Machine learning techniques for intermediate mass gap lepton partner searches at the large hadron collider

Bhaskar Dutta,[1] Tathagata Ghosh,[2] Alyssa Horne,[3,4] Jason Kumar,[5] Sean Palmer,[3,6] Pearl Sandick,[7] Marcus Snedeker,[3,8] Patrick Stengel,[9] and Joel W. Walker[3]

[1]*Department of Physics and Astronomy, Mitchell Institute for Fundamental Physics and Astronomy, Texas A&M University, College Station, Texas 77843, USA*
[2]*Harish-Chandra Research Institute, A CI of Homi Bhabha National Institute, Chhatnag Road, Jhusi, Prayagraj 211019, India*
[3]*Department of Physics, Sam Houston State University, Huntsville, Texas 77341, USA*
[4]*Department of Physics, Michigan Technological University, Houghton, Michigan 49931, USA*
[5]*Department of Physics and Astronomy, University of Hawai'i, Honolulu, Hawaii 96822, USA*
[6]*Department of Physics, New Mexico State University, Las Cruces, New Mexico 88003, USA*
[7]*Department of Physics and Astronomy, University of Utah, Salt Lake City, Utah 84112, USA*
[8]*Department of Physics, East Carolina University, Greenville, North Carolina 27858, USA*
[9]*Istituto Nazionale di Fisica Nucleare, Sezione di Ferrara, via Giuseppe Saragat 1, I-44122 Ferrara, Italy*

(Received 1 October 2023; accepted 22 March 2024; published 11 April 2024)

We consider machine learning techniques associated with the application of a boosted decision tree (BDT) to searches at the Large Hadron Collider (LHC) for pair-produced lepton partners which decay to leptons and invisible particles. This scenario can arise in the minimal supersymmetric Standard Model (MSSM), but can be realized in many other extensions of the Standard Model (SM). We focus on the case of intermediate mass splitting (∼30 GeV) between the dark matter (DM) and the scalar. For these mass splittings, the LHC has made little improvement over LEP due to large electroweak backgrounds. We find that the use of machine learning techniques can push the LHC well past discovery sensitivity for a benchmark model with a lepton partner mass of ∼110 GeV, for an integrated luminosity of 300 fb^{-1}, with a signal-to-background ratio of ∼0.3. The LHC could exclude models with a lepton partner mass as large as ∼160 GeV with the same luminosity. The use of machine learning techniques in searches for scalar lepton partners at the LHC could thus definitively probe the parameter space of the MSSM in which scalar muon mediated interactions between SM muons and Majorana singlet DM can both deplete the relic density through dark matter annihilation and satisfy the recently measured anomalous magnetic moment of the muon. We identify several machine learning techniques which can be useful in other LHC searches involving large and complex backgrounds.

DOI: 10.1103/PhysRevD.109.075018

I. INTRODUCTION

A wide variety of scenarios for physics beyond the Standard Model (BSM) have been probed experimentally with the Large Hadron Collider (LHC). However, as no conclusive evidence of BSM physics has been found yet, focus has turned to scenarios and signatures which are more difficult to probe. One particularly difficult scenario is the pair production of scalar lepton partners of SM fermions ($\tilde{\ell}^{\pm}$), each of which decays to a lepton and an invisible particle ($\tilde{\ell} \to \ell X$) with a ∼30 GeV mass splitting. This scenario can arise in the minimal supersymmetric standard model (MSSM) [1,2], as well as in other BSM models [2,3] often specifically motivated by the need for a viable dark matter candidate and by recent measurements of the anomalous magnetic moment of the muon. This scenario is difficult to probe at the LHC because there is a large SM background from the production of electroweak gauge bosons, decaying to leptons and neutrinos. As a result, current LHC constraints [4,5] on this scenario show little improvement over LEP [6]. Although a variety of new analysis strategies have been proposed, there is no clear-cut strategy for effectively separating signal from background in models with these moderately compressed particle spectra. In this work, we investigate the possibility of using machine learning algorithms to more effectively pick out signal from background.

The standard signature of scalar lepton partner pair production at the LHC is $\bar{\ell}\ell$ plus missing transverse energy E_{T} (MET). However, if the mass splitting is $\lesssim$60 GeV, then the lepton and invisible particles are both soft, and it is difficult to see this signature above the electroweak background [7]. One strategy used to evade the difficulty of soft particles is to search for events in which one or more hard jets are also emitted ($pp \to \tilde{\ell}^*\tilde{\ell} +$ jets). This jet provides a transverse kick to the $\tilde{\ell}^*\tilde{\ell}$ system, yielding harder leptons and larger MET. Even so, distinguishing signal events from the large electroweak background requires additional techniques. If the mass splitting is small ($\lesssim$20 GeV), then the electroweak background can be rejected because the neutrinos produced by W decay typically carry larger MET than carried by the X arising from $\tilde{\ell}$ decay [8]. The efficacy of this search strategy has been enhanced by recent experimental developments, resulting in lower energy thresholds for lepton identification [9]. However, searches at the LHC are still challenging for intermediate mass splittings in the range of ∼20–50 GeV.

Theoretical work has shown that progress can be made for mass splittings in the ∼20–50 GeV range, using a variety of kinematic variables involving the energies of the leptons, jets, and MET, as well as their angular correlations with each other [10]. But the progress is incremental; even applying these strategies, a model with $M_{\tilde{\ell}} = 160$ GeV and $M_{\tilde{\ell}} - M_X = 30$ GeV would be well out of reach of LHC with 300 fb^{-1} luminosity [10]. Moreover, there is no clean set of simple cuts which one can apply to maximize sensitivity, based on an *a priori* principle. Instead, one sequentially applies cuts to many kinematic variables, with each cut (and the order of cuts) being determined largely by trial and error. This is the type of setting in which one might expect machine learning algorithms to greatly improve search sensitivity, by determining the optimal set of cuts to impose on a large number of kinematic variables which can exhibit strong (nonlinear) correlations. Additionally, one might hope that it would be possible to work backward after an optimal set of cuts is found to determine the underlying principle(s) responsible for the efficacy of those cuts. We will see that both of these intuitions are true.

In this work, we consider a benchmark point which is allowed by current constraints (cf. discussion in Sec. III) where the scalar lepton is a partner to the left-handed muon, with $M_{\tilde{\ell}} = 110$ GeV, $M_{\tilde{\ell}} - M_{\chi} = 30$ GeV. We use a boosted decision tree (BDT) [11,12] to classify events as signal or background, finding that a cut based on the BDT classifier can increase the LHC sensitivity to $\gtrsim 5\sigma$ while maintaining a signal-to-background ratio of ∼0.3 with 300 fb^{-1} of data. We identify kinematic variables that dominantly contribute to the signal sensitivity and plot the residual distribution as applicable.

Importantly, we find that direct application of a BDT algorithm to simulated data is not sufficient. Essentially, because some SM background processes have very large rates, application of a BDT algorithm to an uncurated data sample will succeed at removing the largest backgrounds, while leaving subleading backgrounds that, in turn, can limit sensitivity. Instead, we adopt a strategy in which some initial precuts are imposed (in addition to the basic cuts which define the event topology) to reduce leading backgrounds with easily identified characteristics, leaving the BDT free to focus subsequently on more challenging features of the residual leading and subdominant backgrounds.

In a similar vein, we find that generation of the training sample used by the BDT also requires special care, because the most difficult backgrounds to remove lie in regions of phase space with relatively small cross section (though larger than the signal). If these regions of phase space are not adequately sampled by simulation, then the BDT can become overly focused on the details of a few particular events in the training sample. To avoid this problem, we simulate in kinematic tranches, significantly enhancing the fraction of computational time devoted to kinematic regions having smaller cross sections but larger pass rates for standard cuts.

In previous studies, a variety of machine learning techniques have been proposed and implemented for analyses of LHC data in which the BSM signal is particularly challenging to differentiate from the SM background. Following an early proposal for the use of BDTs and neural networks (NNs) in searches for fermionic partners of SM electroweak gauge bosons [13], the pair production of such fermionic partners has been constrained in a BDT analysis of $(\bar{\ell}\ell + \mathrm{MET})$ final states at the LHC [4]. For boosted topologies, both BDT and NN analyses have been shown to enhance the sensitivity of LHC searches to dark matter models with extended Higgs sectors [14]. Analyses of monojet plus missing energy signatures at LHC in several simplified dark matter models with mediators of different spins has demonstrated that NNs are able to efficiently determine if there is a BSM signal within a given dataset when trained on two-dimensional histograms combining information from different kinematic variables [15,16].

In addition to supervised learning applications, less supervised techniques have been proposed for BSM searches at LHC. Weakly supervised techniques have been shown to increase the sensitivity of monojet plus missing energy searches to strongly interacting dark matter models with anomalous jet dynamics [17] and self-supervised contrastive learning has been proposed for anomaly detection searches in dijet events [18]. Complementing these largely model-agnostic techniques, adversarial NNs have also been suggested as an unsupervised approach to improve the invariant mass reconstruction of BSM gauge bosons decaying to leptons and neutrinos [19].

In this work, we favor an application of supervised learning strategies using a BDT due to the relatively

straightforward manner in which the algorithm classifies signal and background events, facilitating a clear interpretation of results. We will focus not only on the search sensitivity which is achieved by application of a BDT, but also on what the BDT teaches us regarding the underlying search strategy, and on improved techniques for applying machine learning to general LHC searches with large and complex backgrounds.

The search strategy we propose is relevant for a variety of phenomenologically motivated extensions of the SM. For example, charged mediator models with a Majorana SM-singlet dark matter candidate and scalar muon partners mediating interactions with the SM can both satisfy the dark matter relic density and account for the anomalous magnetic moment of the muon [20]. With the coupling of the dark matter to the muon and its scalar partner fixed by the SM hypercharge coupling, as in the MSSM, the relic density in such models can be depleted by scalar mediated dark matter annihilation for lepton partner masses $M_{\tilde{\ell}} \lesssim 150$ GeV [1] and by co-annihilation processes involving the scalars for $M_{\tilde{\ell}} \lesssim 1$ TeV [2]. For simplified models with dark matter-scalar-lepton couplings larger than in the MSSM, both production mechanisms can yield the observed dark matter relic density for scalar masses $M_{\tilde{\ell}} \gtrsim 1$ TeV [2,3]. Charged mediator models with the most massive scalar muon partners mentioned above are virtually impossible to detect at LHC due to the falling production cross sections at higher scalar masses. However, there are large regions of unconstrained parameter space which are challenging but feasible to probe for LHC searches with the improved sensitivity of analyses using machine learning techniques such as a BDT.

The plan of this paper is as follows. In Sec. II, we describe our approach and motivation for using a BDT. We outline the scalar muon partner signal and associated backgrounds, along with a summary of previously implemented and proposed analyses using cut-and-count strategies, in Sec. III. The details of the event simulation, selected observables, BDT training, and results of our analysis are described in Sec. IV. We discuss the conventional wisdom which can be gained from our study and applied to others in Sec. V, before summarizing our findings and briefly discussing future work in Sec. VI.

II. BACKGROUND AND MOTIVATION FOR USING A BDT

By its nature, known physics is necessarily more prevalent or more readily visible to conventional experimental techniques than unknown physics, while unknown physics and the opportunity to extend our understanding of fundamental particles and interactions are of greater basic interest. Accordingly, the problem of how one efficiently suppresses known backgrounds in order to enhance the prospective visibility of new processes is of great interest. In a prior study [10], we attempted to systematically separate the signal associated with the production of lepton-partners in intermediate mass gap scenarios from competing SM backgrounds via an iterative process involving the plotting of distributions in relevant kinematic observables and the manual application of event selection cuts. This approach proved effective, although there are a number of ways in which it is potentially suboptimal.

First, it was observed that this process is extremely sensitive to the order in which the cuts are applied, since a variety of useful secondary and tertiary event selections remain obscured until a majority of foreground debris is removed by primary selections. A corollary of this statement is that the possible sequencing of cuts explodes combinatorially, and it is very difficult to be certain that a given sequence of cuts is in any sense optimal. Likewise, there is a great danger that early event selections can be applied too aggressively, precluding more surgically targeted cuts down the line. This is part of a larger problem, that any by-hand procedure is intrinsically *ad hoc* and one-off, while simultaneously remaining extremely labor-intensive. Finally, there is a risk of biasing the analysis in the limit of low statistics, which always ultimately applies as selections become increasingly strict, since cuts are typically engineered and validated with respect to the same collection of events.

