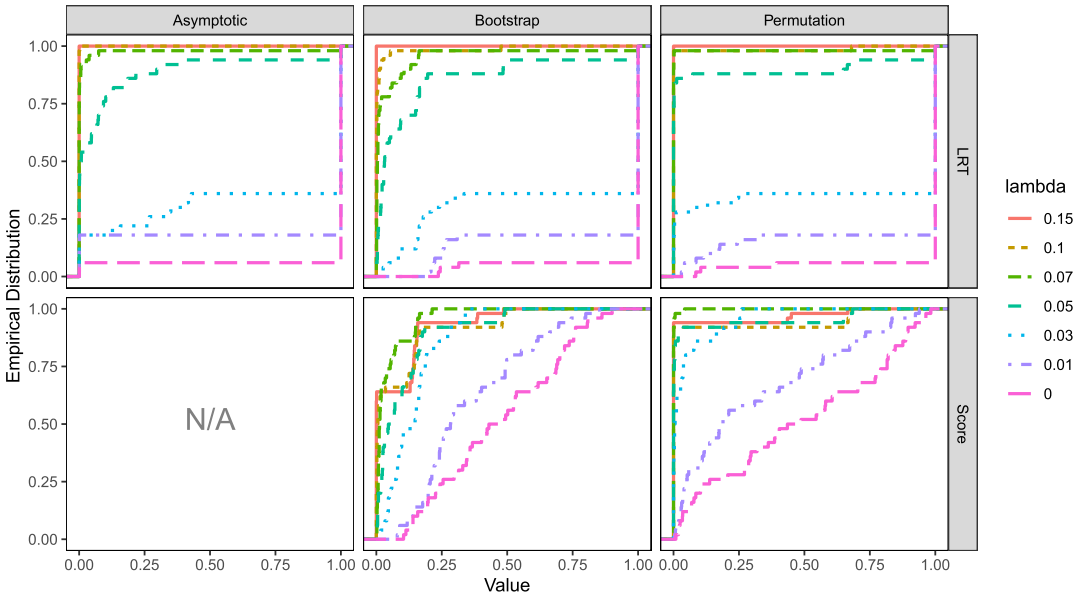


FIG. 4. *Empirical CDF of the p-values given by the supervised tests for different signal strengths* (λ). *The columns in the grid of plots represent the different methods of testing* (*asymptotic*, *bootstrap*, *and permutation*), *and the rows represent the use of the different test statistics* (*LRT and Score*). *Note that the plot for the asymptotic test using the score statistic is missing since we do not consider that test due to the poor quality of the asymptotic approximation*.

experimental data ($\lambda < 0.03$). We omit the plot for the slow permutation method using the LRT, as it has anomalously low power for most λ values due to the reasons mentioned above.

We additionally compared these methods to nearest neighbor (NN) two-sample tests, as introduced in Schilling (1986) and Henze (1988). We considered the nearest neighbor tests, as opposed to other tests, since the signal that we are trying to detect appears as a collection of nearby data points. Hence, tests based on neighborhoods should have better power to detect it. We compared the methods presented in this paper to the NN tests (asymptotic and permutation versions) for a subsample of the data set. We observed that the asymptotic version especially had very poor power. The permutation version performed better but was outperformed by the semisupervised AUC, MCE and LRT methods. Additionally, it was not scalable to extend it to the larger data set. Hence, we concluded that it was computationally impractical to apply it to the larger full data set.

In conclusion, we see that slow permutation method when using the AUC and MCE statistics outperforms the other semisupervised methods and additionally gives comparable performance to the supervised methods in detecting the signal in the experimental data. Note that the power of the slow permutation method using the AUC and MCE statistics is much better than the one using the LRT statistic which is the statistic used by D'Agnolo and Wulzer (2019) and D'Agnolo et al. (2021). So these methods give an improvement over using just the LRT statistic.

6.2.2. *Anomaly detection when data has a misspecified signal*. As mentioned in the Introduction, a model-dependent search that targets one kind of new physics signal will not be powerful to detect a different signal which might actually be present in the data. This is one of the main motivations for model-independent methods. In this section we demonstrate that if the signal model is misspecified in just one dimension, model-independent methods are still able to detect the signal, whereas the model-dependent methods fail to detect the signal.

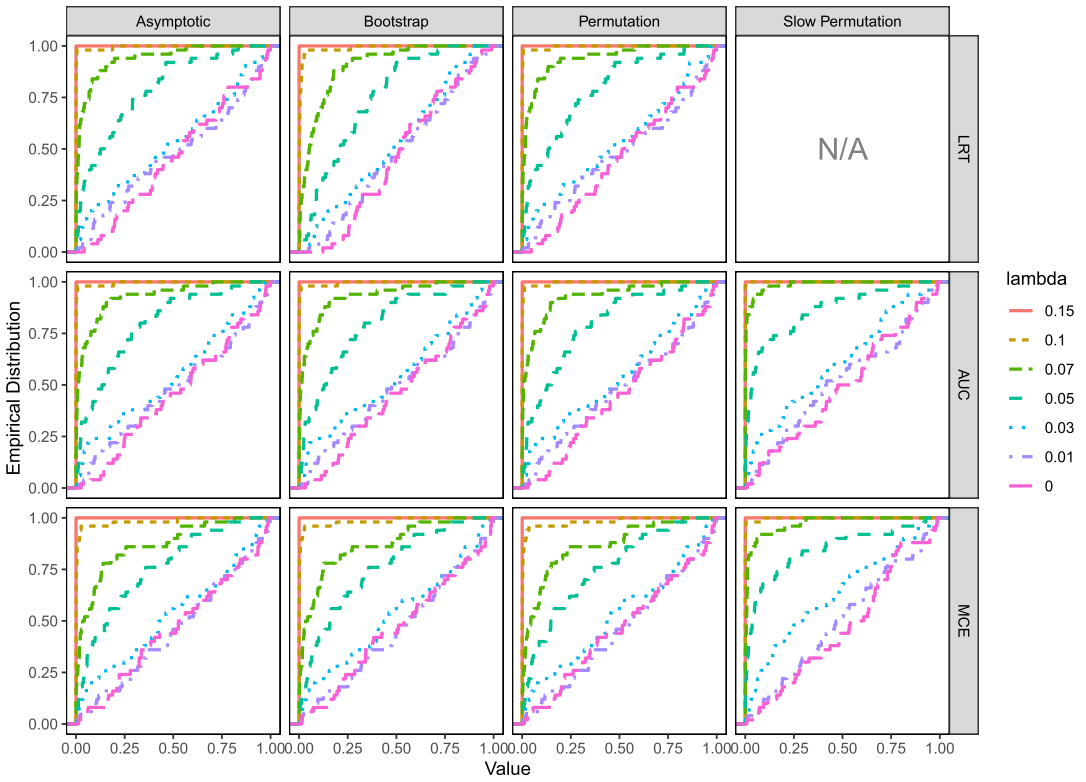


FIG. 5. *Empirical CDF of the p-values, given by the semisupervised tests, for different signal strengths* (λ). *The columns in the grid of plots represent the different methods of testing (asymptotic, bootstrap, permutation, and slow permutation), and the rows represent the use of the different test statistics (LRT, AUC, and MCE). We leave out the plot for slow permutation test using the LRT statistic since, in practice, it demonstrates very poor performance. It suffers from bias caused by overfitting since for slow permutation we consider the in-sample LRT.*

