THE EUROPEAN
PHYSICAL JOURNAL PLUS

Adversarially Learned Anomaly Detection on CMS open data: re-discovering the top quark

O. Knapp[1], O. Cerri[2], G. Dissertori[1], T. Q. Nguyen[2], M. Pierini[3,a] [ID], J. R. Vlimant[2]

[1] Institute for Particle Physics and Astrophysics, ETH Zurich, 8093 Zurich, Switzerland
[2] Division of Physics, Mathematics and Astronomy, California Institute of Technology, 1200 E. California Blvd, Pasadena, CA 91125, USA
[3] Experimental Physics Department, European Organization for Nuclear Research (CERN), 1211 Geneva, Switzerland

Abstract We apply an Adversarially Learned Anomaly Detection (ALAD) algorithm to the problem of detecting new physics processes in proton–proton collisions at the Large Hadron Collider. Anomaly detection based on ALAD matches performances reached by Variational Autoencoders, with a substantial improvement in some cases. Training the ALAD algorithm on 4.4 fb^{-1} of 8 TeV CMS Open Data, we show how a data-driven anomaly detection and characterization would work in real life, re-discovering the top quark by identifying the main features of the $t\bar{t}$ experimental signature at the LHC.

1 Introduction

CERN's Large Hadron Collider (LHC) delivers proton-proton collisions in unique experimental conditions. Not only it accelerates protons to an unprecedented energy (6.5 TeV for each proton beam). It also operates at the highest collision frequency, producing one proton-beam crossing (an "event") every 25 ns. By recording the information of every sensor, the two LHC multipurpose detectors, ATLAS and CMS, generate $\mathcal{O}(1)$ MB of data for each interesting event. The LHC big data problem consists in the incapability of storing each event, that would require writing $\mathcal{O}(10)$ TB/s. In order to deal with this limitation, a typical LHC detector is equipped with a real-time event selection system, the trigger.

The need of a trigger system for data taking has important consequences downstream. In particular, it naturally shapes the data analysis strategy for many searches for new physics phenomena (new particles, new forces, etc.) into a hypothesis test [1]: one specifies a signal hypothesis upfront (in fact, as upfront as the trigger system) and tests the presence of the predicted kind of events (the experimental signature) on top of a background from known physics processes, against the alternative hypothesis (i.e., known physics processes alone). From a data-science perspective, this corresponds to a supervised strategy. This procedure was very successful so far, thanks to well established signal hypotheses to test (e.g., the existence of the Higgs boson). On the other hand, following this paradigm didn't produce so

[a] e-mail: maurizio.pierini@cern.ch (corresponding author)

far any evidence for new-physics signals. While this is teaching us a lot about our Universe,[1] it also raises questions on the applied methodology.

Recent works have proposed strategies, mainly based on Machine Learning (ML), to relax the underlying assumptions of a typical experimental analysis [2–12], extending traditional unsupervised searches performed at colliders [13–20]. While many of these works focus on *off-line* data analysis, Ref. [3] advocates the need to also perform an *on-line* event selection with anomaly detection techniques, in order to be able to save a fraction of new physics events even when the underlying new physics scenario was unforeseen (and no dedicated trigger algorithm was put in place). Selected anomalous events could then be visually inspected (as done with the CMS exotica hotline data stream on early LHC runs in 2010-2012) or be given as input to off-line unsupervised analyses, following any of the strategies suggested in literature.

In this paper, we extend the work of Ref. [3] in two directions: (i) we identify anomalies using an Adversarially Learned Anomaly Detection (ALAD) algorithm [21], which combines the strength of generative adversarial networks [22,23] with that of autoencoders [24–26]; (ii) we demonstrate how the anomaly detection would work in real life, using the ALAD algorithm to re-discover the top quark. To this purpose we use real LHC data, released by the CMS experiment on the CERN Open Data portal [27]. Our implementation of the ALAD model in TensorFlow [28], derived from the original code of Ref. [21], is available on GitHub [29].

This paper is structured as follows: the ALAD algorithm is described in Sect. 2. Its performance is assessed in Sect. 3, repeating the study of Ref. [3]. In Sect. 4 we use an ALAD algorithm to re-discover the top quark on a fraction of the CMS 2012 Open Data (described in "Appendix A"). Conclusions are given in Sect. 5.

2 Adversarially Learned Anomaly Detection

The ALAD algorithm is a kind of Generative Adversarial Network (GAN) [22] specifically designed for anomaly detection. The basic idea underlying GANs is that two artificial neural networks compete against each other during training, as shown in Fig. 1. One network, the generator $G : \mathcal{Z} \rightarrow \mathcal{X}$, learns to generate new samples in the data space (e.g., proton-proton collisions in our case) aiming to resemble the samples in the training set. The other network, the discriminator $D_x : \mathcal{X} \rightarrow [0, 1]$, tries to distinguish real samples from generated ones, returning the score of a given sample to be real, as opposed of being generated by G. Both G and D_x are expressed as neural networks, which are trained against each other in a saddle-point problem:

$$\min_{G} \max_{D_x} \mathbb{E}_{x \sim p_{\mathcal{X}}} [\log D_x(x)] + \mathbb{E}_{z \sim p_{\mathcal{Z}}} [\log (1 - D_x(G(z)))] , \tag{1}$$

where $p_{\mathcal{X}}(x)$ is the distribution over the data space $\mathcal{X}$ and $p_{\mathcal{Z}}(z)$ is the distribution over the latent space $\mathcal{Z}$. The solution to this problem will have the property $p_{\mathcal{X}} = p_G$, where p_G is the distribution induced by the generator [22]. The training typically involves alternating gradient descent on the parameters of G and D_x to maximize for D_x (treating G as fixed) and to minimize for G (treating D_x as fixed).

---

[1] For instance, the amount of information derived from this large number of "unsuccessful" searches has put to question the concept of "natural" new physics models, such as low scale supersymmetry. Considering that the generally prevailing pre-LHC view of particle physics was based on two pillars (the Higgs boson and low-scale natural supersymmetry), these experimental results are shaping our understanding of microscopic physics as much as, and probably even more than, the discovery of the Higgs boson.

Deep learning for anomaly detection [23] is usually discussed in the context of (variational) autoencoders [24,25]. With autoencoders (cf. Fig. 1), one projects the input x to a point z of a latent-space through an encoder network $E : \mathcal{X} \to \mathcal{Z}$. An approximation $D(z) = D(E(x))$ of the input information is then reconstructed through the decoder network, $D : \mathcal{Z} \to \mathcal{X}$. The intuition is that the decoder D can only reconstruct the input from the latent space representation z if $x \sim p_{\mathcal{X}}$. Therefore, the reconstruction for an anomalous sample, which belongs to a different distribution, would typically have a higher reconstruction loss. One can then use a metric D_R defining the output-to-input distance (e.g., the one used in the reconstruction loss function) to derive an anomaly-score A:

$$A(x) \sim D_R(x, D(E(x))). \tag{2}$$

While this is not directly possible with GANs, since a generated $G(z)$ doesn't correspond to a specific x, several GAN-based solutions have been proposed that would be suitable for anomaly detection, as for instance in Refs. [21,23,30–32].