Machine learning offers substantial promise for alleviating negative outcomes associated with each of these objections. Specifically, it can much more thoroughly and efficiently scan the space of available discriminants, while delivering results that are more stable and more reproducible, using a generalizable approach that has increased cross-applicability, and which requires less investment of human capital into repetitive and automatable tasks. Additionally, leading machine learning algorithms offer built-in protection against over fitting. For example, the learning rate can be turned down such that the likelihood of over-applying an early selection and losing sensitivity to future gains is reduced. Likewise, the training and testing samples can be readily isolated, reducing the risk of learning random features of the presented sample that are not replicated in the larger ensemble. Also, the separation between training and testing can typically be "folded" in multiple ways, allowing for cross-validation via replication of the analysis, which helps to quantify the stability of results against statistical fluctuations. Much greater complexity and refinement of the discriminant is available in a machine learning context relative to manual approaches. Additionally, the assignment of a continuous event-by-event classification likelihood represents a much richer category of information than the discretization implicit to a more basic cut-and-count approach.

However, a pressing concern associated with machine learning is that one often sacrifices the ability to investigate *what* was learned, and to develop intuition regarding the nature of the signal and background separation. This is a

key reason that we favor the use of supervised learning with a BDT rather deep learning approaches utilizing more complex algorithms such as NNs.[1] In particular, the input to a BDT is a simple list of numerical data associated with each event, together with an event weight and a known classification in the case of the training sample. The BDT goes through a process that is similar to the design of a manual event selection profile in many regards, but it is strictly systematic, mathematically rigorous, and reproducible. At each stage of training, the optimal discriminant and splitting value is ascertained according to a well-defined loss function; subsequently, each branch of this "tree" can elect its own best choice for the next splitting, and so on. Likewise, the classification score assigned to each terminal leaf of the tree is selected by extremization of the same functional, in the effort to optimally reconcile feature-based predictions with corresponding truth labels. The capacity to differentially select supplementary discriminants for previously separated populations adds a level of flexibility and refinement that is quite challenging to implement by hand.

A number of such trees are successively generated in the "boosting" process, wherein events are reweighted to emphasize the correction of prior missorts in subsequent refinements of the training. By combining a long sequence of relatively shallow trees, a strong discriminant is built from the conjunction of many "weak learners." Adjustable "regularization" factors built into the training objective may be tuned to veto branchings that would add complexity without meaningfully reducing the classification loss, and to slow the rate of learning to limit the danger of over fitting. Training may also be halted prior to completion of the specified maximal tree count if real-time validation against the testing sample indicates an onset of diminishing returns via a plateau in the loss function. Collectively, these measures help to mitigate the so-called "bias-variance" problem, i.e., the problem of over training on features that are not widely generalizable. Ultimately, the final classification score assigned to a given event represents the sum over its leaf scores for all trees. This unbounded value y is typically mapped onto the bounded range $p(y) \in \{0, 1\}$ via application of the sigmoid "logistic" function, or similar. Crucially, the process by which the classification scores are assigned is entirely tractable, and one may output the full set of constructed decision trees if so desired, including the splitting features and values, as well as the selected leaf scores. In addition, a summary report of the "importance" of each feature to the training is readily available, and it is possible to iteratively reduce the dimensionality of the provided feature space using this information, selectively eliminating redundant (highly correlated) and indiscriminant features.

Rareness of the targeted new physics processes, and likewise, of the most competing SM backgrounds (in the tail regions of the background distributions), implies that it is generally computationally impossible to work with simulated events in their naturally occurring proportions. The solution is to keep track of per-event weights, i.e., partial cross sections linked to each event that represent the extent of the final-state phase space for which they are a working proxy. Even within a category of final state particles, e.g., top quark pair production or dibosons plus jets, events representing the higher-energy hard scatterings that are generally of greater interest after an application of selection cuts are generally power-suppressed relative to their softer low-energy counterparts. Accordingly, it becomes very computationally inefficient to simulate enormous quantities of events that are dominated by softer scatterings that will mostly be discarded in order to emphasize harder events in the distribution tails. This can likewise be handled by separating each simulation category into disjoint tranches that are binned in some relevant process scale such as the scalar sum over transverse momentum p_{T}. A significant quantity of rarer events can be generated at lower cost in this way, allowing the analysis to maintain a more uniform level of statistical representation across the phase space. The rare events then simply carry smaller per-event weights, which can be fed into the machine learning in order to direct its attention appropriately toward the most proportionally relevant features at a given stage of the training.

Although more sophisticated types of deep learning such as neural networks (NNs) can potentially outperform BDTs in the classification of signal and background events, previous studies suggest the performance of NNs and BDTs are comparable. In a study of boosted event topologies relevant for dark matter models with extended Higgs sectors [14], BDT and NN implementations yield similar increases in sensitivity relative to a cut-based analysis. A comparison between deep NNs and BDTs in Ref. [13] demonstrates only marginally better performance by NNs in an analysis of scalar lepton pair production, without the boost from a hard jet. A similar comparison in searches for fermionic partners of electroweak gauge bosons shows a more significant improvement in performance by deep NNs compared to BDTs, which is attributed to the lack of high-level kinematic variables, whose larger discriminating power is necessary for an optimal BDT analysis. A dedicated study of NNs implemented for the pair production of scalar leptons in boosted event topologies at LHC is thus an interesting possibility for future work.

III. CHARACTERIZING SIGNAL AND LEADING BACKGROUNDS

For our analysis, we focus on a model in which the scalar lepton $\tilde{\ell}$ is the partner of the left-handed muon, with

[1]While deep NNs are typically considered to be less inherently interpretable than a BDT, there has been significant recent work on both algorithm-agnostic and NN-specific methods to evaluate the explainability of deep learning algorithms (for a recent overview, see Ref. [21]).

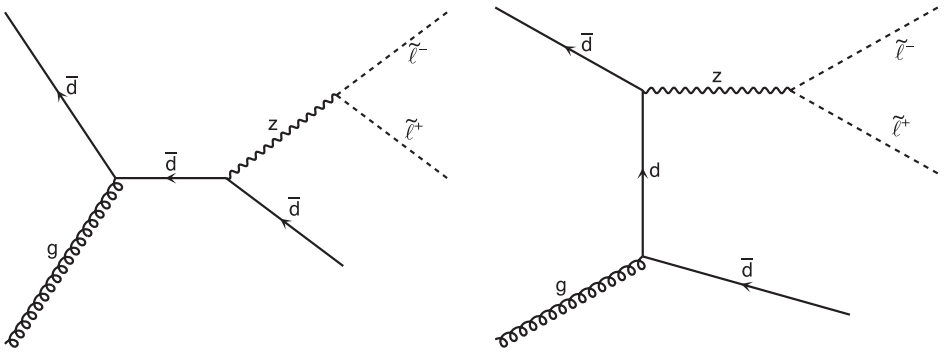

FIG. 1. Two example Feynman diagrams representing signal production with an associated jet.

$M_{\tilde{\ell}} = 110$ GeV and $M_{\tilde{\ell}} - M_X = 30$ GeV ($\ell = \mu$). We assume X is a stable SM-singlet Majorana fermion. This scenario is thus realized in the MSSM, where $\tilde{\ell}$ is a left-handed smuon, X is a bino-like lightest supersymmetric particle (LSP), and other SUSY particles are relatively heavy. However, this scenario can also be realized in a variety of other phenomenologically well-motivated BSM models. The event topology we are interested in is $pp \to \bar{\ell}\ell + \text{MET} + \text{jets}$. This topology can be produced by signal events in which proton collisions result in $\tilde{\ell}^*\tilde{\ell}$ pair production, with each new scalar decaying via $\tilde{\ell} \to \ell X$, with one or more jets. We simulate the signal process using the `MSSM_SLHA2` [22] model distributed with MadGraph [23]. Model parameters are generated consistently using the SUSY-hit [24] package, with benchmark smuon and neutralino masses taking the values indicated previously. In addition to initial-state radiation, jets can also be produced as part of the hard scattering, and we simulate the signal process inclusively, with 0, 1, or 2 final-state jets (with matching). Two example Feynman diagrams (from MadGraph) are shown in Fig. 1.

This region of parameter space is still allowed by analyses of 139 fb^{-1} of ATLAS data [4,8], which did not utilize machine learning techniques. Indeed, for a mass splitting of 30 GeV, the tightest constraint is still from LEP [6], which rules out the region of parameter space with $M_{\tilde{\ell}} \lesssim 97$ GeV for right-handed smuons.[2] A scenario with similar phenomenology is explored in Ref. [25].

The leading SM backgrounds to this signal arise from processes in which the charged leptons are produced from the decay of on-shell weak gauge bosons, with missing energy arising either from neutrinos or jet mismeasurement. The main such processes are

(i) $pp \to Zj$, with $Z \to \bar{\tau}\tau, \bar{\ell}\ell$. If Z decays to $\bar{\tau}\tau$, then the τs decay to ℓ, $\bar{\ell}$ and neutrinos, while if Z decays to $\bar{\ell}\ell$, then MET arises from mismeasurement of the jet energy;
(ii) $pp \to \bar{t}t$, with $t \to bW$, and the Ws decaying to ℓ, $\bar{\ell}$ and neutrinos;
(iii) $pp \to ZZ, W^+W^-$, with the massive gauge boson decays producing ℓ, $\bar{\ell}$ and neutrinos;
(iv) $pp \to \bar{\tau}\tau j$, with the τs decaying to ℓ, $\bar{\ell}$ and neutrinos;
(v) $pp \to WZ$, with the massive gauge boson decays producing a neutrino and three charged leptons, with one charged lepton missed.

Of these background processes, the rate for $pp \to Zj$ dominates by roughly two orders of magnitude over the nearest competitor. But the rates for all of these processes are much larger than for $pp \to \tilde{\ell}^*\tilde{\ell}$. Thus, to obtain significant improvements in sensitivity and signal-to-background ratio, it is necessary to strongly reject the dominant background, while still retaining strong rejection of the subleading backgrounds.

Because $pp \to Zj$ events largely yield a Z near rest, the neutrinos produced from the products of the $Z \to \bar{\tau}\tau$ decay process tend to have low energy. As a result, a requirement of minimum missing energy is successful in reducing this background. The requirement of a minimum missing energy also reduces the background arising from $pp \to Zj$, with $Z \to \bar{\ell}\ell$, since the likelihood that MET arises from jet mismeasurement falls with increasing MET. $pp \to Zj$ events can also be distinguished by the kinematic variables built from the lepton and jet momenta, including the dilepton invariant mass ($M_{\ell\ell}$) and the ditau invariant mass ($M_{\tau\tau}$), which will be described further in the next section. The utility of the $M_{\tau\tau}$ variable arises from the fact that, in the $Z \to \bar{\tau}\tau$ process, the τs are boosted. As a result, the neutrinos produced by τ-decay are largely collinear with the

[2]LEP searches did not attempt to constrain left-handed smuons due to the relatively heavy masses expected in concrete realizations of the MSSM. Since LEP had sufficient luminosity for the right-handed smuon searches to be kinematically limited by the center-of-mass energy of the beamline, the constraints on left-handed smuon masses are expected to be similar due to the larger cross sections relative to right-handed smuon production.

charged leptons, allowing one to reconstruct the mass of the parent particle using transverse momentum conservation. The difficulty lies not so much in removing the Z-related background, but in doing so without compromising one's ability to distinguish signal from the remaining backgrounds. In [10], the cut-and-count strategy for reducing Z-related backgrounds focused at the level of primary selections on:

(i) Rejecting events with $M_{\ell\ell}$ within $M_Z \pm 10$ GeV,
(ii) Rejecting events with $M_{\tau\tau} < 125$ GeV, and
(iii) Rejecting events with MET < 125 GeV.

Additional kinematic variables can be used to discriminate between signal and W- and top-related backgrounds, based on the energy and angular distribution of the decay processes. In [10], a total of seven secondary and tertiary cuts were applied in sequence to define a signal region. We will see that a BDT analysis can provide substantial improvement relative to that approach.