We transform the signal in the experimental data to intentionally make it different from the signal model which makes the signal model misspecified. We consider a transformation in just one variable, transforming

$$\texttt{tau\_pt}^* = \texttt{tau\_pt} - 0.7\big(\texttt{tau\_pt} - \min(\texttt{tau\_pt})\big) \tag{42}$$

in the experimental data. Meanwhile, we do not transform the signal data used by the supervised methods for training the classifier. This simulates a misspecified signal situation, where the signal in the experimental data is not from the same distribution as the signal used by the supervised methods for training. This is to emulate a situation where the signal in the experimental data is different from the signal that the signal model specifies in just one dimension.

We chose this particular transformation for multiple reasons. First, it transforms the variable that has the most marginally different means for the background and the signal data, as demonstrated by Figure 4 in the Supplementary Material (Chakravarti et al. (2023)). Second, the marginal distribution of the signal along this variable is different from the marginal distribution of the background. These two reasons make `tau_pt` a variable that potentially influences the classifier. By transforming the variable, we lower the mean `tau_pt` for the signal in the experimental data, causing the model dependent methods to lose power in detecting the transformed signal.

We now compare the power of the supervised and the semisupervised methods in detecting the transformed signal in the experimental data. The tests are run on 50 random samplings of

TABLE 4

Power of detecting the misspecified signal in the Higgs boson data for each model in percentages. We consider 50 *random samplings of the data and* 1000 *bootstrap and permutation cycles. We perform each test at* $\alpha = 5\%$ *significance level. The last column* ($\lambda = 0$) *represents the Type I error of the methods*

Model	Method	Signal Strength (λ)						
		0.15	0.1	0.07	0.05	0.03	0.01	0
Supervised LRT	Asymptotic	2	10	2	8	8	6	4
	Bootstrap	0	0	0	0	0	0	0
	Permutation	0	0	0	0	0	2	0
Supervised Score	Bootstrap	0	0	0	0	0	0	0
	Permutation	0	0	0	0	0	2	8
Semisupervised LRT	Asymptotic	100	100	100	82	18	4	4
	Bootstrap	100	100	100	60	4	2	0
	Permutation	100	100	100	82	18	4	2
	Slow Perm	100	100	78	22	2	4	6
Semisupervised AUC	Asymptotic	100	100	100	78	16	8	4
	Bootstrap	100	100	100	82	20	10	0
	Permutation	100	100	100	80	20	8	2
	Slow Perm	100	100	100	100	34	10	4
Semisupervised MCE	Asymptotic	100	100	100	66	24	6	4
	Bootstrap	100	100	100	62	16	6	4
	Permutation	100	100	100	62	14	6	4
	Slow Perm	100	100	100	98	22	8	2

the data, and the percentage of times each of the tests rejects the null that there is no signal, is given in Table 4. As before, "Permutation" indicates the faster permutation method from Section 3.2 that uses out-of-sample test statistics for testing, and "Slow Perm" indicates the slower permutation method from Section 3.2, that uses in-sample test statistics for testing and retrains the classifier in every permutation cycle. The significance level for all the tests is considered at $\alpha = 0.05$.

As seen in Table 4, as expected, the supervised methods have no power at all in detecting the transformed signal. An interesting observation from the table is that it appears that the supervised methods are more conservative when there is some signal present compared to when there is no signal present. This occurs as the signal labels that the supervised models are trained on are inconsistent with the signal in the experimental data, when the signal is present.

The semisupervised methods, on the other hand, still have power to detect the transformed signal since they are not trained on the misspecified signal training data. In fact, the semisupervised methods appear to have higher power with the transformed signal than with the original signal, indicating that the transformed signal is somewhat easier to disentangle from the background than the original signal. Comparing the semisupervised methods, we again observe that the slow permutation method, using AUC, has the highest power. We also notice that the asymptotic and permutation tests that use LRT demonstrate power comparable to the AUC. The slow permutation method with MCE also performs comparably to the slow permutation method with AUC. The slow permutation method with LRT has again anomalously low power for smaller λ, due to the bias caused by using the in-sample LRT as the test statistic.

In conclusion, semisupervised methods continue to demonstrate power similar to the previous case where the signal model was not misspecified. On the other hand, the supervised