In this work, we focus on the ALAD method [21], built upon the use of bidirectional-GANs (BiGAN) [33]. As shown in Fig. 1, a BiGAN model adds an encoder $E : \mathcal{X} \to \mathcal{Z}$ to the GAN construction. This encoder is trained simultaneously to the generator. The saddle point problem in Eq. (1) is then extended as follows:

$$\min_{G,E} \max_{D_{xz}} V(D_{xz}, E, G) = \min_{G,E} \max_{D_{xz}} \mathbb{E}_{x \sim p_{\mathcal{X}}}[\log D_{xz}(x, E(x))] \\ + \mathbb{E}_{z \sim p_{\mathcal{Z}}}[\log(1 - D_{xz}(G(z), z)], \tag{3}$$

where D_{xz} is a modified discriminator, taking inputs from both the $\mathcal{X}$ and $\mathcal{Z}$. Provided there is convergence to the global minimum, the solution has the distribution match-
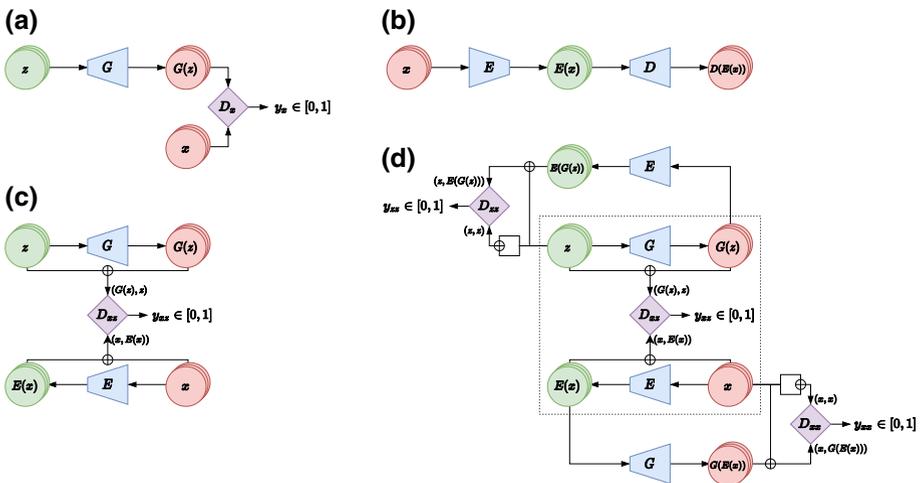

Fig. 1 Comparison of Deep Network architectures: **a** In a GAN, a generator G returns samples $G(z)$ from latent-space points z, while a discriminator D_x tries to distinguish the generated samples $G(z)$ from the real samples x. **b** In an autoencoder, the encoder E compresses the input x to a latent-space point z, while the decoder D provides an estimate $D(z) = D(E(x))$ of x. **c** A BiGAN is built by adding to a GAN an encoder to learn the z representation of the true x, and using the information both in the real space $\mathcal{X}$ and the latent space $\mathcal{Z}$ as input to the discriminator. **d** The ALAD model is a BiGAN in which two additional discriminators help converging to a solution which fulfils the cycle-consistency conditions $G(E(x)) \approx x$ and $E(G(z)) \approx z$. The $\oplus$ symbol in the figure represents a vector concatenation

ing property $p_E(x, z) = p_G(x, z)$, where one defines $p_E(x, z) = p_E(z|x)p_{\mathcal{X}}(x)$ and $p_G(x, z) = p_G(x|z)p_{\mathcal{Z}}(z)$ [33]. To help reaching full convergence, the ALAD model is equipped with two additional discriminators: D_{xx} and D_{zz}. The former discriminator together with the value function

$$V(D_{xx}, E, G) = \mathbb{E}_{x \sim p_{\mathcal{X}}}[\log D_{xx}(x, x)] + \mathbb{E}_{x \sim p_{\mathcal{X}}}[\log (1 - D_{xx}(x, G(E(x))))] \quad (4)$$

enforces the cycle-consistency condition $G(E(x)) \approx x$. The latter is added to further regularize the latent space through a similar value function:

$$V(D_{zz}, E, G) = \mathbb{E}_{z \sim p_{\mathcal{Z}}}[\log D_{zz}(z, z)] + \mathbb{E}_{z \sim p_{\mathcal{Z}}}[\log (1 - D_{zz}(z, E(G(z))))], \quad (5)$$

enforcing the cycle condition $E(G(z)) \approx z$. The ALAD training objective consists in solving:

$$\min_{G,E} \max_{D_{xz}, D_{xx}, D_{zz}} V(D_{xz}, E, G) + V(D_{xx}, E, G) + V(D_{zz}, E, G). \quad (6)$$

Having multiple outputs at hand, one can associate the ALAD algorithm to several anomaly-score definitions. Following Ref. [21], we consider the following four anomaly scores:

– A "Logits" score, defined as: $A_L(x) = \log(D_{xx}(x, G(E(x))))$.
– A "Features" score, defined as: $A_F(x) = ||f_{xx}(x, x) - f_{xx}(x, G(E(x)))||_1$, where $f_{xx}(\cdot, \cdot)$ are the activation values in the last hidden layer of D_{xx}.
– The L_1 distance between an input x and its reconstructed output $G(E(x))$: $A_{L_1}(x) = ||x - G(E(x))||_1$.
– The L_2 distance between an input x and its reconstructed output $G(E(x))$: $A_{L_2}(x) = ||x - G(E(x))||_2$.

We first apply this model to the problem described in Ref. [3], in order to obtain a direct comparison with VAE-based anomaly detection. Then, we apply this model to real LHC data (2012 CMS Open Data), showing how anomaly detection could guide physicists to discover and characterize new processes.

3 ALAD performance benchmark

We consider a sample of simulated LHC collisions, pre-filtered by requiring the presence of a muon with large transverse momentum (p_T)[2] and isolated from other particles. Proton-proton collision events at a center-of-mass energy $\sqrt{s} = 13$ TeV are generated with the PYTHIA8 event-generation library [34]. The generated events are further processed with the DELPHES library [35] to model the detector response. Subsequently the DELPHES *particle-flow* (PF) algorithm is applied to obtain a list of reconstructed particles for each event, the so-called PF candidates.

Events are filtered requiring $p_T > 23$ GeV and isolation[3] Iso < 0.45. Each collision event is represented by 21 physics-motivated high-level features (see Ref. [3]). These input

---

[2] As common for collider physics, we use a Cartesian coordinate system with the z axis oriented along the beam axis, the x axis on the horizontal plane, and the y axis oriented upward. The x and y axes define the transverse plane, while the z axis identifies the longitudinal direction. The azimuth angle ϕ is computed with respect to the x axis. The polar angle θ is used to compute the pseudorapidity $\eta = -\log(\tan(\theta/2))$. The transverse momentum (p_T) is the projection of the particle momentum on the (x, y) plane. We fix units such that $c = \hbar = 1$.

[3] A definition of isolation is provided in "Appendix A."

Table 1 Composition of the cocktail of SM processes

Process	Cross section [nb]	Trigger efficiency (%)	Acceptance (%)	Fraction (%)
$W \to \ell\nu$	58	68	55.6	59.2
QCD	1.6×10^5	9.6	0.08	33.8
$Z \to \ell\ell$	20	77	16	6.7
$t\bar{t}$	0.7	49	37	0.3

The first column gives the production cross section for each process. The trigger efficiency is the fraction of events passing a loose selection on the muon p_T, corresponding to an online selection in the trigger system. The acceptance is the fraction of events passing our selection criteria ($p_T > 23$ GeV and Iso < 0.45). Each fraction is computed with respect to the previous step. The last column gives the composition of the SM-cocktail

features are pre-processed before being fed to the ALAD algorithm. The discrete quantities[4] (q_l, IsEle, N_μ and N_e) are represented through one-hot encoding. The other features are standardized to a zero median and unit variance. The resulting vector, containing the one-hot encoded and continuous features, has a dimension of 39 and is given as input to the ALAD algorithm.

The sample, available on Zenodo [36], consists of the following Standard Model (SM) processes:

– Inclusive W boson production: $W \to \ell\nu$, with $\ell = e, \mu, \tau$ being a charged lepton [37].
– Inclusive Z boson production: $Z \to \ell\ell$ [38].
– Multijet production from Quantum Chromodynamic (QCD) interaction [39].
– $t\bar{t}$ production [40].

A SM cocktail is assembled from a weighted mixture of those four processes, with weights given by the production cross section. This cocktail's composition is given in Table 1.