IV. SIMULATION AND BDT ANALYSIS

A. Event simulation

Monte Carlo training samples were generated for the $\sqrt{s} = 13$ TeV LHC[3] using MadGraph/MadEvent [23], with showering and hadronization in PYTHIA8 [26], and detector simulation in DELPHES [27]. We consider events in which one finds a $\mu^+\mu^-$ pair, exactly one hard central jet ($P_T \geq 30$ GeV, $|\eta| < 2.5$), zero b-tagged jets, MET (≥ 30 GeV), and no hadronic τ decays.[4] We refer to these defining event selections as the "topology cuts", and they correspond to the primary event selections described Ref. [10]. They are distinguished from the further "precuts" to be itemized subsequently, which are also applied manually, and prior to the main BDT analysis.

We consider six major background processes: $\bar{\mu}\mu jjj$, $\bar{\tau}\tau jjj$, $\bar{t}tjj$, $WWjj$, $ZZjj$, $WZjj$. Events are simulated inclusively, with up to two or three additional jets, depending on the process (as summarized in Table I). Specifically, we combine processes with zero up to the specified maximal number of hard isolated jets j at the Feynman diagram level (in MadGraph), and perform jet matching (including the simulation of initial state radiation and related effects) in conjunction with PYTHIA8.[5] Note that $Z+$ jets backgrounds are already included in the $\bar{\mu}\mu jjj$ and $\bar{\tau}\tau jjj$ classes, which also include lepton production through off-shell Z^* and photon γ^* mediators. We do not directly simulate the production of tW, whose final state is quite similar to that of $t\bar{t}$. At production level, the inclusive $t\bar{t}+$ jets [29] background dominates by more than an order of magnitude over that of $tW+$ jets [30], although that rate is offset to some extent by an extra opportunity for a b-jet veto. One significant challenge to the inclusive simulation of tW backgrounds is that they are quite difficult to disentangle from double counting with $t\bar{t}$. The cross sections for signal and background events with the stipulated topology are given in the middle column of Table I. Note that cross sections for the individual background processes vary widely, with all being larger than the signal.

TABLE I. Table of residual cross sections for the listed processes to have the correct topology (middle column) and to pass the precuts (right column). The top row is the signal process, while the remaining rows are the leading backgrounds.

Process	σ (fb) (topology cuts)	σ (fb) (precuts)
$\tilde{\ell}^*\tilde{\ell}$	12.3	1.38
$\bar{\mu}\mu jjj$	12500	3.19
$\bar{\tau}\tau jjj$	596	14.6
$\bar{t}tjj$	66.8	5.49
$ZZjj$	26.6	0.234
$ZWjj$	46.9	0.355
$WWjj$	72.6	5.44

We tranche the simulation of each final state into nonoverlapping bins of phase space in order to ensure statistically reliable population of the kinematic tails. For direct dilepton production (including $\bar{\tau}\tau jjj$), we break the simulation runs up according to the generator-level P_T of the leading lepton. For the case of $\bar{t}tjj$, we tranche on P_T of the top quark. To accomplish this, it is necessary to asymmetrically decay the $\bar{t}$ directly in MadGraph, since generator-level cuts would otherwise be applied to both the particle and antiparticle.[6] The t is decayed subsequently

[3]Additional data collected by ATLAS and CMS up to and beyond $\mathcal{L} = 300$ fb^{-1} will be at higher center-of-mass energy, $\sqrt{s} \lesssim 14$ TeV. We consider $\sqrt{s} = 13$ TeV for ease of comparison to previous analyses and note that the production cross sections for all relevant signal and background processes vary by $\lesssim$10% between $\sqrt{s} = 13$ TeV and $\sqrt{s} = 14$ TeV. All processes are calculated at tree-level since, as discussed in Ref. [10], K-factors which account for higher order corrections have little impact on the projected sensitivity.

[4]We assume, following [4,28], that the effects of pile-up can be reduced by requiring that, for jets with $p_T < 60$ GeV, a sufficient fraction of the tracks associated with the jet point back to the primary vertex, as identified by the jet vertex tagger.

[5]For example, the $\bar{\mu}\mu jjj$ process card includes the instructions `generate p p > l+ l-`, `add process p p > l+ l- j`, `add process p p > l+ l- j j`, and `add process p p > l+ l- j j j`. These processes are simulated by MadGraph and are reliable for hard, well-separated partonic constituents. Showering is performed by PYTHIA8, and is reliable in the complementary limit of soft and/or collinear radiation (where it may generate any number of partonic states). The matching procedure allows each approach to handle its realm of specialization while partitioning the phase space to avoid double counting.

[6]The associated process card instructions are `generate p p > t t, t > w- j`, `add process p p > t t j, t > w- j`, and `add process p p > t t j j, t > w- j`.

in PYTHIA, as usual. For $ZWjj$, we tranche on P_{T} of the Z-boson. The same tranching is applied to $ZZjj$ production, after decaying one of the two Z-bosons leptonically. For $WWjj$, we require at least one of the W-bosons to decay leptonically, and again tranche on the leading (or only) leptonic P_{T}. The signal model is forced to decay to a final state including a di-muon pair and tranched on P_{T} of the leading muon. Several bins of increasing width were simulated for each of these final state discriminants, in order to represent soft, intermediate, and hard multi-TeV scattering processes. Specifically, the numerical boundaries selected for this study were 50, 100, 150, 200, 300, 400, 500, 750, 1000, 1500, 2000, 3000, and 4000 GeV. In this manner, events were generated with a smooth distribution of per-event weights spanning around 6-8 orders of magnitude. We verified explicitly that the associated production cross sections summed over tranches are consistent with those obtained by a single joint simulation. In total, after jet matching but prior to the application of topology cuts, more than 300 million candidate background events were generated, along with over five million events for the signal benchmark. After precuts, the number of surviving event samples passed to the BDT for further analysis is reduced to around a half million signal events and around 200,000 background events. The corresponding cross sections are indicated in the last column of Table I.

B. Kinematic variables

We provide the BDT with 27 variables that are consistent with the final state topology of an opposite-sign, same-flavor dilepton, a hard (non-b) monojet, and MET. This set includes a variety of sophisticated "high level" observables which are known to be effective against various components of the SM background, as well as a number of more basic kinematic inputs referencing the dilepton pair (ℓ_1, ℓ_2), the hard jet j, and/or the MET. The most difficult task of the BDT is to distinguish the decay process $\tilde{\ell} \to \ell X$ from the background processes which involve W-boson decay $W \to \ell\nu$. We thus include several discriminants that are specifically constructed to emphasize differences in the mass or spin of the parent particle or invisible particle. The set of computed observables is summarized in Table II. For a more detailed discussion on why some of these variables are relevant to unearth the underlying physics of the compressed spectra topology, we refer the reader to Ref. [10]. We do not make any effort to remove degeneracies in the feature set since the BDT is intrinsically resilient to such correlations. Event analysis, including application of selection cuts and the computation of observables, is performed with the AEACuS [31,32] package.

Several of the itemized variables require further description. In particular, $\cos\theta^*_{\ell_1\ell_2}$ [33,34] is equal to the cosine of the polar scattering angle in the frame where the pseudorapidities of the leptons are equal and opposite. It is designed to reflect the fact that the angular distribution of intermediary particles with respect to the beam axis in the parton center-of-mass frame is determined by their spin, and that the lepton angular distribution should reflect this heritage. Much of the practical utility of the $\cos\theta^*_{\ell_1\ell_2}$ variable hinges upon its resiliency against longitudinal boosts of the partonic system. This feature is apparent in the definition $\cos\theta^*_{\ell_1\ell_2} \equiv \tanh(\Delta\eta_{\ell_1\ell_2}/2)$, where $\Delta\eta_{\ell_1\ell_2}$ is the pseudorapidity difference between the two leptons. The W-boson associated backgrounds have a distribution in this

TABLE II. Observables delivered to the BDT.

Invariant masses	
$M_{\ell\ell}$	Dilepton system mass
M_j	Jet mass
Mass-like constructions	
M^0_{T2}	Massless invisible hypothesis
M^{100}_{T2}	Massive invisible hypothesis
$M_{\tau\tau}$	Ditau mass
Transverse scales	
E_{T}	Missing transverse energy
H_{T}	Scalar sum of hadronic P_{T}
M_{eff}	Sum of E_{T} and H_{T}
Transverse momentum values	
$P^{\ell_1}_{\mathrm{T}}$	Leading lepton
$P^{\ell_2}_{\mathrm{T}}$	Subleading lepton
P^{j}_{T}	Jet
Scale ratios	
$(M^{100}_{\mathrm{T2}} - 100) \div M^0_{\mathrm{T2}}$	Ratio of excess mass
$P^{\ell_1}_{\mathrm{T}} \div E_{\mathrm{T}}$	Leading lepton to MET
$P^{\ell_2}_{\mathrm{T}} \div E_{\mathrm{T}}$	Subleading lepton to MET
$P^{j}_{\mathrm{T}} \div E_{\mathrm{T}}$	Jet to MET
Azimuthal angular separations	
$\Delta\phi_{\ell_1\not{E}_{\mathrm{T}}}$	Leading lepton to MET
$\Delta\phi_{\ell_2\not{E}_{\mathrm{T}}}$	Subleading lepton to MET
$\Delta\phi_{j\not{E}_{\mathrm{T}}}$	Jet to MET
$\Delta\phi_{\ell_1\ell_2}$	Dilepton system
$\Delta\phi_{\ell_1 j}$	Leading lepton to jet
$\Delta\phi_{\ell_2 j}$	Subleading lepton to jet
Rapidity separations	
$\cos\theta^*$	Dilepton system
$\tanh\lvert\Delta\eta_{\ell_1 j}\rvert$	Leading lepton to jet
$\tanh\lvert\Delta\eta_{\ell_2 j}\rvert$	Subleading lepton to jet
Rapidity values	
η_{ℓ_1}	Leading lepton
η_{ℓ_2}	Subleading lepton
η_j	Jet

variable that is almost flat up to a value around 0.8, whereas distributions for the scalar-mediated signal models more sharply peak at zero, suggesting a clear region of preference, as discussed in Ref. [10]. We construct analogous variables for other differences in pseudorapidity, as applicable.

The M_{T2} [35–37] variable, which was used similarly in Ref. [38], corresponds to the minimal mass of a pair-produced parent which could consistently decay into the visible dilepton system and the observed missing transverse momentum vector sum under a specific hypothesis for the mass of the invisible species. We compute two versions of this variable, M_{T2}^{0} and M_{T2}^{100}, corresponding respectively to a massless hypothesis and a 100 GeV hypothesis. The former is consistent with the MET arising from neutrinos, and the latter is consistent with a dark matter candidate in the neighborhood of our benchmark model.

Finally, the ditau mass variable $M_{\tau\tau}$ [39] attempts to reconstruct the invariant mass of a $\tau\tau$ pair that has decayed leptonically under the hypothesis that the MET is associated with two pair of neutrinos that are emitted collinearly with each of the observed leptons. Momentum conservation in the transverse plane is then sufficient to reconstruct the energy of each neutrino pair, which in turn determines the momentum of each hypothetical τ and allows one to reconstruct the invariant mass of the pair. This quantity may be expressed in closed form [10] as

$$M_{\tau\tau}^2 \equiv -M_{\ell_1\ell_2}^2 \frac{(P_T^{\ell_1} \times P_T^j) \cdot (P_T^{\ell_2} \times P_T^j)}{|P_T^{\ell_1} \times P_T^{\ell_2}|^2}, \tag{1}$$

where $M_{\ell_1\ell_2}^2$ is the invariant squared mass of the visible lepton system. One may confirm that $M_{\tau\tau}^2 = M_Z^2$ if the neutrinos are collinear with the charged leptons produced by τ-decay, with the τs arising from the process $Zj \to \bar{\tau}\tau j$ (see Appendix B for details). Note that $M_{\tau\tau}^2 > 0$ if either $-P_T^j$ or P_T^j lies between $P_T^{\ell_1}$ and $P_T^{\ell_2}$, and is negative if neither do. This makes the ditau mass a good kinematic variable for more generally distinguishing event topology, in addition to rejecting events that involve the process $Z \to \bar{\tau}\tau \to \bar{\ell}\ell + 4\nu$. We define $M_{\tau\tau} \equiv \mathrm{sign}[M_{\tau\tau}^2] \times \sqrt{|M_{\tau\tau}^2|}$.