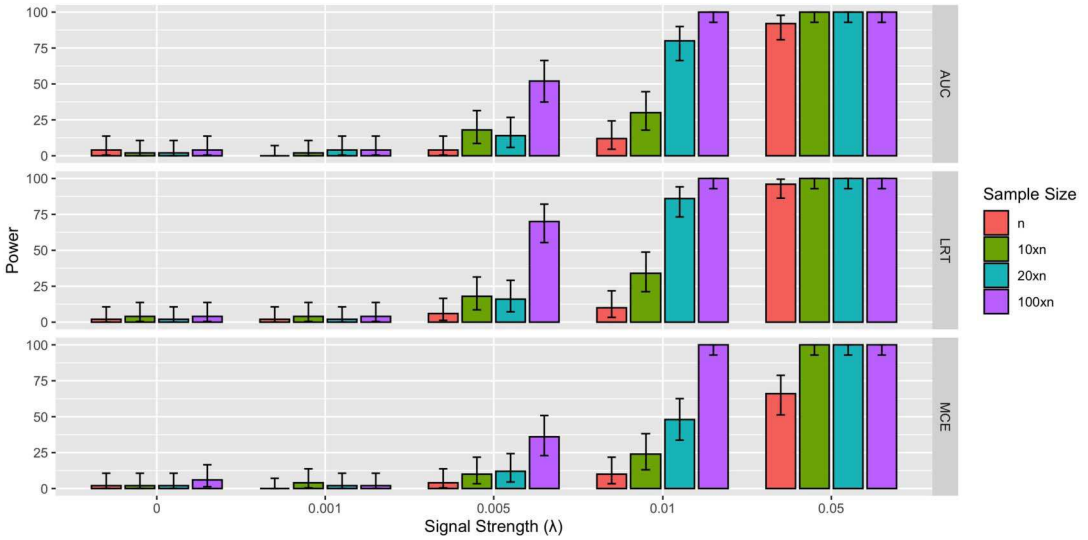


FIG. 6. *Power of the asymptotic model-independent tests using AUC, LRT, and MCE statistics to detect different signal strengths* ($\lambda = 0, 0.001, 0.005, 0.01, 0.05$) *in* 50 *simulations with samples of sizes* n, $10n$, $20n$, *and* $100n$, *where* $n = 2 \times 10^4$. $\lambda = 0$ *represents the Type I error rate. The error bars represent the* 95% *Clopper–Pearson confidence intervals for the power. We note that, for all of the three statistics, the power increases in larger sized samples for* $\lambda = 0.005, 0.01, 0.05$. *For* $\lambda = 0.001$, *none of the tests appear to have power to detect the signal, but further increases in the sample size should provide power in this case as well.*

models do not have any power to detect the transformed signal in the misspecified model case which motivates the use of semisupervised models in such a case.

6.2.3. *Anomaly detection with smaller signal strengths.* So far, the signal strengths that we have considered in this paper have been in the order of 10^{-2}, and the sample sizes have been in the order of 10^4. But in many high-energy physics searches, the signal strengths that the physicists are looking for are in the order of $10^{-3} - 10^{-4}$, and they have much larger sample sizes as well. For example, the true signal strength of the Higgs boson signal in the data set where the two jets are produced is $\lambda = 0.0035$. In this section we show that the power of the model-independent methods to detect signal for smaller signal strengths increases as the sample size increases.

We perform the asymptotic model-independent tests on equally sized background and experimental data sets whose sample sizes vary from 2×10^4 to 2×10^6, with signal strengths λ varying from 0.001–0.05. To make this experiment feasible (due to the lack of availability of sufficient background samples in our original data set), we fit a mixture of eight multivariate Normals to the background data and a mixture of nine multivariate Normals to the signal data in the Higgs boson data set. We then generate samples from the estimated mixtures of Normals to create our simulated data sets for this experiment.

Figure 6 shows that for smaller signal strengths, for example, $\lambda = 0.005$, the power of the model-independent tests increases as the sample size increases. We see that for the largest sample size, $100n = 2 \times 10^6$, the model-independent tests have good power, even for $\lambda = 0.005$, which is comparable to the true Higgs boson signal strength (0.0035) in the two jets channel. Whereas, for the sample size that we used in our previous experiments, $n = 2 \times 10^4$, the model-independent tests have small power for $\lambda = 0.01$ and no power for $\lambda = 0.005$. This demonstrates that when larger data sets are available, the model-independent methods have power to detect even smaller signal strengths. In many search channels, LHC data sets are of the order of millions or billions of events, so these tests should have power for detecting reasonably small signals in those channels.

6.3. *Estimating the Higgs boson signal strength.* In this section we demonstrate the performance of the methods proposed in Section 4 to estimate the signal strength (λ) using the Higgs boson data set. We vary the signal strength from $\lambda = 0.5$ to $\lambda = 0$ and look at the estimates of the signal strength as well as the bootstrapped uncertainty intervals given by Method 4.1 in Section 4.

We consider 100 bootstrap cycles to get the 95% bootstrapped uncertainty intervals. For the estimates we consider $T \in \{0.8, 0.5\}$ and $b \in \{0.01, 0.005\}$, where T is the threshold that indicates the neighborhood of 1 where we fit the Poisson regression model and b is the bin-width of the histogram that we use to estimate the densities. Note that bin-width $b \in \{0.01, 0.005\}$ is equivalent to the number of bins, Bins $\in \{100, 200\}$. To construct the bootstrapped uncertainty intervals, we use three different methods, using $\alpha = 5\%$, and compare them in Figure 7 below.

First, we use the basic bootstrap confidence interval, also known as the reverse percentile interval, which uses empirical quantiles of the bootstrap distribution of $\hat{\lambda}$ in a reverse order to construct the confidence interval (see Davison and Hinkley (1997), equation (5.6), p. 194): $(2\widehat{\lambda} - \lambda^*_{(1-\alpha/2)}, 2\widehat{\lambda} - \lambda^*_{(\alpha/2)})$, where $\lambda^*_{(1-\alpha/2)}$ denotes the $1-\alpha/2$ quantile of the bootstrapped estimates λ^*.

Second, we use the bootstrapped quantiles to form percentile bootstrap confidence intervals (see Davison and Hinkley (1997), equation (5.18), p. 203, Efron and Tibshirani (1993), equation (13.5), p. 171): $(\lambda^*_{(\alpha/2)}, \lambda^*_{(1-\alpha/2)})$.

Third, we use bootstrapped standard errors and normal distribution's quantiles to create bootstrap confidence intervals: $(\widehat{\lambda} - z_{1-\alpha/2}\widehat{\mathrm{se}}_{\lambda^*}, \widehat{\lambda} + z_{1-\alpha/2}\widehat{\mathrm{se}}_{\lambda^*})$, where $z_{1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of the standard normal distribution, and $\widehat{\mathrm{se}}_{\lambda^*}$ is the standard error estimated using the empirical standard deviation of the bootstrapped estimates λ^*.

We additionally present the confidence intervals given by the GLM model (Dobson and Barnett (2018)) itself in Figure 7.

We observe from Figure 7 that the estimates of λ do not vary much with the number of bins used for the histogram estimate of the density. We also notice that thresholding at $T = 0.8$, which is closer to 1, as compared to $T = 0.5$, gives better estimates of λ. We additionally tried $T = 0.9$ on a subsample of the data set. But the estimates in that case were not very stable, and hence we decided to not consider them for the final larger data set. The uncertainty intervals created using bootstrapped quantiles and bootstrapped standard errors appear to be relatively well calibrated when $T = 0.8$, that is, they include the true value of λ, given by the dotted diagonal line, in each of the plots. The uncertainty intervals, given by the GLM model and basic bootstrap, on the other hand, do not contain the true λ for some smaller values of λ. So $\hat{\lambda}$, using $T = 0.8$ with the number of bins either 100 or 200, accompanied with either bootstrapped quantile or bootstrapped standard error uncertainty intervals, appears to give the best performance in estimating and quantifying the uncertainty of the signal strength λ. An additional advantage of the uncertainty intervals using bootstrapped standard errors is that they are centered about the estimate $\hat{\lambda}$.