We train our ALAD model on this SM cocktail and subsequently apply it to a test dataset, containing a mixture of SM events and events of physics beyond the Standard Model (BSM). In particular, we consider the following BSM datasets, also available on Zenodo:

– A leptoquark with mass 80 GeV, decaying to a b quark and a τ lepton: $LQ \to b\tau$ [41].
– A neutral scalar boson with mass 50 GeV, decaying to two off-shell Z bosons, each forced to decay to two leptons: $A \to 4\ell$ [42].
– A scalar boson with mass 60 GeV, decaying to two τ leptons: $h^0 \to \tau\tau$ [43].
– A charged scalar boson with mass 60 GeV, decaying to a τ lepton and a neutrino: $h^\pm \to \tau\nu$ [44].

As a starting point, we consider the ALAD architecture [21] used for the KDD99 [45] dataset[5], which has similar dimensionality as our input feature vector. In this configuration, both the D_{xx} and D_{zz} discriminators take as input the concatenation of the two input vectors, which is processed by the network up to the single output node, activated by a sigmoid function. The D_{xz} discriminator has one dense layer for each of the two inputs. The two intermediate representations are concatenated and passed to another dense layer and then

---

[4] q_l is the charge of the lepton; IsEle is a flag set to 1 (0) if the lepton is an electron (muon); N_μ and N_e are the muon and electron multiplicities, respectively.

[5] The KDD99 dataset is one of the most used datasets for studies related to anomaly detection. The dataset contains real credit card transaction data, labelled as legitimate or fraudulent. The dataset was developed for the Fifth International Conference on Knowledge Discovery and Data Mining.

to a single output node with sigmoid activation, as for the other discriminators. The hidden nodes of the generator are activated by ReLU functions [46], while Leaky ReLU [47] are used for all the other nodes. The slope parameter of the Leaky ReLU function is fixed to 0.2. The network is optimized using the Adam [48] minimizer and minibatches of 50 events each. The training is regularized using dropout layers in the three discriminators.

Starting from this baseline architecture, we adjust the architecture hyperparameters one by one, repeating the training while maximizing a figure of merit for anomaly detection efficiency. We perform this exercise using as anomalies the benchmark models described in Ref. [3] and looking for a configuration that performs well on all of them. To quantify performance, we consider both the area under the receiver operating characteristic (ROC) curve and the positive likelihood ratio LR_+. We define the LR_+ as the ratio between the BSM signal efficiency, i.e., the true positive rate (TPR), and the SM background efficiency, i.e., the false positive rate (FPR). The training is performed on half of the available SM events (3.4M events), leaving the other half of the SM events and the BSM samples for validation. From the resulting anomaly scores, we compute the ROC curve and compare it to the results of the VAE in Ref. [3]. We further quantify the algorithm performance considering the LR_+ values corresponding to an FPR of 10^{-5}.

The optimized architecture, adapted from Ref. [21], is summarized in Table 2. This architecture is used for all subsequent studies. We consider as hyperparameters the number of hidden layers in the five networks, the number of nodes in each hidden layer, and the dimensionality of the latent space, represented in the table by the size of the E output layer.

Having trained the ALAD on the training dataset, we compute the anomaly scores for the validation samples as well as for the four BSM samples, where each BSM process has $O(0.5M)$ samples. Figure 2 shows the ROC curves of each BSM benchmark process, for the four considered anomaly scores. The best VAE result from Ref. [3] is also shown for comparison. In the rest of this paper, we use the L_1 score as the anomaly score. Similar results would have been obtained using any of the other three anomaly scores. Figure 3 compares the A_{L_1} distribution for each BSM process with the SM cocktail. One can clearly see that all BSM processes have an increased probability in the high-score regime compared to the SM cocktail. We further verified that the anomaly score distributions obtained on the SM-cocktail training and validation sets are consistent. This test excludes the occurrence of over-training issues.

The ALAD algorithm outperforms the VAE by a substantial margin on the $A \rightarrow 4\ell$ sample, providing similar performance overall, and in particular for FPR $\sim 10^{-5}$, the working point chosen as a reference in Ref. [3]. We verified that the uncertainty on the TPR at fixed FPR, computed with the Agresti-Coull interval [49], is negligible when compared to the observed differences between ALAD and VAE ROC curves, i.e., the difference is statistically significant.

The left plot in Fig. 4 provides a comparison across different BSM models. As for the VAE, ALAD performs better on $A \rightarrow 4\ell$ and $h^{\pm} \rightarrow \tau\nu$ than for the other two BSM processes. The right plot in Fig. 4 shows the LR_+ values as a function of the FPR ones. The LR_+ peaks at a SM efficiency of $O(10^{-5})$ for all four BSM processes and is basically constant for smaller SM-efficiency values.

4 Re-discovering the top quark with ALAD

In order to test the performance of ALAD on real data, and in general to show how an anomaly detection technique could guide physicists to a discovery in a data-driven manner,

Table 2 Hyperparameters for the ALAD algorithm

Operation	Units	Activation	Batch Norm.	Dropout rate
$E(x)$				
Number of hidden layers	**2**			
Dense	**64**	Leaky ReLU	×	–
Dense	**64**	Leaky ReLU	×	–
Output	**16**	Linear	×	–
$G(z)$				
Number of hidden layers	**2**			
Dense	**64**	ReLU	×	–
Dense	**64**	ReLU	×	–
Output	39	Linear	×	–
$D_{xz}(x, z)$				
Number of hidden layers	**2**			
Only on x				
Dense	**128**	Leaky ReLU	√	–
Only on z				
Dense	**128**	Leaky ReLU	×	0.5
Concatenate outputs				
Dense	**128**	Leaky ReLU	×	0.5
Output	1	Sigmoid	×	–
$D_{xx}(x, \hat{x})$				
Concatenate x and x'				
Number of hidden layers	**1**			
Dense	**128**	Leaky ReLU	×	0.2
Output	1	Sigmoid	×	–
$D_{zz}(z, \hat{z})$				
Concatenate z and z'				
Number of hidden layers	**1**			
Dense	**128**	Leaky ReLU	×	0.2
Output	1	Sigmoid	×	–
Training parameter	Value			
Optimizer	Adam ($\alpha = 10^{-5}$, $\beta_1 = 0.5$)			
Batch size	**50**			
Leaky ReLU slope	0.2			
Spectral norm	√			
Weight, bias init.	Xavier Initializer, Constant(0)			

Parameters in bold have been optimized for. No Dropout layer is applied wherever a dropout rate is not specified

we consider a scenario in which collision data from the LHC are available, but no previous knowledge about the existence of the top quark is at hand. The list of known SM processes contributing to a dataset with one isolated high-p_T lepton would then include W production, Z/γ^* production and QCD multijet events, neglecting more rare processes such as diboson
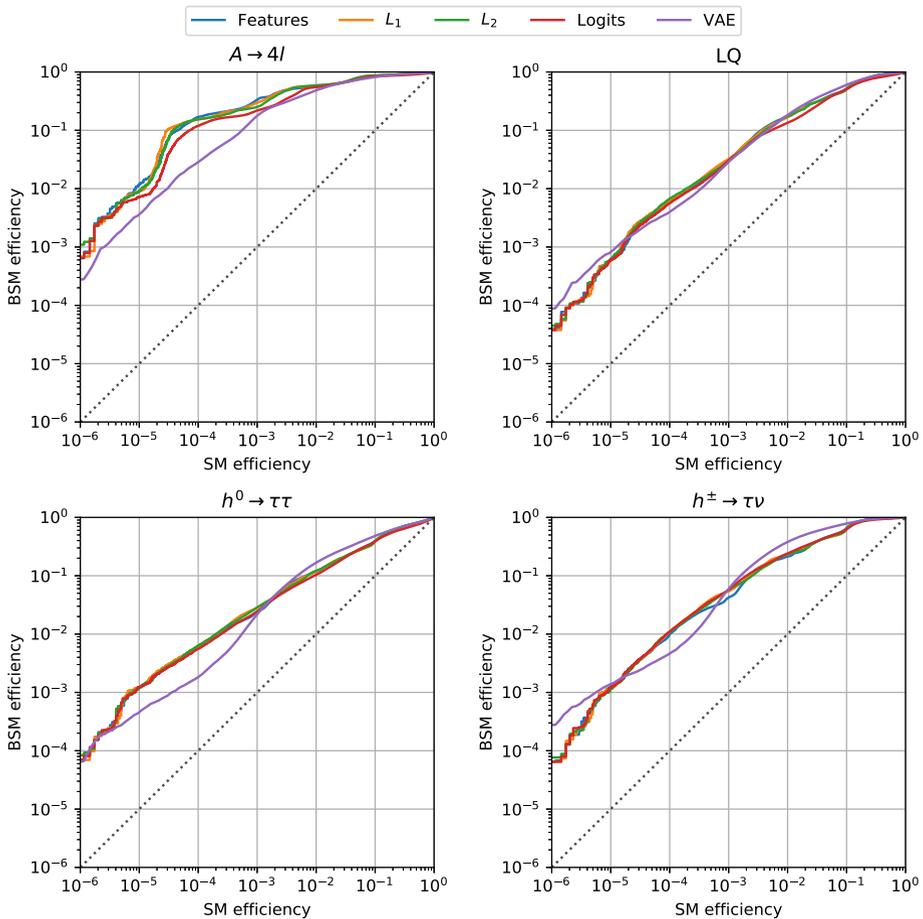
Fig. 2 ROC curves for the ALAD trained on the SM cocktail training set and applied to SM + BSM validation samples. The VAE curve corresponds to the best result of Ref. [3], which is shown here for comparison. The other four lines correspond to the different anomaly score models of the ALAD

