Eur. Phys. J. C (2020) 80:1055
https://doi.org/10.1140/epjc/s10052-020-08629-w

The European
Physical Journal C

Regular Article - Theoretical Physics

Invisible Higgs search through vector boson fusion: a deep learning approach

Vishal S. Ngairangbam[1,2,a], **Akanksha Bhardwaj**[1,b], **Partha Konar**[1,c], **Aruna Kumar Nayak**[3,d]

[1] Theoretical Physics Division, Physical Research Laboratory, Ahmedabad 380009, India
[2] Indian Institute of Technology, Gandhinagar, Gujarat 382424, India
[3] Institute of Physics, Bhubaneswar, Odisha 751005, India

Received: 26 August 2020 / Accepted: 2 November 2020 / Published online: 16 November 2020

Abstract Vector boson fusion proposed initially as an alternative channel for finding heavy Higgs has now established itself as a crucial search scheme to probe different properties of the Higgs boson or for new physics. We explore the merit of deep-learning entirely from the low-level calorimeter data in the search for invisibly decaying Higgs. Such an effort supersedes decades-old faith in the remarkable event kinematics and radiation pattern as a signature to the absence of any color exchange between incoming partons in the vector boson fusion mechanism. We investigate among different neural network architectures, considering both low-level and high-level input variables as a detailed comparative analysis. To have a consistent comparison with existing techniques, we closely follow a recent experimental study of CMS search on invisible Higgs with 36 fb^{-1} data. We find that sophisticated deep-learning techniques have the impressive capability to improve the bound on invisible branching ratio by a factor of three, utilizing the same amount of data. Without relying on any exclusive event reconstruction, this novel technique can provide the most stringent bounds on the invisible branching ratio of the SM-like Higgs boson. Such an outcome has the ability to constraint many different BSM models severely.

1 Introduction

With the emergence of deep learning frameworks, a plethora of machine learning applications have gained immense importance in high-energy physics (HEP) recently, in collider and neutrino physics [1–3]. Supported by substantial multilateral developments in this field, efforts are being poured in to explore different aspects of HEP phenomenology, especially in the context of the Large Hadron Collider (LHC) [4–9]. In recent years, deep learning applications have been widely explored to understand hadronic jets' formation and properties, the most common structured object found in any event at LHC, created from QCD fragmentation and hadronization of fundamental quarks and gluons. More interestingly, boosted heavy particles like Higgs, top or massive gauge bosons can also produce similar jet objects after the hadronization of their decay products. Prior to the advent of deep-learning approaches, the realization that the internal dynamics of different jet objects are dissimilar received intense scrutiny [10–15] looking into the underlying structures as probes for new physics [16–23]. For jet substructure studies, the primary deep-learning approach is to employ calorimeter energy deposits of a jet in $\eta - \phi$ pixel tower converted into the pictorial description of such 'jet-images' [24] as input to Convolutional Neural Network (CNN) [25–27]. Very successful n-prong taggers are developed for Z /W bosons [28] and the top tagging [29–31] by utilizing this idea, which is further extended to discriminate between quark and gluons [27]. Contrary to jet-images, various other approaches have also been explored for the input space. These include looking for the optimal basis of substructure variables in N-body phase space [32], forming the jet-spectra with two-point correlations at different angular ranges [33,34], and making an analogy of collider events with natural language thereby using recursive neural networks for feature extraction [35]. Deep Neural Networks (DNN) have established their importance for classification of signal and background using low/high-level variables [36–45]. Although there are some studies [46–49] of utilizing the inclusive event information at hadron colliders as input for deep-learning neural networks, their full potential are yet to be explored extensively. For the benefit of the readers, many more such exciting approaches

[a] e-mail: vishalng@prl.res.in (corresponding author)
[b] e-mail: akanksha@prl.res.in
[c] e-mail: konar@prl.res.in
[d] e-mail: nayak@iopb.res.in

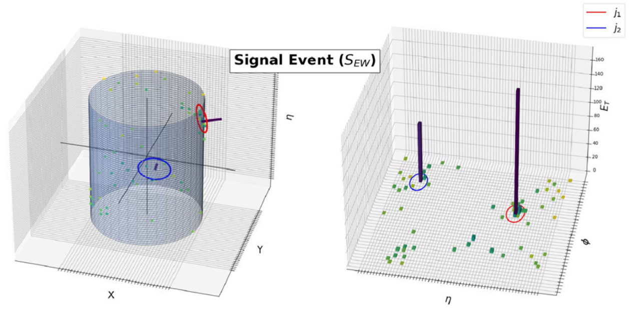


Fig. 1 The figure shows a 3D depiction of a prototype signal event originated from an electroweak VBF Higgs production in a naive detector geometry in left plot. The same event is flattened in a convenient $\eta - \phi$ plane in right plot, where the transverse projection of calorimeter energy deposits in different pixels are drawn. Two reconstructed primary jets are shown with color circles, and corresponding transverse energy deposits are visible from height of the bars

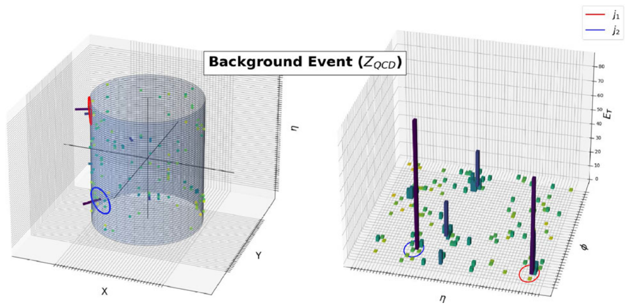


Fig. 2 Same as Fig. 1, but for a prototype background event originated from a $Z(\nu\bar{\nu}) + jets$ production, where the jets originate from QCD vertices

in the machine learning framework can be followed in the recent review [50–52] and references therein.