6.4. *Interpreting the classifier using active subspace methods.* We demonstrate the application of the active subspace methods for a random simulation (one of the 50 simulations in Section 6.2), which detects the signal at significance level $\alpha = 0.05$, when $\lambda = 0.15$, that is, 15% of the experimental data is from the signal sample. We consider $\lambda = 0.15$, since the random forests demonstrate good power in detecting the signal for that signal strength (Table 3).

We then use Method 5.1, presented in Section 5, to find the active subspace of the fitted semisupervised classifier. The second step of the algorithm requires us to choose a linear smoother as well as a smoothing parameter for it. We choose a Gaussian kernel smoother

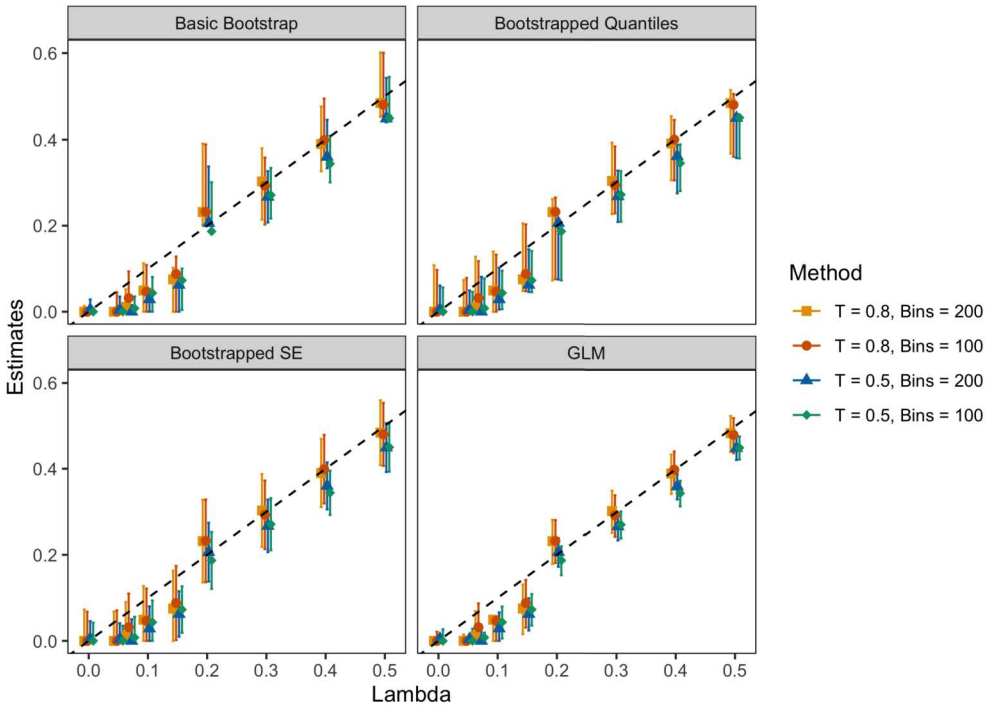


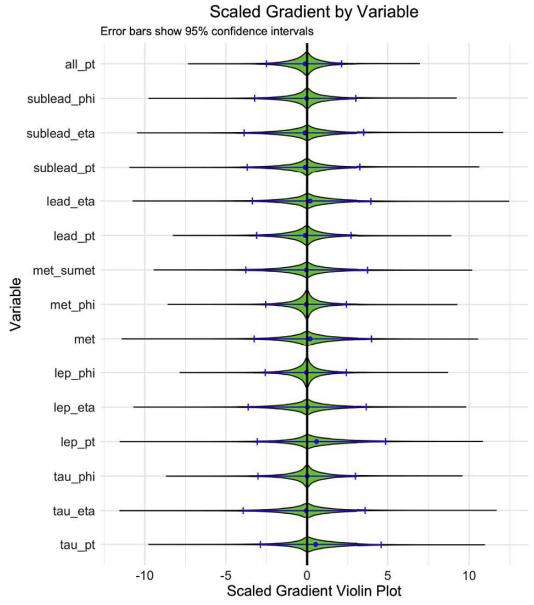
FIG. 7. *Estimates of the signal strength* (λ), *along with the* 95% *uncertainty intervals, of the Higgs boson in the experimental data. The true* λ, *given by the dotted diagonal line, varies from* 0 *to* 0.5. *T specifies the threshold that controls the size of the neighborhood of* 1 *that is considered in the Poisson regression model, and Bins specifies the number of bins considered in the histogram when estimating the density of* $\widehat{g_q}(1)$, *as described in Section* 4.

as the linear smoother and $h = 0.5$ as the smoothing parameter. The smoothing parameter h is used to scale the standard deviations of the variables, which is then used as the standard deviation of the multivariate Gaussian kernel, that is, $\hat{sd}(Z)/h$ is considered as the standard deviation of the multivariate Gaussian kernel. The smoothing parameter selection process is described in the Supplementary Material (Chakravarti et al. (2023)).

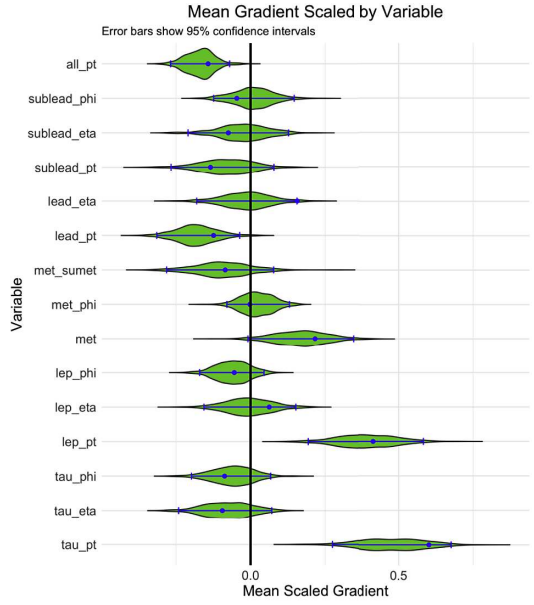
Figure 8 gives us the active variables given by the mean standardized gradient, the first eigenvector, and the second eigenvector. The higher eigenvectors do not contain much information and have been included in the Supplementary Material (Chakravarti et al. (2023)). The figure additionally gives the distribution of the bootstrapped active subspace estimates as violin plots. These help us construct 95% uncertainty intervals for the active subspace variables using the bootstrapped empirical percentiles. We consider 500 bootstrap cycles to get the 95% bootstrapped uncertainty intervals.