production. Top-quark pair production, the "unknown anomalous process," represents ∼ 1% of the total dataset.

We consider a fraction of LHC events collected by the CMS experiment in 2012 and corresponding to an integrated luminosity of about 4.4 fb^{-1} at a center-of-mass energy $\sqrt{s} = 8$ TeV. For each collision event in the dataset, we compute a vector of physics-motivated high-level features, which we give as input to ALAD for training and inference. Details on the dataset, event content, and data processing can be found in "Appendix A."

We select events applying the following requirements:

– At least one lepton with $p_T > 23$ GeV and PF isolation Iso < 0.1 within $|\eta| < 1.4$.
– At least two jets with $p_T > 30$ within $|\eta| < 2.4$.

This selection is tuned to reduce the expected QCD multijet contamination to a negligible level, which avoids problems with the small size of the available QCD simulated dataset. This selection should not be seen as a limiting factor for the generality of the study: in real
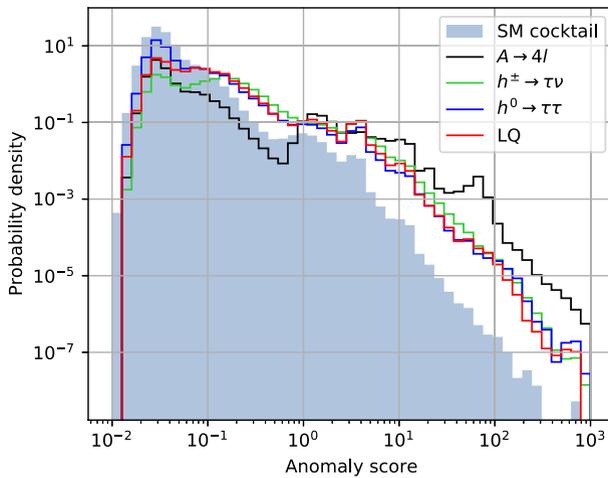
Fig. 3 Distribution for the A_{L_1} anomaly score. The "SM cocktail" histogram corresponds to the anomaly score for the validation sample. The other four distributions refer to the scores of the four BSM benchmark models
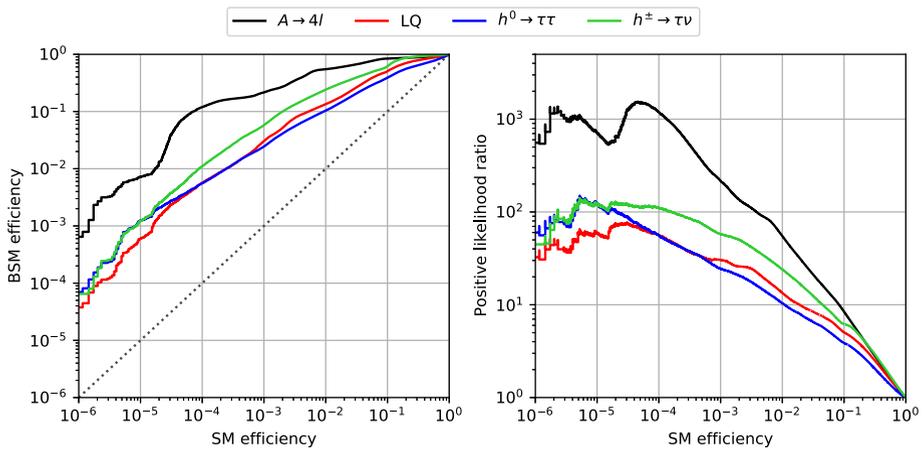
Fig. 4 Left: ROC curves for each BSM process obtained with the ALAD L_1-score model. Right: LR$_+$ curves corresponding to the ROC curves on the left

life, one would apply multiple sets of selection requirements on different triggers, in order to define multiple datasets on which different anomaly detection analyses would run.

Our goal is to employ ALAD to isolate $t\bar{t}$ events as due to a rare (and pretended to be) unknown process in the selected dataset. Unlike the case discussed in Ref. [3] and Sect. 3, we are not necessarily interested in a pre-defined algorithm to run on a trigger system. Given an input dataset, we want to isolate its outlier events without explicitly specifying which tail of which kinematic feature one should look at. Because of this, we don't split the data into a training and a validation dataset. Instead, we run the training on the full dataset. The training is performed fixing the ALAD architecture to the values shown in Table 2. In our procedure, we implicitly rely on the assumption that the anomalous events are rare, so that their modeling is not accurately learned by the ALAD algorithm.
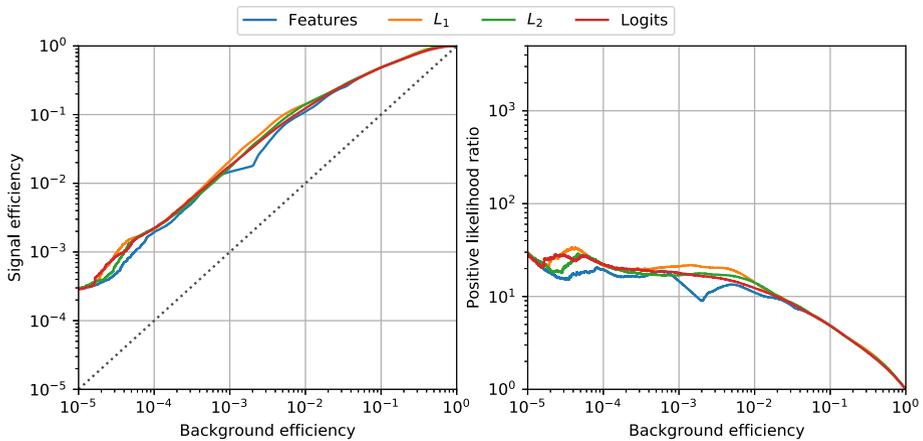
Fig. 5 Left: ROC curves for each anomaly score, where the signal efficiency is the fraction of $t\bar{t}$ (signal) events passing the anomaly selection, i.e., the true positive rate (TPR). The background efficiency is the fraction of background events passing the selection, i.e. the false positive rate (FPR). Right: Positive likelihood ratio (LR_+) curves corresponding to the ROC curves in the left

In order to evaluate the ALAD performance, we show in Fig. 5 the ROC and LR_+ curves on labelled Monte Carlo (MC) simulated data, considering the $t\bar{t}$ sample as the signal anomaly and the SM W and Z production as the background. An event is classified as anomalous whenever the L_1 anomaly score is above a given threshold. The threshold is set such that the fraction of selected events is about 10^{-3}. The anomaly selection results in a factor-20 enhancement of the $t\bar{t}$ contribution over the SM background, for the anomaly-defining FPR threshold.