C. Training

Since it is challenging to generate enough simulation data to represent LHC data in natural proportions, simulated events are weighted based on the cross section of that background class as described previously. Imbalances in the proportional representation of applicable backgrounds lead to another difficulty in a BDT analysis: training samples are often dominated by backgrounds which are easily rejected by intuitively straightforward cuts, such as a dilepton mass cut around the Z-mass. The BDT may thus learn very well how to reject a large set of background events for which a BDT was not really needed, since a simple cut would have done just as well. On the other hand, the BDT may fail to adequately learn how to discriminate more difficult subleading backgrounds, which may likewise be undersampled. Even if undersampling is mitigated by procedures similar to those described in the prior subsection, training may be dominated by a small subset of events carrying large per-event weights.

To address this issue, we impose a series of cuts, in addition to those which define the event topology, to cull data before the curated data is analyzed by the BDT. These event selections are applied to the training dataset and also the test dataset, uniformly. The goal of these initial precuts is to ensure that the surviving background event classes have roughly similar size, with a magnitude which is no more than $\sim\mathcal{O}(10\text{–}100)$ larger than the signal. Essentially, these cuts have done the "easy work," allowing the BDT to focus on the hard work. Moreover, these cuts provide an initial reduction in background based on principles which are known *ab initio*, allowing us to determine which kinematic variables the BDT uses to remove the more difficult backgrounds.

The precuts we use are[7]:

(i) *Veto events with* $M_{\mu\mu} \in 91 \pm 10$ GeV. This cut dramatically reduces the WZj, ZZj and $\mu^+\mu^- j$ backgrounds by removing events in which the $\mu^+\mu^-$ pair arise from Z-decay.

(ii) *Require* MET > 110 GeV. This cut reduces backgrounds, such as $\mu^+\mu^- j$, in which MET arises from jet mismeasurement. In addition, this cut also serves the dual purpose of acting as a trigger [41].

(iii) *Require* $\cos\theta^*_{\mu\mu} < 0.5$. This cut is effective in reducing the WWj background, because it preferentially selects events in which the parent of the muon is spin-0, rather than spin 1.

The cross sections for signal and background events which pass these precuts are given in the right column of Table I. After these precuts, the data is better suited for delivery to the BDT for training on the discrimination of signal from background, and better outcomes are achieved. In particular, the maximal signal-to-background ratio is approximately doubled using this approach, and the window for optimization of the classification score is significantly broadened, implying improved stability of the result.

Our BDT analysis and the generation of associated graphics are done with the MINOS [31,32] package, which calls XGBoost [42] and MATPLOTLIB [43] on the back end. We train the BDT on 2/3 of the simulated event sample, reserving 1/3 for validation of outcomes on a statistically independent sample. We rotate the portion of data reserved for validation in order to characterize statistical variations

[7]While several analyses of LHC data (e.g., [4,30,40]) have utilized a similar window cut on the invariant mass of the dilepton system around the mass of the Z-boson prior to implementing a BDT analysis, a more comprehensive investigation of precuts in such analyses has not been explored in previous studies.

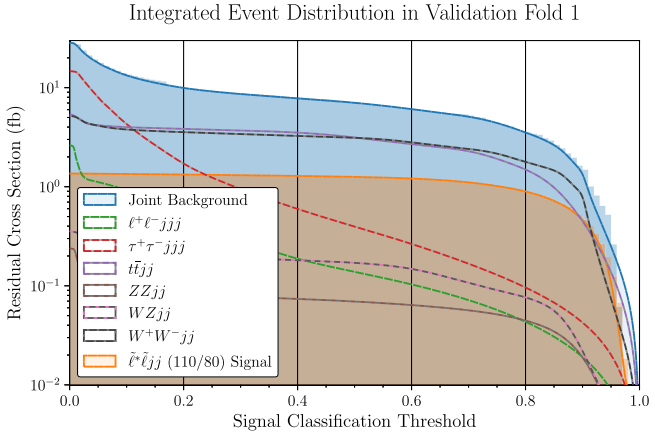

FIG. 2. Plot illustrating the residual cross section for signal and background (as labeled) as a function of the BDT classification score after precuts. Corresponding plots for additional folds of training and evaluation samples can be found in Appendix A.

between each such data "fold." We configure the BDT to build up to 50 trees with a maximal depth of 5 levels per tree. Early stopping is enabled to halt training if the loss function is not improved over the course of the five most recent training epochs, i.e., by the generation of the five most recent trees.

As described in Ref. [42], the loss function for the individual trees of the ensemble is regularized by the quadratic sum of the weights on each leaf multiplied by an "L2" regularization hyperparameter, which we fix as $\lambda = 0.01$. The minimum loss reduction for new splittings in a tree is set to $\gamma = 0$. The contribution of each new tree added to the aggregated prediction of the ensemble is scaled by the learning rate, which we set to $\eta = 0.5$. XGBoost hyperparameters were scanned manually around the values used in the study. One criterion for selection was stability of the results, which do not exhibit strong sensitivity to small variations of these parameters.

We find that certain of the hyperparameters employed by XGBoost have extensive (rather than intensive) scaling with the total sample weight. As such, we find that hyperparameter selections and training outcomes will be less sensitive to variations of the sample size if one normalizes the overall weight to unity in the training. We do this separately for each of the signal and background classes in order to balance the training datasets. Physical per-sample weights are used to interpret the results of training.

Any event analyzed by the BDT is given a continuous decimal score between 0 and 1, with 1 being most signal-like and 0 being most background-like. In Fig. 2, we plot the cross section for signal (orange) and background (as labeled) events to be classified with a score greater than a given threshold. Two similar plots made with a different choice of the training and evaluation samples are given in Appendix A. The far left of the plot (where the signal threshold is zero) corresponds to the case where the BDT results are ignored, and yield the cross section for the signal and each background to pass the precuts. The cross section for signal events to pass the precuts is ∼1.3 fb, indicating that, with an integrated luminosity of ∼300 fb^{-1}, ∼400 signal events are expected to pass the precuts and be analyzed by the BDT. For each of the six backgrounds analyzed, the cross section for events to pass the precuts is in all cases $\lesssim$20 fb, with the largest cross section being for the $\bar{\tau}\tau jjj$ background. In particular, note that the $ZZjj$ and $WZjj$ backgrounds have been effectively eliminated by the precuts; the analysis of the BDT is largely irrelevant for those backgrounds, as $S/B \gg 1$ even before the BDT score is included.

We classify an event as signal-like if its score exceeds a given a threshold, and as background-like if otherwise. Although more complicated methods which do not involve a binary assignment of signal-like or background-like status are possible [44–47], and likely will provide better sensitivity, we will find that this simple method is sufficient to already provide substantial improvement in sensitivity. Moreover, the use of a simple binary assignment will facilitate one of our primary goals, which is understanding the physics considerations which underlie the discrimination of signal from the various backgrounds.

D. Results

Although the $\bar{\tau}\tau jjj$ background is the largest after the precuts, the BDT has little difficulty in discriminating this background from signal. For this background, as well as $\overline{\ell}\ell jjj$, the BDT can effectively suppress the background with little loss in signal cross section (see Fig. 2, for a signal classification threshold $\gtrsim$0.4). The greatest difficulty which the BDT finds is in discriminating the $\bar{t}tjj$ and $WWjj$ backgrounds from signal. We may thus expect that the competition between signal and the $\bar{t}tjj$ and $WWjj$ backgrounds will dominate our sensitivity (and associated signal to background ratio).

We see similarly from Fig. 3 that the BDT provides a strong ability to discriminate signal from all backgrounds, though this discriminating power is weakest for the $\bar{t}tjj$ and $WWjj$ backgrounds. For the $\bar{t}tjj$ and $WWjj$ cases, although the BDT analysis exhibits good discriminating power, the large magnitude of these backgrounds relative to the signal cross section implies that, at best, the ratio of signal-to-background is $\mathcal{O}(1)$. This level of discrimination arises only when the signal classification threshold is high enough (see Fig. 2) that only $\lesssim$150 signal or background events are selected.

We thus see that optimal performance of the BDT implies a reduction of background down to an event rate comparable to signal, with only a factor of ∼3 reduction of the signal event rate. With ∼150 signal events and ∼3 times as many background events, one would expect that a sensitivity of $\gtrsim 6\sigma$ could be obtained with an ∼0.3 signal to background ratio. This intuition is confirmed by the results of Fig. 4, in which we present the number of signal events S (blue), the

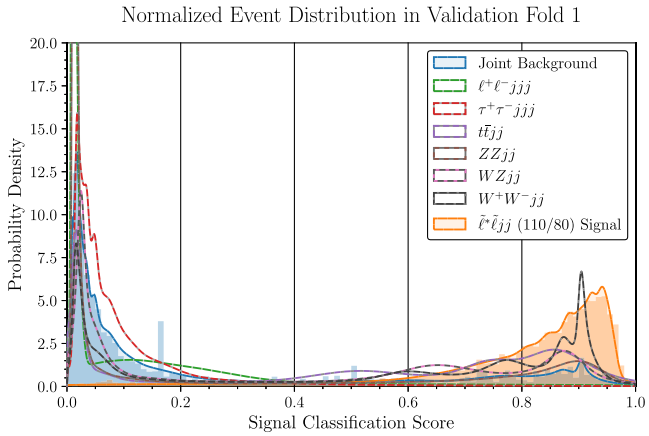

FIG. 3. Density plot representing the separation of signal from each background (as labeled) by the boosted decision tree after precuts. For visual clarity, the densities are separately normalized such that the joint background density is not equivalent to the sum of the individual background densities. Corresponding plots for additional folds of training and evaluation samples can be found in Appendix A.

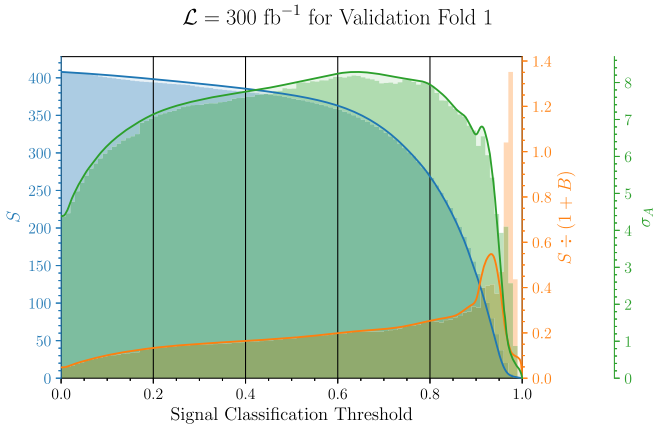

FIG. 4. The number of signal events (S, blue), signal-to-background ration ($S \div (1+B)$, orange) and signal significance (σ_A, green) as a function of the signal classification threshold. Each curve represents an interpolation which smooths out the sharp features in the shaded region, which arise from small event numbers. Corresponding plots for additional folds of training and evaluation samples can be found in Appendix A.

signal-to-background ratio ($S \div (1+B)$, orange), and the signal significance (σ_A, green)[8] as a function of the signal classification threshold, assuming an integrated luminosity of 300 fb^{-1}. Note that shaded regions show sharp features which arise when the number of simulated events which exceed the signal classification threshold is small, and can thus vary dramatically as the threshold increases. These sharp features arise when S is small and do not represent reliable estimates, as they may be artifacts of the size of the simulated dataset. We instead focus on the solid curves, which represent an interpolation of the boundary of the shaded regions which smooths out these sharp features. In Appendix A, the two supplementary plots present analyses with different choices of the sample used for training and evaluation. All three panels are roughly consistent, indicating that our result is robust to variations in the training set. In particular, for a signal classification threshold of ∼0.9, we find that a scenario with $M_{\tilde{\ell}} = 110$ GeV, $M_X = 80$ GeV can be detected with $\gtrsim 6\sigma$ significance and a signal-to-background ratio of ∼0.3, with $S \sim 100$. This represents a marked improvement over the results obtained from the cut-based approach used in [10], in which only 3.0σ sensitivity (with a signal-to-background ratio of 0.2) could be obtained with the same luminosity (see Table III).[9] Moreover, we see that using the cuts imposed in [10] as precuts, with subsequent signal classification by a BDT, does not provide significantly improved sensitivity. Evidently the cut-based approach used in this earlier analysis results in too a large reduction in the number of signal events for optimal signal classification.

It is important to include the effects of systematic uncertainties on the sensitivity of this search. We do not provide a quantitative estimate, as the systematic uncertainties need not be Gaussian. We instead use a qualitative assessment: provided the signal-to-background ratio is substantially larger than the estimated systematic uncertainties, one expects that an underestimated background cannot duplicate the effects of a signal discovered with high statistical significance. An estimate by ATLAS of the background of a slepton search suggests systematic uncertainties at the level of ∼17% [4,8]. As the signal-to-background ratio obtained in this analysis is substantially larger than this ($S/B \sim 0.3$), one expects to be able to detect the presence of signal with sufficient luminosity.