Figure 8(a) additionally provides the violin plots of the standardized gradient estimates $\hat{\beta}_j(Z_i)/\widehat{sd}(\hat{\beta}_j(Z_i))$ given by the local linear smoother at every point in the combined test data $Z_i \in \mathcal{X}_2 \cup \mathcal{W}_2$. We notice that the violin plots are not very informative since they are all symmetric about zero; that is, they don't appear to give any information about the slope of the classifier surface.

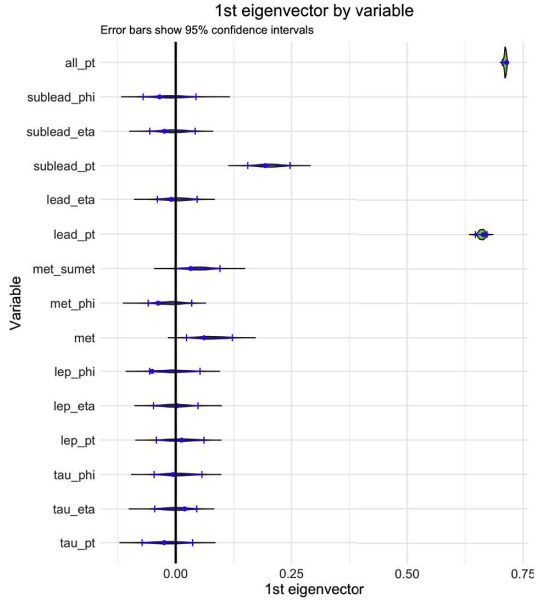
The mean standardized gradient in Figure 8(b) gives the direction in which the classifier output changes most rapidly on average. We see that jointly higher transverse momentums of the hadronic tau (`tau_pt`) and the remaining lepton (`lep_pt`) contribute the most to an increase in the classifier output. Increase in these transverse momentums, combined with a decrease in the transverse momentum of the leading jet (`lead_pt`) and the scalar sum of the transverse momentum of all the jets (`all_pt`), leads to an increase in the classifier output. This implies that the detected signal events display higher momentums of the hadronic tau and



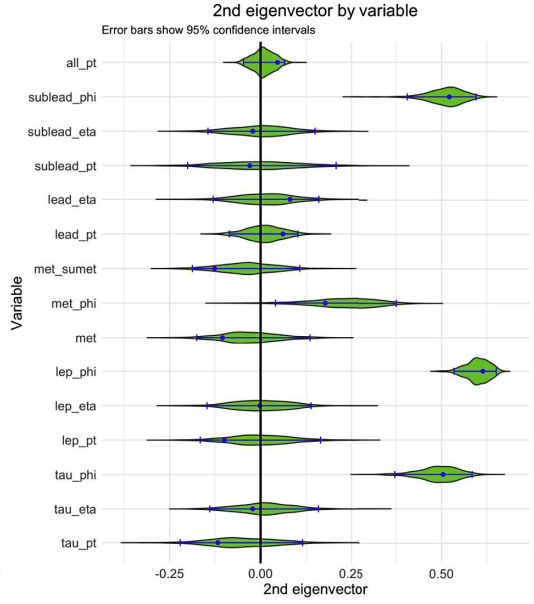
(a) Standardized Gradient Distribution



(b) Mean Standardized Gradient



(c) First Eigenvector



(d) Second Eigenvector

FIG. 8. *The active subspace variables for the classifier trained on data with signal strength* $\lambda = 0.15$ *computed using a local linear smoother that uses the gaussian kernel with smoothing parameter* $h = 0.5$. (*a*) *gives the distribution of the standardized gradients* $\hat{\beta}_j(Z_i)/\widehat{sd(\hat{\beta}_j(Z_i))}$, *given by the local linear smoother at every point in the combined test data* $Z_i \in \mathcal{X}_2 \cup \mathcal{W}_2$. *The dots denote the mean. In* (*b*), (*c*), *and* (*d*), *the violin plot and the dashes give the bootstrapped empirical distribution and the bootstrapped uncertainty intervals computed using the empirical quantiles, respectively, for the standardized mean, the first eigenvector, and the second eigenvector. In* (*b*) *the dots give the mean standardized gradient similar to* (*a*). *In* (*c*) *and* (*d*), *the dots represent the first eigenvector and the second eigenvector computed on the combined test data, respectively.*

the other lepton, lower momentums of the leading jet and lower scalar sum of the momentums of all the jets, as compared to background events. Note that the mean gradient vector not only captures the dependency of the classifier on each variable individually but also characterizes the multivariable dependencies that influence the classifier.

The first eigenvector gives the first principal component of the gradients which demonstrates the relationship between the variables that causes the most variability in the gradients of the classifier. The first eigenvector, as seen in Figure 8(c), indicates that when `all_pt`, the scalar sum of the transverse momentums, `lead_pt`, the transverse momentum of the leading jet, `sublead_pt`, the transverse momentum of the subleading jet, and `met`, the missing transverse energy, jointly change in the same direction, it causes the most variation in the classifier. This means that these variables together, help the classifier separate the signal from the background. The second eigenvector, that is, the second principal component of the gradients, as seen in Figure 8(d), appears to capture the relationship between the azimuth angle ϕ of the different objects in the event that most influences the classifier. Recall that the ϕ values of the objects have been transformed to denote the difference in the angle between each object and the leading jet. So the second eigenvector indicates that the ϕ angle between the leading jet and all the other four objects (the subleading jet, the hadronic tau, the lepton, and the missing transverse energy) influences the classifier. This has an appealing physical interpretation in that it indicates that the azimuthal orientation between the leading jet and the rest of the event is a useful feature for separating the signal from the background.

REMARK. Note that, for any eigenvector $M_{\cdot j}$ of the standardized gradients of the trained classifier, $-M_{\cdot j}$ is also an eigenvector. This causes the violin plots of the bootstrapped eigenvectors to be systematically symmetric about zero. To solve this problem, we first take the variable that has the largest absolute eigenvector value, that is, $k_j = \arg\max_i |\hat{M}_{ij}|$. Then we fix the sign of the bootstrapped eigenvectors $\hat{M}^*_{\cdot j}$ such that the sign of the k_jth variable matches, that is, $\operatorname{sgn}(\hat{M}^*_{k_j j}) = \operatorname{sgn}(\hat{M}_{k_j j})$. This process has been followed in Figures 8(c) and 8(d) to handle the problem.

So the active subspace methods provide an algorithm to interpret the semisupervised classifier once it has detected a signal in the experimental data. The methods imply that the classifier that detects the Higgs boson is positively influenced by `tau_pt` and `lep_pt` and negatively influenced by `all_pt` and `lead_pt`. Additionally, it is also influenced by joint changes in the values of `all_pt`, `lead_pt`, `lsubead_pt`, and `lmet`. The classifier is also influenced by the difference between the azimuth angle ϕ of the leading jet and those of the other four objects, indicating that this might be an important feature for detecting the signal events. Finally, we note that, without techniques like these, it would have been very difficult to understand what kind of a signal the semisupervised classifier has detected within the high-dimensional feature space.