Figure 6 shows the distributions of a subset of the input quantities before and after anomaly selection:

- H_T—The scalar sum of the transverse momenta (p_T) of all jets having $p_T > 30$ GeV and $|\eta| < 2.4$.
- M_J—The invariant mass of all jets entering the H_T sum.
- N_J—The number of jets entering the H_T sum.
- N_B—The number of jets identified as originating from a b quark.

More details are given in "Appendix A." These are the quantities that will become relevant in the rest of the discussion. The corresponding distributions for MC simulated events are shown in Fig. 7, where the $t\bar{t}$ contribution to the selected anomalous data is highlighted. At this stage, we don't attempt a direct comparison of simulated distributions to data, since we couldn't apply the many energy scale and resolution corrections, normally applied for CMS studies. Anyhow, such a comparison is not required for our data-driven definition of anomalous events.

In order to quantify the impact of the applied anomaly selection on a given quantity, we consider the bin-by-bin fraction of accepted events. We identify this differential quantity as anomaly selection transfer function (ASTF). When compared to the expectation from simulated MC data, the dependence of ASTF values on certain quantities allows one to characterize the nature of the anomaly, indicating in which range of which quantity the anomalies cluster together.
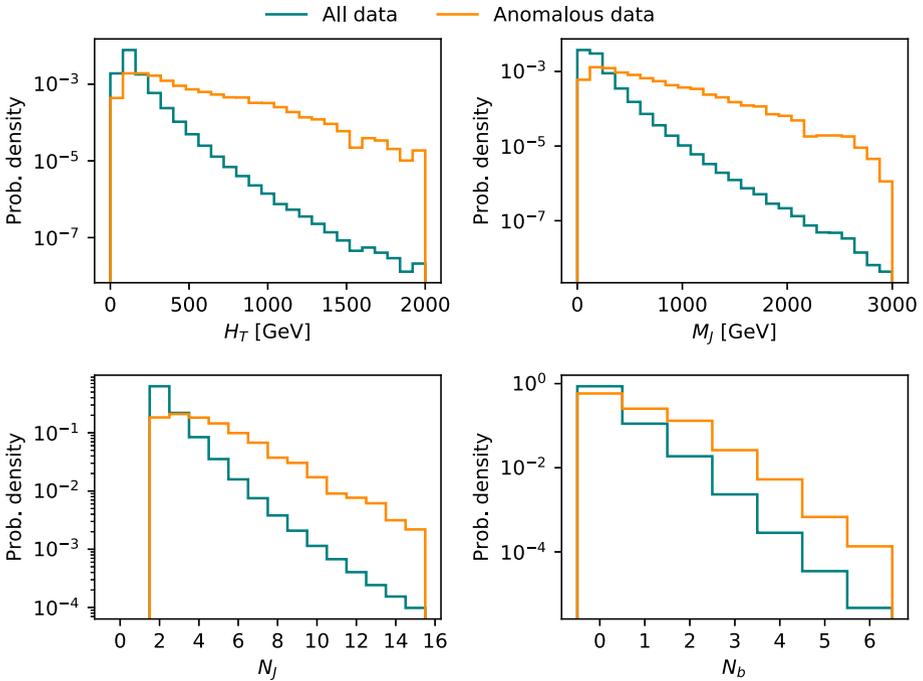
Fig. 6 Event distribution on data, before and after the anomaly selection: H_T (top-left), M_J (top-right), N_J (bottom-left), and N_b (bottom-right) distributions, normalized to unit area. A definition of the considered features is given in "Appendix A"

Figure 8 shows the ASTF for the data and the simulated SM processes (W and Z production) defining the background data. The comparison between the data and the background ASTF suggests the presence of an unforeseen class of events, clustering at large number of jets. An excess of anomalies is observed at large jet multiplicity, which also induces an excess of anomalies at large values of H_T and M_J. Notably, a large fraction of anomalous events has jets originating from b quarks. This is the first time that a MC simulation enters our analysis. We stress the fact that this comparison between data and simulation is qualitative. At this stage, we don't need the MC-predicted ASTF values to agree with data in absence of a signal, since we are not attempting a background estimate like those performed in data-driven searches at the LHC. For us, it is sufficient to observe qualitative differences in the dependence of ASTFs on specific features. Nevertheless, a qualitative difference like those we observe in Fig. 8 could still be induced by systematic differences between data and simulation. That would still be an anomaly, but not of the kind that would lead to a discovery. In our case, we do know that the observed discrepancy is too large to be explained by a wrong modeling of W and Z events, given the level of accuracy reached by the CMS simulation software in Run I. In a real-life situation, one would have to run further checks to exclude this possibility.