Since the main work of the BDT (beyond the precuts, whose intuition we understand) is in rejecting the $\bar{t}tjj$ and $WWjj$ backgrounds, we can ask which kinematic variables the BDT used in that analysis. These results are shown in Fig. 5 which indicate which kinematic variables dominate the "total gain"[10] when a BDT is trained only against a single background. This figure thus gives an indication of which variables are most useful in discriminating signal from a particular background. In particular, the most

[8]We use the Asimov formula $\sigma_A = \sqrt{2[(S+B)\ln(1+S/B) - S]}$ for the signal significance, which reduces to $\sim S/\sqrt{B}$ for $S/B \ll 1$ [48].

[9]The signal sensitivity estimated in [10] is reported as 4.4σ with a signal-to-background ratio of 0.3. In this work, we have improved the background modeling and statistics, resulting in a lower signal significance and signal-to-background ratio when implementing the same series of cuts proposed for intermediate mass gaps.

[10]This measure of relative importance compares the summed reduction in the loss function that is attributable to leaf splittings associated with each available feature. It is sensitive to how often a feature is used across nodes, how many events pass through each such node, and the weights carried by those events.

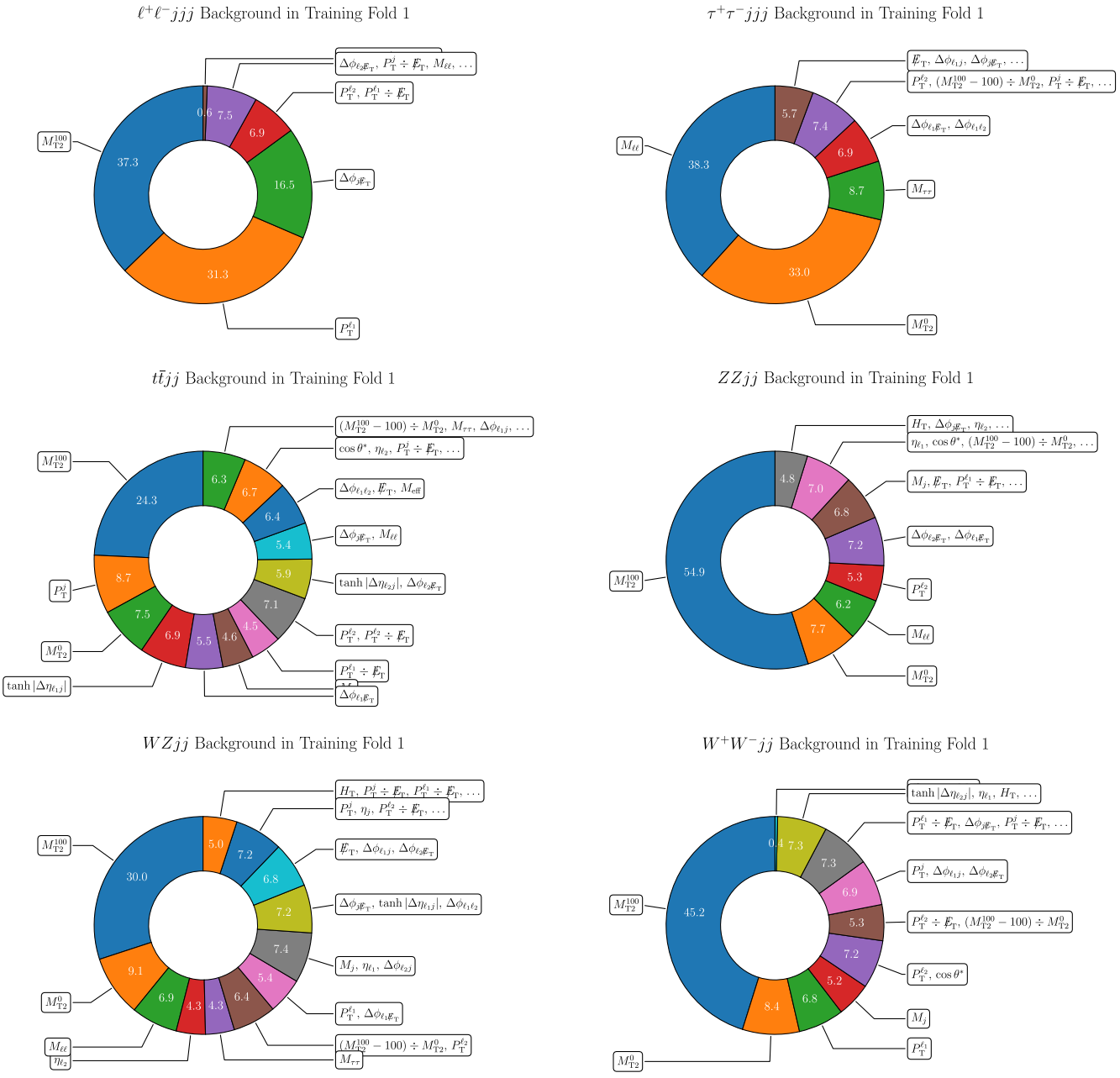

FIG. 5. Plots indicating the relative importance of various kinematic features to the separation of signal from background by the boosted decision tree after level 1 pre-selections, respectively (left-to-right then top-to-bottom) for $\ell\ell jjj$, $\tau\tau jjj$, $t\bar{t}jj$, $ZZjj$, $WZjj$, and $WWjj$. Note that, in each panel, results are shown for a BDT which is trained only with signal and the particular background process in question.

important kinematic variable for rejecting the $ttjj$ and $WWjj$ backgrounds is M_{T2}^{100}, which is defined, as in Ref. [35], as the minimal mass of a parent particle under the assumption that the parent particles are pair produced and each decays to a lepton and an invisible particle with a mass of 100 GeV. Note that this variable was also used in the slepton search analysis presented in Ref. [4], though with a different purpose.

The distributions of M_{T2}^{100} for signal and $\bar{t}tjj$ and $WWjj$ background events (after precuts) are given in Fig. 6. We can see from these distributions that M_{T2}^{100} contributes significant ability to discriminate between signal and background. In particular, background events are biased toward higher values of M_{T2}^{100}, since for background events, the invisible particles are neutrinos which are much lighter than the ansatz of 100 GeV.

TABLE III. Summary of results for our benchmark mass spectrum (i.e., $M_{\tilde{\ell}} = 110$ GeV and $M_{\tilde{\ell}} - M_X = 30$ GeV) after implementing different event selections. For each set of precuts, we show the number of signal events, the signal to background ratio and the signal sensitivity both before and after application of the BDT. When only considering the precuts, we average the respective metrics between the 3 folds of the data. In contrast, the respective BDT signal score thresholds are optimized to show as broad a range of σ_A possible between the different folds while maintaining $S \div (1+B) \geq 0.15$. The first event selection summarizes the results of the event selection with precuts described in Sec. IV C. The second event selection adds a precut on M_{T2}^{100} and the third event selection implements precuts corresponding to that which is implemented for the intermediate mass gap scenarios in previous work [10]. Note the average number of events after precuts between all 3 folds of the data in this final event selection is smaller that the upper end of the range of numbers of signal events after application of the BDT because of the variation between folds at a signal score threshold of 0.00, which is equivalent to only applying the precuts.

	This work		This work + $M_{\mathrm{T2}}^{100} < 130$ GeV		Intermediate mass gaps in [10]	
Event selection	Precuts	BDT	Precuts	BDT	Precuts	BDT
BDT signal threshold	⋯	0.47–0.90	⋯	0.34–0.88	⋯	0.00–0.71
Events at $\mathcal{L} = 300$ fb^{-1}	413	107–386	405	121–387	48	26–51
$S \div (1+B)$	0.05	0.15–0.36	0.06	0.15–0.31	0.20	0.18–0.35
σ_A	4.4	5.8–8.2	4.8	5.6–8.2	3.0	2.5–3.7

However, these distributions do not provide a clear sense of how M_{T2}^{100} is correlated with other variables, which is the type of information one would expect to be learned by a BDT. To consider this question, we assume that the only relevant background processes are $ttjj$ and $WWjj$. That is, we ignore the $WZjj$ and $ZZjj$ background classes (which are essentially eliminated by the precuts) and the $\mu\mu jjj$ and $\tau\tau jjj$ background classes (which are easily eliminated by the BDT). After the precuts, we see from Fig. 3 that we are left with ∼400 signal events, and ∼3300 background, roughly evenly split between $ttjj$ and $WWjj$. We then consider a simple cut in which we accept only events with $M_{\mathrm{T2}}^{100} < 130$ GeV. This cut admits almost all signal events, but reduces each class of background events by roughly $1/3$. This simple cut-and-count analysis would yield a signal-to-background ratio of $S/B \sim 0.18$, with a signal significance of $\sim 8.5\sigma$. Indeed, we see that a $1/3$ reduction in background, with a small reduction in signal, is roughly what the BDT achieves with a signal classification threshold set at ∼0.6 (see Fig. 2), yielding approximately the significance and signal-to-background ratio we estimated with a cut-and-count analysis (see Fig. 4). The deeper correlations found by the BDT (at least in regard to the $ttjj$ and $WWjj$ backgrounds) lead to an improvement in the signal-to-background ratio of roughly a factor of 2 as the signal classification threshold is raised to > 0.9. Note that this improvement in the signal-to-background ratio is essential for overcoming systematic uncertainties, as described above.

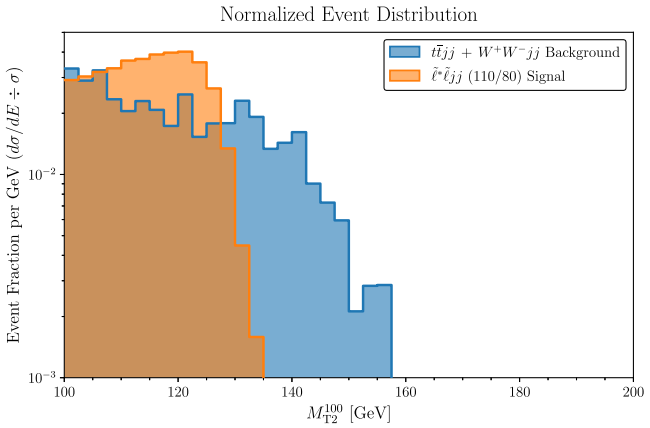

FIG. 6. The distribution of M_{T2}^{100} for events passing the precuts for signal (orange), as well as the combined $ttjj$ and $WWjj$ backgrounds (blue).

Since the imposition of a simple cut such as $M_{\mathrm{T2}}^{100} < 130$ GeV seems to replicate at least part of the BDT analysis, one might ask if imposing this as a precut could improve the performance of the BDT, by allowing it to focus on less obvious correlations. To test this possibility, we ran a second analysis in which $M_{\mathrm{T2}}^{100} < 130$ GeV was imposed as a additional precut. The results are fairly similar to analysis without this additional precut (see Table III). It appears that, because the $ttjj$ and $WWjj$ backgrounds are not excessively large compared to the other backgrounds and are within an order of magnitude of the signal, after applying the precuts described in Sec. IV C, the BDT was not overly focused on reducing them, and thus did not gain significantly in performance when the $ttjj$ and $WWjj$ backgrounds were reduced through an additional precut.

E. Other benchmarks

Until now, we have considered a single benchmark parameter point: $M_{\tilde{\ell}} = 110$ GeV, $M_X = 80$ GeV. Here

TABLE IV. Summary of results for a variety of different mass spectra. For each combination of scalar mass $M_{\tilde{\ell}}$ and mass splitting $\Delta M = M_{\tilde{\ell}} - M_X$, we show (from top to bottom) the optimal signal score threshold for the BDT, the number of signal events expected at $\mathcal{L} = 300$ fb^{-1}, $S \div (1+B)$ and σ_A. For each case, we test the BDT trained on the benchmark spectrum (i.e., $M_{\tilde{\ell}} = 110$ GeV and $\Delta M = 30$ GeV) and the BDT retrained specifically for each spectrum. The respective BDT signal score thresholds are optimized to show as broad a range of σ_A possible while maintaining $S \div (1+B) \geq 0.15$.