7. Conclusion. In this paper we studied model-independent anomaly detection tests, using semisupervised classifiers, that can detect the presence of signal events hidden within background events in high energy particle physics data sets. Additionally, we proposed methods to estimate the signal strength and to identify the active subspace affecting the classifier most strongly, leading to an understanding of the detected signal. We demonstrated the performance of the methods and also compared the proposed tests with comparable model-dependent supervised methods as well as nearest neighbor two-sample tests on a data set related to the search for the Higgs boson at the Large Hadron Collider.

We presented multiple model-independent methods that search for the signal without assuming any knowledge of the signal model. By not assuming any signal model, we retain

the ability to detect unknown and unexpected signals. This is an important capability for future searches at the Large Hadron Collider, where model-dependent searches have so far not yielded evidence of physics beyond the Standard Model. We used a semisupervised classifier to distinguish the experimental data from the background data and used the performance of the classifier to perform a test to detect a significant difference between the two data sets. We proposed three test statistics that can be used for the test: the likelihood ratio test (LRT) statistic, the area under the curve (AUC) statistic, and the misclassification error (MCE) statistic.

We compared the use of the different test statistics as well as the use of different techniques for obtaining the null distribution in building the test that has the most power in detecting the signal. We compared the power of the methods to detect the Higgs boson at different signal strengths and showed that a version of the proposed AUC method has power that is competitive with the model-dependent methods. So even when the signal model is correctly assumed by the model-dependent methods, the proposed model-independent methods appear to still have competitive power to detect the presence of the signal. However, when the signal model is incorrectly assumed or misspecified, the model-dependent methods can totally miss the signal, whereas the proposed model-independent ones are still able to detect the signal, as demonstrated in our experiments. In particular, the proposed methods demonstrate the ability to find new particles without specific a priori knowledge of their properties.

As described in Section 3.1, model-dependent methods currently in use in experimental high-energy physics, are slightly different from the ones presented in Section 3.1. In this paper to make the model-dependent methods comparable to the proposed model-independent ones, we consider high-dimensional model-dependent tests. In current practice though, a threshold is placed on the supervised classifier output to select a subset of the experimental data that is richer in signal events. Then the high-dimensional data set is transformed into a univariate one, and the transformed data are fitted using a one-dimensional mixture model consisting of signal and background components. This is used to construct a profile likelihood ratio test (ATLAS Collaboration and CMS Collaboration (2011)), where some of the nuisance parameters are related to the one-dimensional background model. Dauncey et al. (2015) proposed a discrete profiling method for this test, based on considering the choice of the background functional form as a discrete nuisance parameter which is profiled in an analogous way to continuous nuisance parameters.

These considerations motivate multiple avenues for future exploration. We could construct an analogous model-independent version of the approach described above, where we first use the semisupervised classifier output to select a signal-rich subset of the experimental data, then transform the selected data to a one-dimensional space, and finally perform a test in the one-dimensional space using a mixture model. For well-chosen transformations, this could increase the power of the test. More interestingly, for many signals the signal distribution corresponding to a transformation into the invariant mass variable is predicted by quantum mechanics to be the Cauchy distribution (also known as the Breit–Wigner distribution in particle physics; see Chapter 49 in Particle Data Group (2020)), whose parameters, under the model-independent scenario, are unknown to us. It would be interesting to build this knowledge into the semisupervised method. A straightforward approach would be to simply use the Cauchy distribution, convolved with a model for detector smearing, in the one-dimensional mixture model, but it might also be possible to incorporate this knowledge into the training of the high-dimensional semisupervised classifier.

Another important avenue of future work would be to find ways to account for systematic uncertainties in the background training data. As described above, the current supervised techniques account for systematics using profile likelihoods for one-dimensional summary statistics, with the systematic variations in the one-dimensional space parameterized using potentially hundreds or thousands of nuisance parameters. This should also be feasible for

one-dimensional semisupervised tests. However, it might also be possible to use profiling in the high-dimensional semisupervised likelihood ratio tests. For example, if we had two plausible background samples from two different MC generators, we could use optimal transport (Peyré, Cuturi et al. (2019)) to find the geodesic between the samples. We could then profile over the parameter corresponding to the location on the geodesic which would account for the systematic uncertainty corresponding to the envelope of background models spanned by the two MC generators.

Acknowledgments. The authors would like to thank the anonymous referees, the Associate Editor, and the Editor for their extensive, thoughtful and constructive comments that greatly improved the quality of this paper.

Funding. This work was supported in part by NSF awards PHY-2020295, DMS-2053804, and DMS-2113684.

SUPPLEMENTARY MATERIAL

Supplement to "Model-independent detection of new physics signals using interpretable semisupervised classifier tests" (DOI: 10.1214/22-AOAS1722SUPPA; .pdf). The supplementary material contains the proof of Theorem 4.1, some of the proposed algorithms from Section 3.2, and details about the exploratory data analysis of the Higgs boson data used in the experiments in Section 6. It additionally describes the selection of the smoothing parameter for the active subspace methods.

R code (DOI: 10.1214/22-AOAS1722SUPPB; .zip). R code files for the algorithms and the experiments in the paper.

REFERENCES

ADAM-BOURDARIOS, C., COWAN, G., GERMAIN, C., GUYON, I., KÉGL, B. and ROUSSEAU, D. (2015). The Higgs boson machine learning challenge. In *Proceedings of the NIPS* 2014 *Workshop on High-Energy Physics and Machine Learning* (G. Cowan, C. Germain, I. Guyon, B. Kégl and D. Rousseau, eds.). *Proceedings of Machine Learning Research* **42** 19–55. PMLR, Montreal, Canada.

ANDREASSEN, A., NACHMAN, B. and SHIH, D. (2020). Simulation assisted likelihood-free anomaly detection. *Phys. Rev. D* **101** 095004.

ATLAS COLLABORATION (2012). Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett. B* **716** 1–29.

ATLAS COLLABORATION (2014). Dataset from the ATLAS Higgs Boson Machine Learning Challenge 2014. CERN Open Data Portal.

ATLAS COLLABORATION (2019). A strategy for a general search for new phenomena using data-derived signal regions and its application within the ATLAS experiment. *Eur. Phys. J. C* **79** 120.

ATLAS COLLABORATION AND CMS COLLABORATION (2011). LHC Higgs Combination Group, Procedure for the LHC Higgs boson search combination in Summer 2011. Technical Report, CMS-NOTE-2011-005.