Starting from the ASTF plots of Fig. 8, we define a post-processing selection (PPS), aiming to isolate a subset of anomalous events in which the residual SM background could be suppressed. In particular, we require $N_J \geq 6$ and $N_b \geq 2$. Figure 9 shows the distributions of some of the considered features after the PPS. According to MC expectations, almost all background events should be rejected. The same should apply to background events in data.
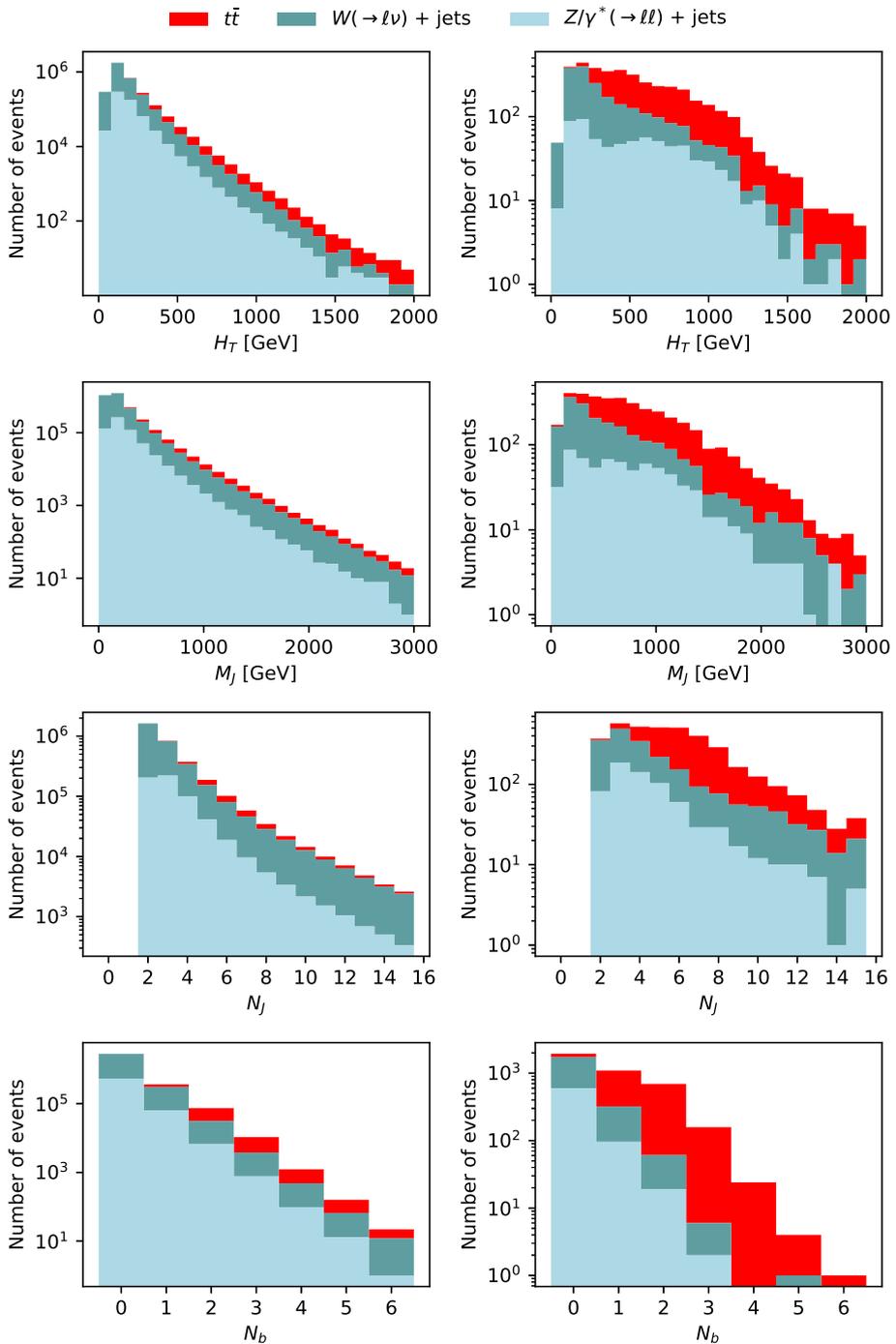
Fig. 7 Event distribution from simulation before (left) and after (right) applying the anomaly selection. From top to bottom: H_T, M_J, N_J, and N_b. Distributions are normalized to an integrated luminosity $\int L = 4.4\,\mathrm{fb}^{-1}$
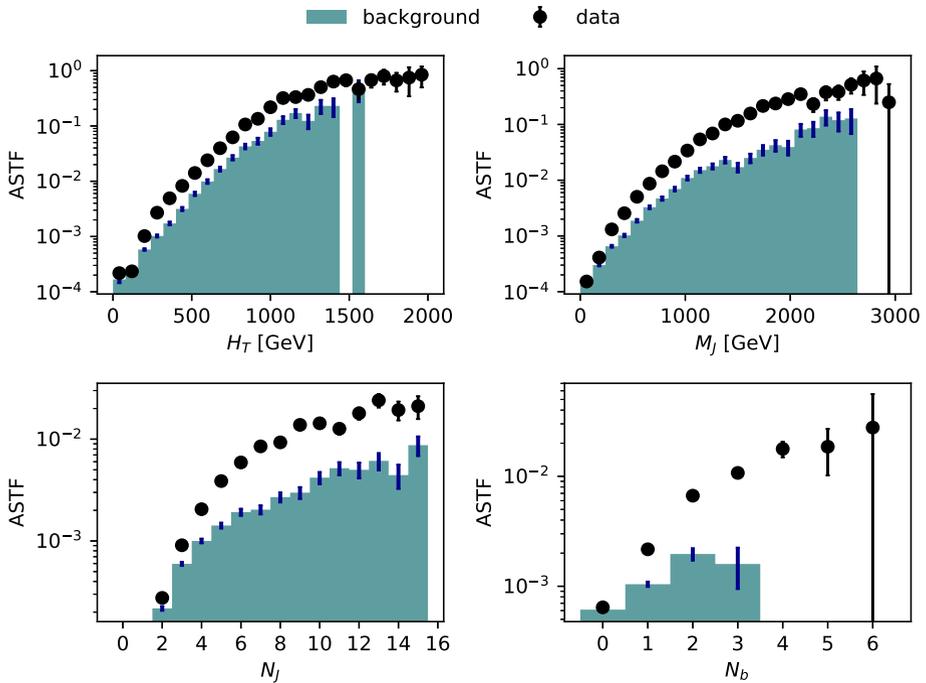
Fig. 8 ASTF ratios for H_T (top-left), M_J (top-right), N_J (bottom-left), and N_b (top-right) for data. The filled area shows the same ratio, computed on simulated background data (W and Z events). The uncertainty bands represent the statistical uncertainties on the ASTF ratios

Instead, a much larger number of events is observed. This discrepancy points to the presence of an additional process, characterized by many jets (particularly coming from b jets). As a closure test, we verified that the agreement between the observed distributions in data and the expectation from MC simulation is restored once the $t\bar{t}$ contribution is taken into account.

In summary, the full procedure consists of the following steps:

1. Define an input dataset with an event representation which is as generic as possible.
2. Train the ALAD (or any other anomaly detection algorithm) on it.
3. Isolate a fraction α of the events, by applying a threshold on the anomaly score.
4. Study the ASTF on as many quantities as possible, in order to define a post-processing selection that allows one to isolate the anomaly.
5. (Optionally) once a pure sample of anomalies is isolated, a visual inspection of the events could also guide the investigation, as already suggested in Ref.[3].

The event selection could happen in the trigger system of the experiment. Similarly, one could foresee a list of ASTF plots to be produced in real-time, e.g., as part of the Data Quality Monitoring system of the experiment. While the algorithm training could happen online (e.g., using a sample of randomly selected events, to remove bias induced by the online selection), the proposed strategy would benefit from a real-time training workflow, which could be set up with dedicated hardware resources. There could be a cost in terms of trigger throughput, that could be compensated by foreseeing dedicated scouting streams [50–52] with reduced event content, tailored to the computation of the ASTF ratios in offline analyses. In this case, the ASTF ratio plots could be produced offline.
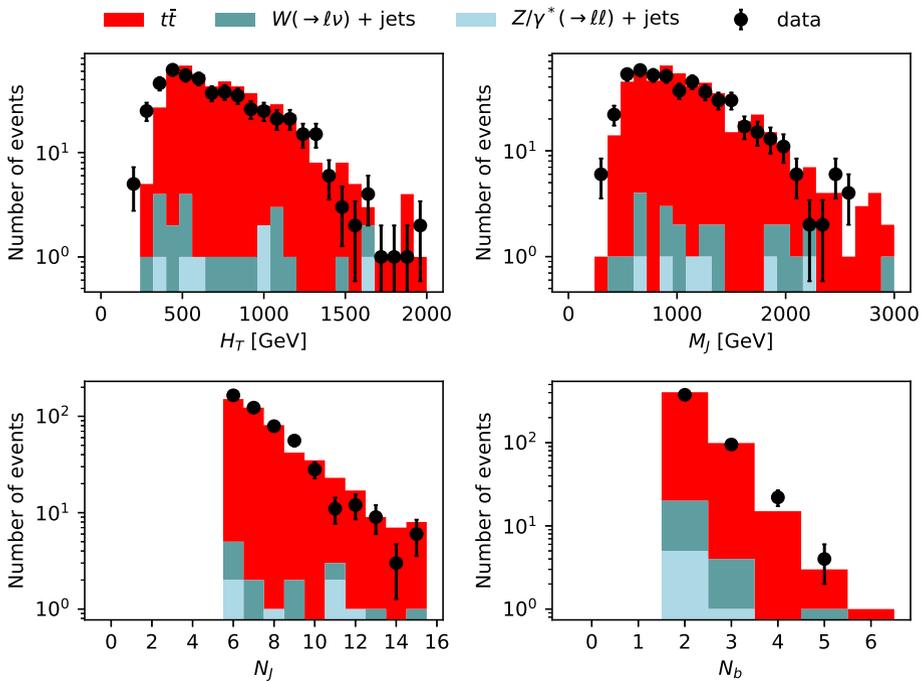
Fig. 9 Data distribution for H_T (top-left), M_J (top-right), N_J (bottom-left), and N_b (top-right) after the post processing selection. The filled histograms show the expectation from MC simulation, normalized to an integrated luminosity of $\int L = 4.4\,\text{fb}^{-1}$), including the $t\bar{t}$ contribution

Once the aforementioned procedure is applied, one would gain an intuition about the nature of the new rare process. For instance, one could study the ASTF distribution using as a reference quantity on the x axis the run period at which an event was taken. Anomalies clustering on specific run periods would most likely be due to transient detector malfunctioning or experimental problems of other nature (e.g., a bug in the reconstruction software). If instead the significant ASTF ratios point to events with a lepton, large jet multiplicity, with an excess of b-jets, one might have discovered a new heavy particle decaying to leptons and b-jets. In a world with no previous knowledge of the top quark, some bright theorist would explain the anomaly proposing the existence of a third up-type quark with unprecedented heavy mass.