	$M_{\tilde{\ell}} = 110$ GeV		$M_{\tilde{\ell}} = 160$ GeV		$M_{\tilde{\ell}} = 300$ GeV	
	Benchmark	Retrained	Benchmark	Retrained	Benchmark	Retrained
$\Delta M = 30$ GeV	0.47–0.90	⋯	0.91–0.92	0.90–0.94	⋯	⋯
	107–386	⋯	33–55	13–43	⋯	⋯
	0.15–0.36	⋯	0.18–0.32	0.15–0.39	< 0.15	< 0.15
	5.8–8.2	⋯	2.5–3.1	1.8–2.9	⋯	⋯
$\Delta M = 40$ GeV	0.47–0.91	0.23–0.87	⋯	0.90–0.92	⋯	⋯
	68–375	163–496	⋯	24–49	⋯	⋯
	0.15–0.36	0.15–0.35	< 0.15	0.15–0.25	< 0.15	< 0.15
	4.2–7.9	6.7–9.1	⋯	1.9–3.0	⋯	⋯
$\Delta M = 50$ GeV	0.59–0.90	0.13–0.88	⋯	0.79–0.93	⋯	⋯
	67–320	172–575	⋯	38–156	⋯	⋯
	0.15–0.23	0.15–0.42	< 0.15	0.15–0.35	< 0.15	< 0.15
	3.7–7.3	7.6–10.7	⋯	3.0–5.4	⋯	⋯
$\Delta M = 60$ GeV	0.62–0.90	0.05–0.90	⋯	0.54–0.96	⋯	0.96
	60–295	252–691	⋯	21–273	⋯	14–16
	0.15–0.21	0.15–1.01	< 0.15	0.15–1.16	< 0.15	0.15–0.26
	3.3–6.9	9.9–14.9	⋯	3.6–7.9	⋯	1.5–1.9

we consider the sensitivity of an LHC search with 300 fb^{-1} integrated luminosity to models with different choices for the masses. In Table IV, we consider $M_{\tilde{\ell}} =$ 110,160,300 GeV, with $M_{\tilde{\ell}} - M_X = 30,40,50,60$ GeV. In each case, we either use BDT trained against the original benchmark model, or against the model to be analyzed (in all cases, the original precuts are used). For each analysis, we present an optimal range for the signal classification thresholds, and the range of the number of signal events, signal to background ratio, and signal significance as one varies the signal classification threshold within its optimal range. We only present results for $S \div (1+B) \geq 0.15$, in order to ensure that the analysis is robust against likely systematic uncertainties [4,8].

For cases with $M_{\tilde{\ell}} = 110$ GeV and larger mass gaps, $M_{\tilde{\ell}} - M_X > 30$ GeV, we see that the BDT trained on the benchmark spectrum, $M_{\tilde{\ell}} = 110$ GeV and $M_{\tilde{\ell}} - M_X = 30$ GeV, maintains robust sensitivity to scalar muon production even with some degradation as the mass gaps become larger. Alternatively, when training the BDT individually for each benchmark, we see that the sensitivity is enhanced at larger mass splittings due to the improved separation of signal from background in the kinematic distributions. Although mass gaps $M_{\tilde{\ell}} - M_X = 50, 60$ GeV are currently excluded for $M_{\tilde{\ell}} = 110$ GeV, the signal significance we project for the BDT analysis of these cases is considerably larger than in the most stringent LHC constraints from Ref. [4], which exclude mass splittings of $\sim$50 GeV at $\sim 2\sigma$. Similar dependencies of the sensitivity in the BDT analysis on the mass splitting can be seen for $M_{\tilde{\ell}} = 160$ GeV, however the BDT trained on the benchmark mass spectrum is only sensitive to the scenario with the same mass splitting of $M_{\tilde{\ell}} - M_X = 30$ GeV. When the BDT is retrained for each mass spectrum, we see the BDT could be sensitive to the full range of mass splittings from 30 GeV to 60 GeV, none of which are currently constrained by LHC searches.

For $M_{\tilde{\ell}} = 300$ GeV, we never find $S \div (1+B)$ much larger than 0.15, implying that a BDT search using these precuts might be difficult even at higher luminosity. But it is worth noting that, for such a large lepton partner mass, the cross section for signal events to pass the precuts is much smaller than each of the backgrounds, with the total background well over two orders of magnitude larger than the signal, after imposing the precuts. The precuts were imposed precisely to eliminate this hierarchy, and we see that the precuts chosen to be adequate for $M_{\tilde{\ell}} = 110$ GeV do not achieve this purpose if the lepton partner mass is much heavier.

Better prospects would likely lie in developing more aggressive precuts to study this higher mass range (likely using higher luminosity). But training a BDT after the imposition of more aggressive precuts would require a much larger sample of simulated events. The difficulty lies in ensuring that the training dataset adequately samples the tails of the kinematic distributions, where it may be most difficult to distinguish signal events from background events. As one imposes more aggressive precuts, one will

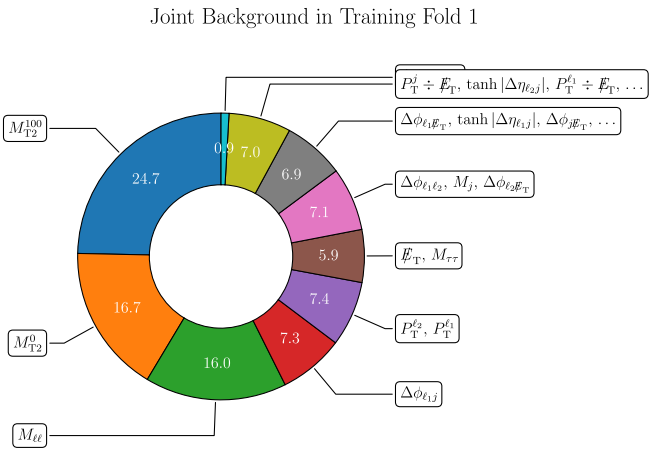

FIG. 7. Plot indicating the relative importance of various kinematic features to the separation of signal from the joint (all source) background. This training occurs application of the Sec. IV C precuts.

find that fewer events in the training sample will pass the precuts. As a result, one may find that at the tails of the kinematic distributions, training is dominated by only a few events, leading to unreliable BDT performance which is biased by the idiosyncracies of the training dataset. To apply precuts designed for much higher luminosities while avoiding this difficulty, one should generate a much more extensive set of simulated data. Although that is beyond the scope of this work, it would be an interesting topic of future study.

F. Restricted training

In this subsection we assess the question of how much classification power remains if one restricts the set of available variables for training. Specifically, we start from the precuts described in Sec. IV C, while passing only a portion of the variables in Table II to the BDT. Specifically, we test restricted training for just the most effective variables (as indicated by averaging feature importance to total gain over the three folds for training of signal against the joint background) and also for a selection of "low-level" variables (including E_{T}, H_{T}, the transverse momentum values, as well as the azimuthal and rapidity separation variables).

Figure 7 illustrates the feature importance (for fold 1) with joint training against all backgrounds after the application of precuts. We use these results (including folds 2 and 3) to select the top performing variables for restricted training. The top performing variables for joint training after precuts are all high-level variables associated with reconstructed masses (or lower mass bounds), namely M_{T2}^{100}, M_{T2}^{0}, and $M_{\ell\ell}$, in that order. These three features account for more than half of the total gain, and are important across all three training folds. By contrast, no other feature has an average importance above about 6%, and more variability is observed across the folds. For these reasons, we include just these three features in the restricted training for top-performing variables. This has the added benefit of precluding any direct overlap with the restricted training on low-level features.

We find that the best results are indeed obtained using the full ensemble of both high- and low-level observables. The peak signal-to-background ratio (using the interpolated measure) is reduced by about 40% for each of the restricted trainings relative to the case of training on all variables. In terms of statistical significance, the high-level variables almost match the performance of the full set, whereas the low-level observables give up about 40% in this regard as well. Since the success of this analysis hinges critically on elevating the signal-to-background ratio above the systematic uncertainty noise floor, it becomes all the more important to supplement the top-performing high-level observables with the larger number of weaker correlations provided by the low-level observables.

V. CONVENTIONAL WISDOM FOR BDTS

There are several lessons that have been learned during the course of this study that would seem to represent a type of conventional wisdom which is substantially transferable in other similar applications.

The approach described in Sec. II works reasonably well "out of the box," but we have observed that it is possible to achieve significantly better separation between signal and background using a hybrid technique that involves significant human input in conjunction with training of the BDT. By contrast, the emerging understanding in the deep-learning community is that one should generally just "get out of the way" and let the training proceed with minimal supervision and data curation. However, that intuition is apparently less applicable in the context of more basic types of machine learning applications such as boosted decision trees. The upside is that this additional investment of effort comes with a very substantially expanded and auditable inventory of precisely what and how the machine was learning, while still typically accessing a level of feature separation that meaningfully exceeds what is available with manual feature selections after maximal effort.

In a similar vein, the BDT seems to take maximal advantage of high-level features that have been constructed to efficiently encode physics features that are expected to be relevant. Again, by contrast, deep-learning approaches are known to work exceptionally well on raw low-level data, and extensive human curation is generally not necessary, or may even be detrimental. This is because a neural network of sufficient complexity can

essentially mimic any mathematical transformation, and training will drive the development of this function in an optimized way. By contrast, shallow decision trees are always splitting on single features, and they tend to operate most effectively when the available features are preconfigured to very densely encode the most relevant parameterization of available information. For example, we generally find that reconstructed masses and composite constructions like M_{T2} or $\cos\theta^*$ pack much more punch than isolated kinematic features. Even simple operations, like taking differences or ratios of scales or angular coordinates that are expected to have more discriminating power as relative values than absolute values can increase efficiency.

Additionally, a very deep layer of background will be quite difficult for the BDT to "see through," and it is helpful to remove "obvious" backgrounds prior to engaging the BDT. Along these lines, it is often the case that the most dominant backgrounds can be the simplest to defeat, and it seems that this should always be done up front when it is possible. For example, the cross section of single Z events is extremely large, and its magnitude will typically overwhelm any search for final states that feature opposite-sign same-flavor dileptons. However, the vast majority of these dilepton events will cluster around the resonant Z mass, and the residual cross section can be reduced dramatically with a simple window cut exclusion. Of course, the BDT can learn such a cut on its own. However, it seems to become "exhausted" in the process, and it becomes much harder for it to see the more subtle separations which are subsequently critical to achieving a successfully refined training if the dominant fraction of the input event weights are telling the BDT to pay attention to the mass window as a first priority. Not only should "obvious" discriminants be applied as manual precuts, it is advisable that the hardness of these cuts be selected so that the residual cross section associated with each type of background are brought approximately into parity.

On a related note, for practical reasons described previously, the starting cross section of signal events is always much smaller than that of relevant backgrounds. We find that the behavior of the BDT will be much more uniform and predictable if one separately normalizes the sum of signal and background weights to unity prior to training. This is because several of the hyperparameters associated with the BDT implementation, in particular those associated with regulation of the objective, have extensive (rather than intensive) scaling properties. This normalization helps to ensure that intuition developed regarding useful settings of these parameters has improved cross-applicability to other contexts. In fact, useful hyperparameter values generically tend to be "order one" in this normalization. Of course, one must then rescale back to physical cross sections when making predictions for rates in a real world experimental environment.

Since we are only really interested here in classification of categories that are somewhat difficult to separate by elementary means, the BDT will presumably need to access rather narrow and subtle features in order to train successfully. Likewise, since we are only really interested in enhancing the visibility of extremely rare processes, it is necessary to filter away the vast majority of competing background processes. These realities would appear to usher in a fundamental dilemma common to all relevant classification problems in collider physics. This is that the training is only successful in the regime where it has become statistically limited, and therefore less reliable. In other words, "harder cuts" leave fewer events, and it becomes successively more difficult to validate that the selected cuts are generalizable. This concern must be balanced against the priorities identified previously, such that any precuts will not preclude effective training. Specifically, there must be a statistically significant sampling of events on hand for the BDT to process. Of course, if it is possible to reduce all relevant background via manual selections, then it may not be necessary to further employ machine learning at all. However, overly aggressive precuts can additionally prevent the BDT from acting in a more surgical manner, and potentially achieving comparable background reductions while retaining a greater density of signal. Although the BDT is technically classifying rather than cutting, the same dilemma ultimately applies to its training process as well, and one often in practice contemplates an effective selection cut on some threshold of the resulting classification score.