BACH, S., BINDER, A., MONTAVON, G., KLAUSCHEN, F., MÜLLER, K.-R. and SAMEK, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **10** e0130140. https://doi.org/10.1371/journal.pone.0130140

BEHNKE, O., KRÖNINGER, K., SCHOTT, G. and SCHÖRNER-SADENIUS, T. (2013). *Data Analysis in High Energy Physics*: *A Practical Guide to Statistical Methods*. Wiley, New York.

BHAT, P. C. (2011). Multivariate analysis methods in particle physics. *Annu. Rev. Nucl. Part. Sci.* **61** 281–309.

BÖHNING, D., DIETZ, E., SCHAUB, R., SCHLATTMANN, P. and LINDSAY, B. G. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Ann. Inst. Statist. Math.* **46** 373–388.

BOSTRÖM, H. (2008). Calibrating random forests. In 2008 *Seventh International Conference on Machine Learning and Applications* 121–126.

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.

CASA, A. and MENARDI, G. (2018). Nonparametric semisupervised classification for signal detection in high energy physics. ArXiv preprint. Available at arXiv:1809.02977.

CDF COLLABORATION (2008). Model-independent and quasi-model-independent search for new physics at CDF. *Phys. Rev. D* **78** 012002.

CDF COLLABORATION (2009). Global search for new physics with $2.0\,fb^{-1}$ at CDF. *Phys. Rev. D* **79** 011101.

CHAKRAVARTI, P., KUUSELA, M., LEI, J. and WASSERMAN, L. (2023). Supplement to "Model-independent detection of signals using interpretable semi-supervised classifier tests." https://doi.org/10.1214/22-AOAS1722SUPPA, https://doi.org/10.1214/22-AOAS1722SUPPB

CHANDOLA, V., BANERJEE, A. and KUMAR, V. (2009). Anomaly detection: A survey. *ACM Comput. Surv.* **41** 1–58.

CHOUDALAKIS, G. (2008). Model independent search for new physics at the Tevatron. ArXiv preprint. Available at arXiv:0805.3954.

CMS COLLABORATION (2012). Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys. Lett. B* **716** 30–61.

CMS COLLABORATION (2017). MUSiC, a model unspecific search for new physics, in pp collisions at $\sqrt{s} = 8$ TeV. CMS Physics Analysis Summary CMS-PAS-EXO-14/016.

CMS COLLABORATION (2020). MUSiC, a model unspecific search for new physics, in pp collisions at sqrt (s) = 13 TeV. Technical Report, Technical report CMS-PAS-EXO-19-008, CERN, Geneva.

COLLINS, J., HOWE, K. and NACHMAN, B. (2018). Anomaly detection for resonant new physics with machine learning. *Phys. Rev. Lett.* **121** 241803. https://doi.org/10.1103/PhysRevLett.121.241803

COLLINS, J. H., HOWE, K. and NACHMAN, B. (2019). Extending the search for new resonances with machine learning. *Phys. Rev. D* **99** 014038.

CONSTANTINE, P. G. (2015). *Active Subspaces*: *Emerging Ideas for Dimension Reduction in Parameter Studies*. *SIAM Spotlights* **2**. SIAM, Philadelphia, PA. MR3486165 https://doi.org/10.1137/1.9781611973860

CONSTANTINE, P. G., DOW, E. and WANG, Q. (2014). Active subspace methods in theory and practice: Applications to Kriging surfaces. *SIAM J. Sci. Comput.* **36** A1500–A1524. MR3233940 https://doi.org/10.1137/130916138

CONSTANTINE, P. G., EMORY, M., LARSSON, J. and IACCARINO, G. (2015). Exploiting active subspaces to quantify uncertainty in the numerical simulation of the HyShot II scramjet. *J. Comput. Phys.* **302** 1–20. MR3404505 https://doi.org/10.1016/j.jcp.2015.09.001

COWAN, G., CRANMER, K., GROSS, E. and VITELLS, O. (2011). Asymptotic formulae for likelihood-based tests of new physics. *Eur. Phys. J. C* **71** 1554.

CRANMER, K. (2015). Practical statistics for the LHC. ArXiv preprint. Available at arXiv:1503.07622.

CRANMER, K., PAVEZ, J. and LOUPPE, G. (2015). Approximating likelihood ratios with calibrated discriminative classifiers. ArXiv preprint. Available at arXiv:1506.02169.

CUI, C., ZHANG, K., DAULBAEV, T., GUSAK, J., OSELEDETS, I. and ZHANG, Z. (2020). Active subspace of neural networks: Structural analysis and universal attacks. *SIAM J. Math. Data Sci.* **2** 1096–1122. MR4168214 https://doi.org/10.1137/19M1296070

D'AGNOLO, R. T. and WULZER, A. (2019). Learning new physics from a machine. *Phys. Rev. D* **99** 015014.

D'AGNOLO, R. T., GROSSO, G., PIERINI, M., WULZER, A. and ZANETTI, M. (2021). Learning multivariate new physics. *Eur. Phys. J. C* **81** 1–21.

D'AGNOLO, R. T., GROSSO, G., PIERINI, M., WULZER, A. and ZANETTI, M. (2022). Learning new physics from an imperfect machine. *Eur. Phys. J. C* **82** 1–37.

D0 COLLABORATION (2012). Model independent search for new phenomena in pp (bar) collisions at $\sqrt{s} =$ 1.96 TeV. *Phys. Rev. D* **85**.

DAUNCEY, P., KENZIE, M., WARDLE, N. and DAVIES, G. (2015). Handling uncertainties in background shapes: The discrete profiling method. *J. Instrum.* **10** P04015.

DAVISON, A. C. and HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application*. *Cambridge Series in Statistical and Probabilistic Mathematics* **1**. Cambridge Univ. Press, Cambridge. MR1478673 https://doi.org/10.1017/CBO9780511802843

DOBSON, A. J. and BARNETT, A. G. (2018). *An Introduction to Generalized Linear Models*, 4th ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR3890007

DORIGO, T. and DE CASTRO, P. (2020). Dealing with nuisance parameters using machine learning in high energy physics: A review. ArXiv preprint. Available at arXiv:2007.09121.

EFRON, B. and TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap*. *Monographs on Statistics and Applied Probability* **57**. CRC Press, New York. MR1270903 https://doi.org/10.1007/978-1-4899-4541-9

FRIEDMAN, J. H. (2003). On multivariate goodness-of-fit and two-sample testing. In *PHYSTAT* 2003, *SLAC*, *Stanford*, *California*.