5 Conclusions

We presented an application of ALAD to the search for new physics at the LHC. Following the study presented in Ref. [3], we show how this algorithm matches (and in some cases improves) the performance obtained with variational autoencoders. The ALAD architecture also offers practical advantages with respect to previously proposed approach, based on VAEs. Besides providing an improved performance in some cases, it offers an easier training procedure. Good performance can be achieved using a standard MSE loss, unlike the previously proposed VAE

model, for which good performance was obtained only after a heavy customization of the loss function.

We train the ALAD algorithm on a sample of real data, obtained by processing part of the 2012 CMS Open data. On these events, we show how one could detect the presence of $t\bar{t}$ events, a $\sim$ 4% population (after selection requirements are applied) pretended to originate from a previously unknown process. Using the anomaly score of the trained ALAD algorithm, we define a post-selection procedure that let us isolate an almost pure subset of anomalous events. Furthermore, we present a strategy based on ASTF distributions to characterize the nature of an observed excess and we show its effectiveness on the specific example at hand. Further studies should be carried on to demonstrate the robustness of this strategy for more rare processes.

For the first time, our study shows with real LHC data that anomaly detection techniques can highlight the presence of rare phenomena in a data-driven manner. This result could help promoting a new class of data-driven studies in the next run of the LHC, possibly offering additional guidance in the search for new physics.

As for the VAE model, the promotion of these algorithms to the trigger system of the LHC experiments could allow to mitigate for the limited bandwidth of the experiments. Unlike the VAE model, the strategy presented in this paper would also require a substantial fraction of normal events to be saved (through pre-scaled triggers), in order to compute the ASTF ratios with sufficient precision. The procedure could run in the trigger system, or on dedicated low-rate data streams. Putting in place this strategy for the LHC Run III could be an interesting way to extend the physics reach of the LHC experiments.

Data Availability Statement This manuscript has associated data in a data repository. [Authors' comment: the Delphes datasets is released on Zenodo and available at https://zenodo.org/communities/mpp-hep. The CMS Open Data are released on the CERN Open Data portal at https://opendata.cern.ch/search?experiment=CMS.]

Appendix

A CMS single muon open data

The re-discovery of the top quark, described in Sect. 4, is performed on a sample of real LHC collisions, collected by the CMS experiment and released on the CERN Open Data portal [27]. The collision data correspond to the SingleMu dataset in the Run2012B period [53]. This dataset results from the logic OR of many triggers requiring a reconstructed muon. Most of the events were collected by an inclusive isolated-muon trigger, with requirements very similar to those applied as a pre-selection in Sect. 3 [3].

Once collected, the raw data recorded by the CMS detector were processed by a global event reconstruction algorithm, based on Particle Flow (PF) [54]. This processing step gives as output a list of so-called PF candidates (electrons, muons, photons, charged hadrons and neutral hadrons), reconstructed combining measurements from different detectors to achieve maximal accuracy.

While the ALAD training is performed directly on data, samples from MC simulation are also used in the study, in order to validate the model and its findings on a labelled sample. To this purpose, we considered samples of W+jets [55–57], Z/γ^*+jets [58–61], and $t\bar{t}$ [62] events. No sample of QCD multijet events was considered, since its contribution was found to be negligible after the baseline requirements on the muon and on two additional jets (see Sect. 4). These MC samples were generated by the CMS collaboration with different libraries and processed by a full Geant4 simulation [63]. The same reconstruction software used on data was applied to the output of the simulation, so that the same lists of PF candidates are available in this case.

Following a procedure similar to that of Ref. [3], we take as input the lists of PF candidates and compute a set of physics motivated features, which are used as input to train the ALAD algorithm:

We consider 14 event-related quantities:

- H_T—The scalar sum of the transverse momenta (p_T) of all jets having $p_T > 30$ GeV and $|\eta| < 2.4$.
- M_J—The invariant mass of all jets entering the H_T sum.
- N_J—The number of jets entering the H_T sum.
- N_B—The number of jets identified as originating from a b quark.
- p_T^{miss}—The missing transverse momentum defined as $p_T^{\mathrm{miss}} = \left| -\sum_q \mathbf{p}_T^q \right|$, where the sum goes over all PF candidates.
- $p_{T,TOT}^{\mu}$—The vector sum of the $\mathbf{p}_T$ of all PF muons in the event having $p_T > 0.5$ GeV.
- M_μ—The combined invariant mass of all muons entering the above sum in $p_{T,TOT}^{\mu}$.
- N_μ—The number of muons entering the sum in $p_{T,TOT}^{\mu}$.
- $p_{T,TOT}^{e}$—The vector sum of the $\mathbf{p}_T$ of all PF electrons in the event having $p_T > 0.5$ GeV.
- M_e—The combined invariant mass of all electrons entering the above sum in $p_{T,TOT}^{e}$.
- N_e—The number of electrons entering the sum in $p_{T,TOT}^{e}$.
- N_{neu}—The number of all neutral hadron PF-candidates.
- N_{ch}—The number of all charged hadron PF-candidates.
- N_γ—The number of all photon PF-candidates.

We further consider 10 quantities, specific to the highest-p_T lepton in the event:

- p_T^ℓ – The lepton p_T.
- η_ℓ – The lepton pseudorapidity.

- q_ℓ – The lepton charge (either -1 or $+1$)
- Iso^ℓ_{ch} – The lepton isolation, defined as the ratio between the scalar sum of the p_T of the other charged PF candidates with angular distance $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2} < 0.3$ from the lepton, and the lepton p_T.
- Iso^ℓ_{neu} – Same as Iso^ℓ_{ch} but with the sum going over all neutral hadrons.
- Iso^ℓ_γ – Same as Iso^ℓ_{ch} but with the sum going over all photons.
- Iso – The total isolation given by $\text{Iso} = \text{Iso}^\ell_{ch} + \text{Iso}^\ell_{neu} + \text{Iso}^\ell_\gamma$.
- M_T – The combined transverse mass of the lepton and the E_T^{miss} system, which is given by

$$M_T = \sqrt{2 p_T^\ell E_T^{\text{miss}}(1 - \cos\Delta\phi)}.$$

.
- $p_{T,\parallel}^{\text{miss}}$ – The parallel component of p_T^{miss} with respect to the lepton.
- $p_{T,\perp}^{\text{miss}}$ – The orthogonal component of p_T^{miss} with respect to the lepton.
- IsEle – A flag set to 1 if the lepton is an electron, 0 if it is a muon.