VI. CONCLUSION

We have considered the LHC sensitivity to a scenario of scalar lepton partner pair-production, with each partner decaying to a muon and an invisible particle with an intermediate mass splitting ($M_{\tilde{\ell}} - M_X \sim 30$ GeV). This is a scenario for which current LHC sensitivity has not exceeded LEP, owing to the large electroweak background. We have used an analysis based on a boosted decision tree (BDT), and have found that with 300 fb^{-1} luminosity, the LHC could achieve $\gtrsim 5\sigma$ sensitivity to currently allowed models ($M_{\tilde{\ell}} \sim 110$ GeV), with a signal to background ratio $\gtrsim 0.3$. The BDT analysis could also exclude spectra with larger mass gaps $M_{\tilde{\ell}} - M_X \sim 50$ GeV and the same scalar mass with a significance $\gtrsim 4$ times larger than the most recent LHC analysis [4]. With the same luminosity, LHC could exclude models with $M_{\tilde{\ell}}$ as large as 160 GeV. But for a larger mass splitting ($\sim$60 GeV), the LHC could provide $> 5\sigma$ evidence for models with $M_{\tilde{\ell}} \sim 160$ GeV (also allowed by analyses of current data). The projected sensitivity of cut-based analyses in previous theoretical studies suggested that LHC could only be sensitive to models with mass splittings $\sim$30–60 GeV for scalar masses

not much large than ∼110 GeV [10]. Thus, a BDT analysis in scalar lepton partner searches could definitively probe realizations of the MSSM in which scalar muons with $M_{\tilde{\ell}} \lesssim 150$ GeV [1] mediate interactions that can both deplete the relic density through dark matter annihilation and account for the anomalous magnetic moment of the muon.

The most difficult backgrounds to discriminate are $\bar{t}t$ and $WW +$ jets. The kinematic variable relied on most heavily by the BDT in discriminating signal from these background is M_{T2}^{100}. Nevertheless, the BDT makes good use of less obvious variables to simultaneously suppress the $\mu\mu +$ jets and $\tau\tau +$ jets backgrounds while further reducing $\bar{t}t$ and $WW +$ jets, producing an optimal signal-to-background ratio which is a factor of ∼6 better than one would get by simply imposing an additional cut on M_{T2}^{100}.

In performing this analysis, we have learned several lessons regarding the application of BDTs to LHC analyses which may be applied to other searches. In particular, we found that the BDT is much more effective if obvious precuts are applied which reduce each surviving background to roughly equal event numbers, so that the BDT is not overly dominated by any one background class. Essentially, there is little gained by having a BDT relearn things we already know. On the other hand, while the imposition of precuts which reduce very large backgrounds improves the BDT performance, the application of even obvious precuts which have the effect of removing backgrounds which are only comparable to the signal has little effect on the BDT's performance.

There any many avenues open for further study. In particular, it would be interesting to explore the optimal tradeoff between using precuts to reduce backgrounds before the application of a BDT, as opposed to simply allowing a BDT to analyze as large a dataset as possible. It would also be interesting to investigate other approaches to the problem of backgrounds with widely disparate cross sections, such as changes to the BDT hyperparameters.

We have found that it is difficult to properly train a BDT if the most difficult backgrounds to remove are swamped by larger (though more tractable) backgrounds. The essential problem is that the simulation data which one uses for training are necessarily only a fraction of the vast LHC dataset. A subleading background may be undersampled, leading a BDT to focus its training on quirks of the training set, rather than robust features. But this is a more general danger which could prove challenging for similar applications of other machine learning techniques. It is difficult to ensure that the toughest backgrounds are sufficiently well sampled in simulation if one has not already characterized the backgrounds which are easy or difficult to discriminate. It would be interesting to explore the importance of sufficient background sampling in the training of deep learning algorithms such as neural networks.

Our analysis has in a sense been optimized for the intermediate mass gap range, as we have trained the BDT with a single BSM model with mass splitting of 30 GeV. Unsurprisingly, a BDT trained on only this BSM model does far worse at identifying the presence of new physics if the mass splitting is either significantly larger or smaller, because of the changes to the underlying features of the kinematic distributions which drive this sensitivity (see, for example [10]). It would be interesting to train a BDT with a range of mass splittings, to determine the trade-off between efficiency and generality of the analysis. In connection with this question, it would also be interesting to consider alternative choices in the training procedure. For example, one could consider training BDTs separately to distinguish signal from each of the leading backgrounds, with the individual scores combined to yield an overall signal classifier. In this work, we used a signal classifier threshold to make a binary classification of an event as either signal or background, but it would be interesting to consider alternative approaches.

ACKNOWLEDGMENTS

We are grateful to Alexander Khanov, Xerxes Tata, and Evelyn Thomson for useful discussions. We thank Jordan Pittman, Stephanie Samperio, Conrad Albrecht, and Joseph Hansen for technical assistance. For facilitating portions of this research, B. Dutta, J. Kumar and J. W. Walker wish to acknowledge the Center for Theoretical Underground Physics and Related Areas (CETUP*), The Institute for Underground Science at Sanford Underground Research Facility (SURF), and the South Dakota Science and Technology Authority for hospitality and financial support, as well as for providing a stimulating environment. B. D. is supported in part by DOE Grant No. DE-SC0010813. T. G. is supported by the funding available from the Department of Atomic Energy (DAE), Government of India for Harish-Chandra Research Institute (HRI). J. K. would like to thank the organizers of SUSY 2023 for their hospitality. J. K. is supported in part by DOE Grant No. DE-SC0010504. P. Sandick is supported in part by NSF Grant No. PHY-2014075. P. Stengel would like to thank the organizers of Cosmology 2023 in Miramare for their hospitality. P. Stengel is funded by the Instituto Nazionale di Fisica Nucleare (INFN) through the project of the InDark INFN Special Initiative: "Neutrinos and other light relics in view of future cosmological observations" (n. 23590/2021). J. W. W. is supported in part by the National Science Foundation under Grant No. NSF PHY-2112799.

APPENDIX A: ADDITIONAL VALIDATION FOLDS

Plots representing additional validation folds with different selections of the events used for training and evaluation are provided here. Figure 8 shows the residual signal and

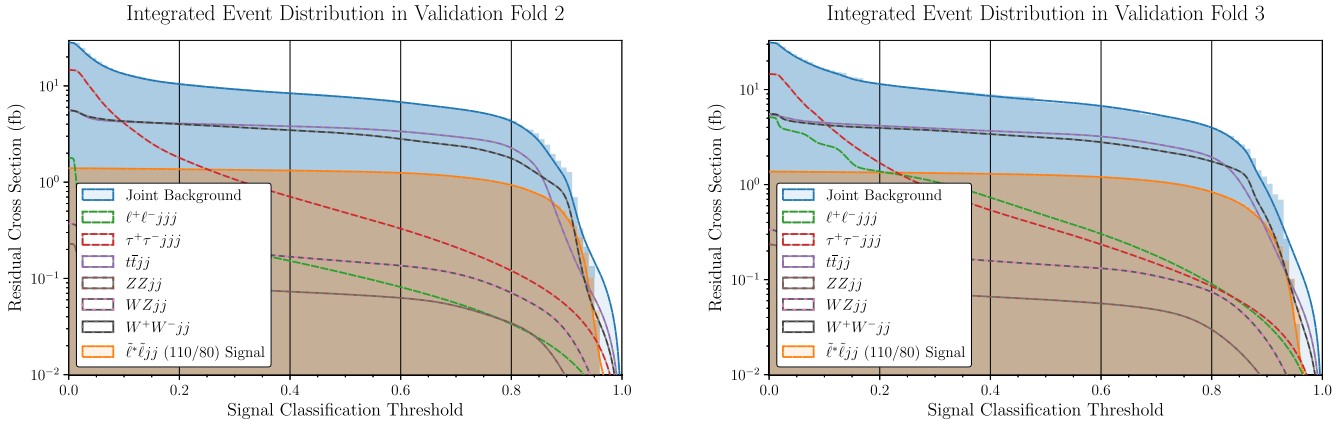

FIG. 8. Same as Fig. 2, but for different choices of training and evaluation samples. Plots illustrating the residual cross section for signal and background (as labeled) as a function of the BDT classification score after precuts.

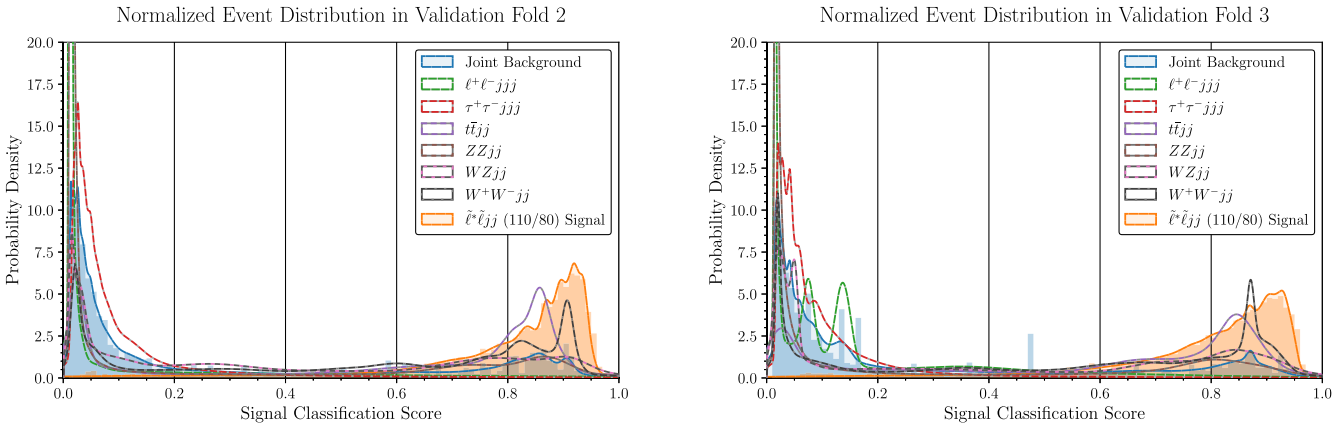

FIG. 9. Same as Fig. 3, but for different choices of training and evaluation samples. Density plots representing the separation of signal from each background (as labeled) by the boosted decision tree after precuts. For visual clarity, the densities are separately normalized such that the joint background density is not equivalent to the sum of the individual background densities.

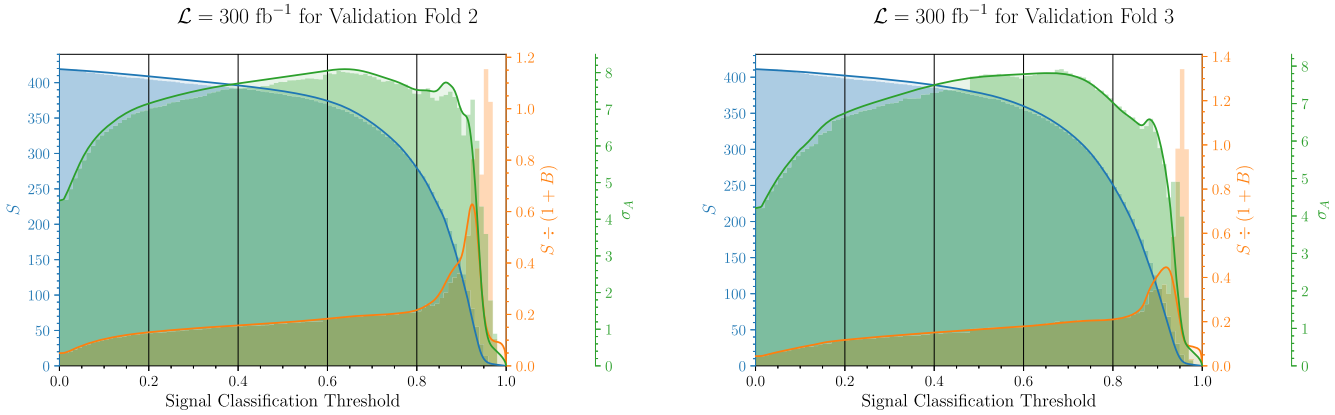

FIG. 10. Same as Fig. 4, but for different choices of training and evaluation samples. The number of signal events (S, blue), signal-to-background ration ($S \div (1+B)$, orange) and signal significance (σ_A, green) as a function of the signal classification threshold. Each curve represents an interpolation which smooths out the sharp features in the shaded region, which arise from small simulated event numbers.

background cross-section as a function of the classification score threshold. Figure 9 shows the probability density distribution of the classification score for signal and background samples. Figure 10 shows the evolution of the signal event count, signal to background ratio, and signal significance as a function of the classification score threshold.