GHOSH, A., NACHMAN, B. and WHITESON, D. (2021). Uncertainty-aware machine learning for high energy physics. *Phys. Rev. D* **104** 056026.

GHOSH, J. K. and SEN, P. K. (1985). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. II* (*Berkeley, Calif.*, 1983). *Wadsworth Statist./Probab. Ser.* 789–806. Wadsworth, Belmont, CA. MR0822065

GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A. and BENGIO, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems* 2672–2680.

GRÖMPING, U. (2009). Variable importance assessment in regression: Linear regression versus random forest. *Amer. Statist.* **63** 308–319. MR2751747 https://doi.org/10.1198/tast.2009.08199

H1 COLLABORATION (2004). A general search for new phenomena in ep scattering at HERA. *Phys. Lett. B* **602** 14–30.

HANLEY, J. A. and MCNEIL, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143** 29–36.

HANLEY, J. A. et al. (1989). Receiver operating characteristic (ROC) methodology: The state of the art. *Crit Rev Diagn Imaging* **29** 307–335.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2722294 https://doi.org/10.1007/978-0-387-84858-7

HENZE, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *Ann. Statist.* **16** 772–783. MR0947577 https://doi.org/10.1214/aos/1176350835

ISHWARAN, H. (2007). Variable importance in binary regression trees and forests. *Electron. J. Stat.* **1** 519–537. MR2357716 https://doi.org/10.1214/07-EJS039

KASIECZKA, G., NACHMAN, B., SHIH, D., AMRAM, O., ANDREASSEN, A., BENKENDORFER, K., BORTOLATO, B., BROOIJMANS, G., CANELLI, F. et al. (2021). The LHC Olympics 2020: A community challenge for anomaly detection in high energy physics. ArXiv preprint. Available at arXiv:2101.08320.

KIM, I., LEE, A. B. and LEI, J. (2019). Global and local two-sample tests via regression. *Electron. J. Stat.* **13** 5253–5305. MR4043073 https://doi.org/10.1214/19-EJS1648

KIM, I., RAMDAS, A., SINGH, A. and WASSERMAN, L. (2021). Classification accuracy as a proxy for two-sample testing. *Ann. Statist.* **49** 411–434. MR4206684 https://doi.org/10.1214/20-AOS1962

KNUTESON, B. (2000). Ph.D. thesis, University of California at Berkeley.

KUUSELA, M., VATANEN, T., MALMI, E., RAIKO, T., AALTONEN, T. and NAGAI, Y. (2012). Semi-supervised anomaly detection–towards model-independent searches of new physics. In *Journal of Physics*: *Conference Series* **368** 012032. IOP Publishing, Bristol.

LEI, J., G'SELL, M., RINALDO, A., TIBSHIRANI, R. J. and WASSERMAN, L. (2018). Distribution-free predictive inference for regression. *J. Amer. Statist. Assoc.* **113** 1094–1111. MR3862342 https://doi.org/10.1080/01621459.2017.1307116

LYONS, L. and WARDLE, N. (2018). Statistical issues in searches for new phenomena in high energy physics. *J. Phys. G, Nucl. Part. Phys.* **45** 033001.

METZ, C. E. (1978). Basic principles of ROC analysis. In *Seminars in Nuclear Medicine* **8** 283–298. Elsevier, Amsterdam.

NACHMAN, B. (2020). A guide for deploying deep learning in LHC searches: How to achieve optimality and account for uncertainty. *SciPost Phys.* **8** Paper No. 090. MR4196320 https://doi.org/10.21468/scipostphys.8.6.090

NACHMAN, B. and SHIH, D. (2020). Anomaly detection with density estimation. *Phys. Rev. D* **101** 075042.

NELDER, J. A. and WEDDERBURN, R. W. (1972). Generalized linear models. *J. R. Stat. Soc., A* **135** 370–384.

NEWCOMBE, R. G. (2006). Confidence intervals for an effect size measure based on the Mann–Whitney statistic. Part 2: Asymptotic methods and evaluation. *Stat. Med.* **25** 559–573. MR2222114 https://doi.org/10.1002/sim.2324

NICULESCU-MIZIL, A. and CARUANA, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*. *ICML* 2005 625–632. Association for Computing Machinery, New York, NY, USA.

PARTICLE DATA GROUP (2020). Review of particle physics. *PTEP* **2020** 083C01.

PEYRÉ, G., CUTURI, M. et al. (2019). Computational optimal transport: With applications to data science. *Found. Trends Mach. Learn.* **11** 355–607.

RADOVIC, A., WILLIAMS, M., ROUSSEAU, D., KAGAN, M., BONACORSI, D., HIMMEL, A., AURISANO, A., TERAO, K. and WONGJIRAD, T. (2018). Machine learning at the energy and intensity frontiers of particle physics. *Nature* **560** 41–48. https://doi.org/10.1038/s41586-018-0361-2

REISS, R.-D. (1993). *A Course on Point Processes*. *Springer Series in Statistics*. Springer, New York. MR1199815 https://doi.org/10.1007/978-1-4613-9308-5

SCHILLING, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *J. Amer. Statist. Assoc.* **81** 799–806. MR0860514

SHRIKUMAR, A., GREENSIDE, P. and KUNDAJE, A. (2017). Learning important features through propagating activation differences. ArXiv preprint. Available at arXiv:1704.02685.

SOHA, A. L. (2008). General searches for new physics. In 34*th International Conference on High Energy Physics*.

STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 479–498. MR1924302 https://doi.org/10.1111/1467-9868.00346

STROBL, C., BOULESTEIX, A.-L., KNEIB, T., AUGUSTIN, T. and ZEILEIS, A. (2008). Conditional variable importance for random forests. *BMC Bioinform.* **9** 307. https://doi.org/10.1186/1471-2105-9-307

SUNDARARAJAN, M., TALY, A. and YAN, Q. (2017). Axiomatic attribution for deep networks. ArXiv preprint. Available at arXiv:1703.01365.

VAN DER LAAN, M. J. (2006). Statistical inference for variable importance. *Int. J. Biostat.* **2** Art. 2. MR2275897 https://doi.org/10.2202/1557-4679.1008

VATANEN, T., KUUSELA, M., MALMI, E., RAIKO, T., AALTONEN, T. and NAGAI, Y. (2012). Semi-supervised detection of collective anomalies with an application in high energy particle physics. In *The* 2012 *International Joint Conference on Neural Networks* (*IJCNN*) 1–8. IEEE, New York.

WILLIAMSON, B. D., GILBERT, P. B., SIMON, N. R. and CARONE, M. (2020). A unified approach for inference on algorithm-agnostic variable importance. ArXiv preprint. Available at arXiv:2004.03683.