References

1. ATLAS and CMS Collaborations, *Procedure for the LHC Higgs boson search combination in Summer 2011* (2011), aTL-PHYS-PUB-2011-011, CMS-NOTE-2011-005. https://inspirehep.net/literature/1196797. Accessed May 2020
2. C. Weisser, M. Williams (2016). arXiv:1612.07186
3. O. Cerri, T.Q. Nguyen, M. Pierini, M. Spiropulu, J.R. Vlimant, JHEP **05**, 036 (2019). arXiv:1811.10276
4. R.T. D'Agnolo, A. Wulzer, Phys. Rev. D **99**, 015014 (2019). arXiv:1806.02350
5. A. De Simone, T. Jacques, Eur. Phys. J. C **79**, 289 (2019). arXiv:1807.06038
6. M. Farina, Y. Nakai, D. Shih, Phys. Rev. D **101**, 075021 (2020). arXiv:1808.08992
7. J.H. Collins, K. Howe, B. Nachman, Phys. Rev. Lett. **121**, 241803 (2018). arXiv:1805.02664
8. A. Blance, M. Spannowsky, P. Waite, JHEP **10**, 047 (2019). arXiv:1905.10384
9. J. Hajer, Y.Y. Li, T. Liu, H. Wang, Phys. Rev. D **101**, 076015 (2020). arXiv:1807.10261
10. T. Heimel, G. Kasieczka, T. Plehn, J.M. Thompson, SciPost Phys. **6**, 030 (2019). arXiv:1808.08979
11. J.H. Collins, K. Howe, B. Nachman, Phys. Rev. D **99**, 014038 (2019). arXiv:1902.02634
12. R.T. D'Agnolo, G. Grosso, M. Pierini, A. Wulzer, M. Zanetti (2019). arXiv:1912.12155
13. F.D. Aaron et al. (H1), Phys. Lett. B **B674**, 257 (2009). arXiv:0901.0507
14. T. Aaltonen et al. (CDF), Phys. Rev. D **79**, 011101 (2009). arXiv:0809.3781
15. V.M. Abazov et al. (D0), Phys. Rev. D **85**, 092015 (2012). arXiv:1108.5362
16. *MUSiC, a Model Unspecific Search for New Physics, in pp Collisions at $\sqrt{s}$ = 8 TeV* (2017). cMS-PAS-EXO-14-016. https://cds.cern.ch/record/225665. Accessed May 2020
17. M. Aaboud et al. (ATLAS), Eur. Phys. J. **C79**, 120 (2019)
18. B. Nachman, D. Shih, Phys. Rev. D **101**, 075042 (2020). arXiv:2001.04990
19. A. Andreassen, B. Nachman, D. Shih, Phys. Rev. D **101**, 095004 (2020)
20. O. Amram, C.M. Suarez (2020). arXiv:2002.12376
21. H. Zenati, M. Romain, C.S. Foo, B. Lecouat, V.R. Chandrasekhar (2018). arXiv:1812.02288
22. I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio (2014). arXiv:1406.2661
23. T. Schlegl, P. Seeböck, S.M. Waldstein, U. Schmidt-Erfurth, G. Langs (2017). arXiv:1703.05921
24. D.P. Kingma, M. Welling, ArXiv e-prints (2013). arXiv:1312.6114
25. J. An, S. Cho, Spec. Lect. IE **2**, 1 (2015)
26. C. Zhou, R.C. Paffenroth, *Anomaly detection with robust deep autoencoders*, in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York, NY, USA, 2017), p. 665-674. https://doi.org/10.1145/3097983.3098052
27. Open Data Repository. http://opendata.cern.ch/search?page=1&size=20&experiment=CMS. Accessed May 2020

28. M. Abadi et al., *TensorFlow: Large-scale machine learning on heterogeneous systems* (2015). http://tensorflow.org/. Accessed May 2020
29. O. Knapp, *Adversarially learned anomaly detection for the lhc*. gitHub: https://github.com/oliverkn/alad-for-lhc. Accessed May 2020
30. A. Creswell, A.A. Bharath, CoRR abs/1611.05644 (2016). arXiv:1611.05644
31. Y. Wu, Y. Burda, R. Salakhutdinov, R.B. Grosse, CoRR , abs/1611.04273 (2016). arXiv:1611.04273
32. E. Gherbi, B. Hanczar, J.C. Janodet, W. Klaudel, An encoding adversarial network for anomaly detection, in *11th Asian Conference on Machine Learning (ACML 2019)* (Nagoya, Japan, 2019), vol. 101 of *JMLR: Workshop and Conference Proceedings*, p. 1–16. https://hal.archives-ouvertes.fr/hal-02421274
33. J. Donahue, P. Krähenbühl, T. Darrell (2016). arXiv:1605.09782
34. T. Sjöstrand, S. Ask, J.R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C.O. Rasmussen, P.Z. Skands, Comput. Phys. Commun. **191**, 159–177 (2015)
35. J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, M. Selvaggi, (DELPHES 3), JHEP **02**, 057 (2014). arXiv:1307.6346
36. M. Pierini et al., mPP Zenodo community. https://zenodo.org/communities/mpp-hep/?page=1&size=20
37. O. Cerri, T. Nguyen, M. Pierini, J.R. Vlimant, *New physics mining at the large hadron collider*: $w \rightarrow \ell\nu$: (2020). https://doi.org/10.5281/zenodo.3675199
38. O. Cerri, T. Nguyen, M. Pierini, J.R. Vlimant, *New physics mining at the large hadron collider*: $z \rightarrow \ell\ell$: (2020). https://doi.org/10.5281/zenodo.3675203
39. O. Cerri, T. Nguyen, M. Pierini, J.R. Vlimant, *New physics mining at the large hadron collider: Qcd multijet production* (2020). https://doi.org/10.5281/zenodo.3675210
40. O. Cerri, T. Nguyen, M. Pierini, J.R. Vlimant, *New physics mining at the large hadron collider*: $t\bar{t}$ *production:* (2020). https://doi.org/10.5281/zenodo.3675206
41. O. Cerri, T. Nguyen, M. Pierini, J.R. Vlimant, *New physics mining at the large hadron collider*: Lq$\rightarrow b\tau$: (2020). https://doi.org/10.5281/zenodo.3675196
42. O. Cerri, T. Nguyen, M. Pierini, J.R. Vlimant, *New physics mining at the large hadron collider*: $a \rightarrow 4\ell$: (2020). https://doi.org/10.5281/zenodo.3675159
43. O. Cerri, T. Nguyen, M. Pierini, J.R. Vlimant, *New physics mining at the large hadron collider*: $h^0 \rightarrow \tau\tau$: (2020). https://doi.org/10.5281/zenodo.3675190
44. O. Cerri, T. Nguyen, M. Pierini, J.R. Vlimant, *New physics mining at the large hadron collider*: $h^+ \rightarrow \tau\nu$: (2020). https://doi.org/10.5281/zenodo.3675178
45. S.J. Stolfo, W. Fan, W. Lee, A. Prodromidis, P.K. Chan, Cost-based modeling for fraud and intrusion detection: results from the JAM Project, in *Proceedings of the 2000 DARPA Information Survivability Conference and Exposition* (IEEE Computer Press), p. 130–144
46. V. Nair, G.E. Hinton, Rectified linear units improve restricted Boltzmann machines, in *ICML* (2010). https://www.cs.toronto.edu/~fritz/absps/reluICML.pdf
47. A.L. Maas, A.Y. Hannun, A.Y. Ng, Rectifier nonlinearities improve neural network acoustic models, in *ICML Workshop on Deep Learning for Audio, Speech and Language Processing* (2013)
48. D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (2015). arXiv:1412.6980
49. A. Agresti, B.A. Coull, Am. Stat. **52**, 119–126 (1998)
50. D. Anderson, PoS **ICHEP2016**, 190 (2016)
51. S. Mukherjee, Data scouting: a new trigger paradigm, in *5th Large Hadron Collider Physics Conference* (2017). arXiv:1708.06925
52. J. Duarte, Fast reconstruction and data scouting, in *4th International Workshop Connecting The Dots 2018 (CTD2018) Seattle, Washington, USA, March 20–22, 2018* (2018). arXiv:1808.00902
53. CMS Collaboration (2017). http://opendata.cern.ch/record/6021. Accessed May 2020
54. A. Sirunyan et al. (CMS), JINST **12**, P10003 (2017). arXiv:1706.04965
55. CMS Collaboration (2017). http://opendata.cern.ch/record/9863. Accessed May 2020
56. CMS Collaboration (2017). http://opendata.cern.ch/record/9864. Accessed May 2020
57. CMS Collaboration (2017). http://opendata.cern.ch/record/9865. Accessed May 2020
58. CMS Collaboration (2017). http://opendata.cern.ch/record/7719. Accessed May 2020
59. CMS Collaboration (2017). http://opendata.cern.ch/record/7721. Accessed May 2020
60. CMS Collaboration (2017). http://opendata.cern.ch/record/7722. Accessed May 2020
61. CMS Collaboration (2017). http://opendata.cern.ch/record/7723. Accessed May 2020
62. CMS Collaboration (2017). http://opendata.cern.ch/record/9588. Accessed May 2020
63. S. Agostinelli et al. (GEANT4), Nucl. Instrum. Methods **A506**, 250 (2003)