APPENDIX B: $M_{\tau\tau}$

In the process $pp \to Zj \to j\bar{\tau}\tau$ the final state has two boosted τs with $(P^{\tau_1}+P^{\tau_2})^2 = M_Z^2$, and $\vec{P}_T^{\tau_1} + \vec{P}_T^{\tau_2} + \vec{P}_T^{j} = 0$. If each τ decays leptonically, with all products emitted in the approximately forward direction, then the observable particles will be two leptons, $\ell_{1,2}$, satisfying $\vec{P}_T^{\tau_i} = (1+\xi_i)\vec{P}_T^{\ell_i}$, where ξ_i is the ratio of momentum carried by the pair of nearly collinear neutrinos to that of the visible lepton. Since the $\vec{P}_T^{j,\ell_1,\ell_2}$ are observable, conservation of transverse momentum gives two equations for two unknowns: $(1+\xi_1)\vec{P}_T^{\ell_1} + (1+\xi_2)\vec{P}_T^{\ell_2} + \vec{P}_T^{j} = 0$. One generally expects that one can solve for $\xi_{1,2}$, allowing one to reconstruct the mass of the parent Z-boson using $M_Z^2 = (1+\xi_1)(1+\xi_2)M_{\ell\ell}^2$, where $M_{\ell\ell}$ is the dilepton invariant mass. Specifically, following the discussion in [10], we demonstrate here that $M_{\tau\tau}$ from Eq. (1) does indeed reconstruct the Z-boson mass under these circumstances. Note that the conservation law implies $\vec{P}_T^{\tau_1} \times \vec{P}_T^{j} = -\vec{P}_T^{\tau_2} \times \vec{P}_T^{j} = -\vec{P}_T^{\tau_1} \times \vec{P}_T^{\tau_2}$. From this, we have the following:

$$\begin{aligned} M_{\tau\tau}^2 &= -M_{\ell\ell}^2 \frac{(\vec{P}_T^{\ell_1} \times \vec{P}_T^{j}) \cdot (\vec{P}_T^{\ell_2} \times \vec{P}_T^{j})}{|\vec{P}_T^{\ell_1} \times \vec{P}_T^{\ell_2}|^2} \\ &= -M_{\ell\ell}^2 (1+\xi_1)(1+\xi_2) \frac{(\vec{P}_T^{\tau_1} \times \vec{P}_T^{j}) \cdot (\vec{P}_T^{\tau_2} \times \vec{P}_T^{j})}{|\vec{P}_T^{\tau_1} \times \vec{P}_T^{\tau_2}|^2} \\ &= M_{\ell\ell}^2 (1+\xi_1)(1+\xi_2) \frac{(\vec{P}_T^{\tau_1} \times \vec{P}_T^{\tau_2}) \cdot (\vec{P}_T^{\tau_1} \times \vec{P}_T^{\tau_2})}{|\vec{P}_T^{\tau_1} \times \vec{P}_T^{\tau_2}|^2} \\ &= M_{\ell\ell}^2 (1+\xi_1)(1+\xi_2) = M_Z^2 \end{aligned} \tag{B1}$$

[1] K. Fukushima, C. Kelso, J. Kumar, P. Sandick, and T. Yamamoto, MSSM dark matter and a light slepton sector: The incredible bulk, Phys. Rev. D **90**, 095007 (2014).

[2] J. T. Acuña, P. Stengel, and P. Ullio, Minimal dark matter model for muon g-2 with scalar lepton partners up to the TeV scale, Phys. Rev. D **105**, 075007 (2022).

[3] J. T. Acuña, P. Stengel, and P. Ullio, Neutron star heating in dark matter models for muon g-2 with scalar lepton partners up to the TeV scale, arXiv:2209.12552.

[4] ATLAS Collaboration, Search for direct pair production of sleptons and charginos decaying to two leptons and neutralinos with mass splittings near the W-boson mass in $\sqrt{s} = 13$ TeV pp collisions with the ATLAS detector, J. High Energy Phys. 06 (2023) 031.

[5] A. M. Sirunyan *et al.* (CMS Collaboration), Search for supersymmetric partners of electrons and muons in proton-proton collisions at $\sqrt{s} = 13$ TeV, Phys. Lett. B **790**, 140 (2019).

[6] LEPSUSYWG, ALEPH, DELPHI, L3, OPAL Collaborations, Combined LEP Selectron/Smuon/Stau Results, 183–208 GeV, Report No. LEPSUSYWG/04-01.1, 2004, https://lepsusy.web.cern.ch/lepsusy/www/sleptons_summer04/slep_final.html.

[7] G. Aad *et al.* (ATLAS Collaboration), Search for electroweak production of charginos and sleptons decaying into final states with two leptons and missing transverse momentum in $\sqrt{s} = 13$ TeV pp collisions using the ATLAS detector, Eur. Phys. J. C **80**, 123 (2020).

[8] G. Aad *et al.* (ATLAS Collaboration), Searches for electroweak production of supersymmetric particles with compressed mass spectra in $\sqrt{s} = 13$ TeV pp collisions with the ATLAS detector, Phys. Rev. D **101**, 052005 (2020).

[9] ATLAS Collaboration, Identification of very-low transverse momentum muons in the ATLAS experiment, Report No. ATL-PHYS-PUB-2020-002.

[10] B. Dutta, K. Fantahun, A. Fernando, T. Ghosh, J. Kumar, P. Sandick, P. Stengel, and J. W. Walker, Probing squeezed bino-slepton spectra with the Large Hadron Collider, Phys. Rev. D **96**, 075037 (2017).

[11] J. Friedman, T. Hastie, and R. Tibshirani, Special invited paper. Additive logistic regression: A statistical view of boosting, Ann. Stat. **28**, 337 (2000).

[12] J. H. Friedman, Greedy function approximation: A gradient boosting machine, Ann. Stat. **29**, 1189 (2001).

[13] P. Baldi, P. Sadowski, and D. Whiteson, Searching for exotic particles in high-energy physics with deep learning, Nat. Commun. **5**, 4308 (2014).

[14] A. Dey, J. Lahiri, and B. Mukhopadhyaya, LHC signals of a heavy doublet Higgs as dark matter portal: Cut-based approach and improvement with gradient boosting and neural networks, J. High Energy Phys. 09 (2019) 004.

[15] C. K. Khosa, V. Sanz, and M. Soughton, Using machine learning to disentangle LHC signatures of dark matter candidates, SciPost Phys. **10**, 151 (2021).

[16] E. Arganda, A. D. Medina, A. D. Perez, and A. Szynkman, Towards a method to anticipate dark matter signals with deep learning at the LHC, SciPost Phys. **12**, 063 (2022).

[17] T. Finke, M. Krämer, M. Lipp, and A. Mück, Boosting mono-jet searches with model-agnostic machine learning, J. High Energy Phys. 08 (2022) 015.

[18] B. M. Dillon, R. Mastandrea, and B. Nachman, Self-supervised anomaly detection for new physics, Phys. Rev. D **106**, 056005 (2022).
[19] V. K. MB, A. K. Nayak, and A. K. Radhakrishnan, Invariant mass reconstruction of heavy gauge bosons decaying to τ leptons using machine learning techniques, Eur. Phys. J. C **84**, 219 (2024).
[20] T. Ghosh, C. Kelso, J. Kumar, P. Sandick, and P. Stengel, Simplified dark matter models with charged mediators, arXiv:2203.08107.
[21] A. Das and P. Rad, Opportunities and challenges in explainable artificial intelligence (XAI): A survey, arXiv:2006.11371.
[22] B. C. Allanach, C. Balazs, G. Belanger, M. Bernhardt, F. Boudjema, D. Choudhury, K. Desch, U. Ellwanger, P. Gambino, R. Godbole *et al.*, SUSY Les Houches Accord 2, Comput. Phys. Commun. **180**, 8 (2009).
[23] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations, J. High Energy Phys. 07 (2014) 079.
[24] A. Djouadi, M. M. Muhlleitner, and M. Spira, Decays of supersymmetric particles: The Program SUSY-HIT (SUspect-SdecaY-Hdecay-InTerface), Acta Phys. Pol. B **38**, 635 (2007), https://doi.org/10.48550/arXiv.hep-ph/0609292.
[25] S. Ashanujjaman and S. P. Maharathy, Probing compressed mass spectra in the type-II seesaw model at the LHC, Phys. Rev. D **107**, 115026 (2023).
[26] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, An introduction to PYTHIA8.2, Comput. Phys. Commun. **191**, 159 (2015).
[27] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi (DELPHES 3 Collaboration), DELPHES3, A modular framework for fast simulation of a generic collider experiment, J. High Energy Phys. 02 (2014) 057.
[28] ATLAS Collaboration, Tagging and suppression of pileup jets, Report No. ATLAS-CONF-2014-018.
[29] A. M. Sirunyan *et al.* (CMS Collaboration), Measurement of the $t\bar{t}$ production cross section using events with one lepton and at least one jet in pp collisions at $\sqrt{s} = 13$ TeV, J. High Energy Phys. 09 (2017) 051.
[30] A. M. Sirunyan *et al.* (CMS Collaboration), Measurement of the production cross section for single top quarks in association with W bosons in proton-proton collisions at $\sqrt{s} = 13$ TeV, J. High Energy Phys. 10 (2018) 117.
[31] J. W. Walker, AEACuS, RHADAManTHUS, and MInOS: Automated Tools for Collider Event Analysis, Plotting, and Machine Learning, (2023), https://github.com/joelwwalker/AEACuS.
[32] J. Walker, Automated collider event selection, plotting, & machine learning with AEACuS, RHADAManTHUS, & MInOS, *Proc. Sci.*, CompTools2021 (2022) 027.
[33] A. Barr, Measuring slepton spin at the LHC, J. High Energy Phys. 02 (2006) 042.
[34] K. Matchev *et al.*, Measuring slepton spin at the LHC, J. High Energy Phys. 11 (2012) 006.
[35] C. G. Lester and D. J. Summers, Measuring masses of semiinvisibly decaying particles pair produced at hadron colliders, Phys. Lett. B **463**, 99 (1999).
[36] H. C. Cheng and Z. Han, Minimal kinematic constraints and m(T2), J. High Energy Phys. 12 (2008) 063.
[37] A. J. Barr, B. Gripaios, and C. G. Lester, Transverse masses and kinematic constraints: From the boundary to the crease, J. High Energy Phys. 11 (2009) 096.
[38] Z. Han and Y. Liu, MT2 to the rescue—searching for sleptons in compressed spectra at the LHC, Phys. Rev. D **92**, 015010 (2015).
[39] R. K. Ellis, I. Hinchliffe, M. Soldate, and J. J. van der Bij, Higgs decay to $\tau^+\tau^-$: A possible signature of intermediate mass Higgs bosons at the SSC, Nucl. Phys. **B297**, 221 (1988).
[40] G. Aad *et al.* (ATLAS Collaboration), Measurement of the $t\bar{t}t\bar{t}$ production cross section in pp collisions at $\sqrt{s} =$ 13 TeV with the ATLAS detector, J. High Energy Phys. 11 (2021) 118.
[41] ATLAS Collaboration, Trigger menu in 2018, Report No. ATL-DAQ-PUB-2019-001.
[42] T. Chen and G. Carlos, XGBoost: A scalable tree boosting system, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York, 2016), pp. 785–794.
[43] J. D. Hunter, MATPLOTLIB: A 2D graphics environment, Comput. Sci. Eng. **9**, 90 (2007).
[44] S. Whiteson and D. Whiteson, Machine learning for event selection in high energy physics, Eng. Appl. Artif. Intell. **22**, 1203 (2023).
[45] C. W. Murphy, Class imbalance techniques for high energy physics, SciPost Phys. **7**, 076 (2019).
[46] E. Arganda, A. D. Perez, M. de los Rios, and R. M. Sandá Seoane, Machine-learned exclusion limits without binning, Eur. Phys. J. C **83**, 1158 (2023).
[47] C. K. Khosa, V. Sanz, and M. Soughton, A simple guide from machine learning outputs to statistical criteria in particle physics, SciPost Phys. Core **5**, 050 (2022).
[48] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, Asymptotic formulae for likelihood-based tests of new physics, Eur. Phys. J. C **71**, 1554 (2011); **73**, 2501(E) (2013).