Taking an analogy from jet-image classification, we use the full calorimeter image to study the invisible Higgs production in association with a pair of jets. Vector-boson fusion (VBF) production of color singlet particles provide a unique signature in hadron colliders. First studied in reference [53–55], they are characterized by the presence of two hard jets in the forward regions with a large rapidity gap, and a relative absence of hadronic activity in the central regions, when the singlet particle decays non-hadronically. For illustration, the left panel of Fig. 1 shows an event of a Higgs produced in VBF channel decaying invisibly in a simplistic tower geometry, while the same event is mapped in a flattened (η, ϕ) plane by rolling out the ϕ-axis, with the height of the bars corresponding to the magnitude of the transverse projection of calorimeter energy deposits in each pixel. In order to highlight the differences with non-VBF processes, it is instructive to show one such example in Fig. 2. This is a representative event from $Z(\nu\bar{\nu}) + jets$ background, where the jets arise from QCD vertices, which inherently has a much higher hadronic activity in the central regions between the two leading jets. Even though the rapidity gap vanishes when the singlet particle decays hadronically, the absence of color connection between the two forward jets and the central region persists and has been used in the experimental analysis [56], in searches of the Higgs boson decaying to bottom quarks. The VBF process was proposed as the most important mechanism for heavy Higgs searches [57] thanks to a much slower fall in cross-section compared to the s-channel mediated process. Usefulness for intermediate to light mass scalar was also subsequently realized [58] due to its unique signature at the collider. VBF process holds great importance to measure Higgs coupling with gauge bosons and fermions as it allows independent observations of Higgs decay like $h^0 \rightarrow WW$ [59], $h^0 \rightarrow \tau\tau$ [60]. Therefore, it also plays a vital role in determining anomalous coupling to vector boson [61,62] or the CP properties of the Higgs [63,64]. Its clean features make it the most sensitive channel for searching invisible decay of the Higgs boson [65] and in search for physics beyond the standard model [66–68]. As the Higgs can decay invisibly only through a pair of Z bosons producing neutrinos with minuscule branching ratio in the Standard Model (SM), observation of any significant deviation can provide a strong indication towards a theory beyond the Standard Model (BSM) [69]. Hence, this search plays a crucial role to constrain many BSM scenarios, like dark-matter [70–74], massive neutrinos [75,76], supersymmetric [77,78], and extra-dimensional models [79,80].

Although being one of the most promising channels, the production of invisible Higgs is challenging to probe as only a few observables can be constructed over the unique features of VBF. Ensuring a color quiet central region by so-called 'central jet veto,' and rather specific choices related to the jets, the separation in pseudorapidity $|\Delta\eta_{jj}|$ and the dijet invariant mass m_{jj} are the significant ones. A central jet veto essentially discards events with additional jets in the region between the two forward tagging jets. Electroweak VBF production of Higgs can satisfy such criteria naturally with excellent efficiency. These same criteria can also ensure the elimination of vast QCD backgrounds up to a large extent, where jets are produced with a massive W or Z boson decaying (semi)invisibly. Finally, the much weaker electroweak backgrounds coming in the form of VBF production of W or Z, become the dominant factor for such study. However, we must note at the same time that there is a significant drop in signal contribution from other dominant non-VBF Higgs production modes, such as higher-order in α_s correction to gluon fusion initiated processes for Higgs productions [81].

A natural order of inquiry, therefore, calls for the investigation of whether deep machine learning vision in the form of CNN, together with other neural networks, can have the ability to recognize the characteristics of VBF by learning from the data itself. Networks would map the probability distribution functions to characterize each process by utilizing low-level or high-level variables. Moreover, we would like to understand how useful these learned features are or how they correlate with our traditional characteristics of VBF. Finally, there can be enough scope to engage this very sophisticated tool to get some hybrid output in terms of maximizing the efficiency in selecting signal events, rather than classifying in separate clusters as VBF and non-VBF types.

While the present study can easily be extended for other decay modes of Higgs, we choose the invisible channel for our study to showcase the importance of deep learning quantitatively using different neural network architectures. We propose to study the full event topology of VBF by examining the calorimeter tower-image using CNN, which utilizes the low-level variables. We also consider the performance of event classification using dense Artificial Neural Networks (ANN), which employ high-level variables. In total, we investigate seven different neural network architectures and provide a comparative study of the performance of networks. The performance of networks is quantified in terms of expected constraints on the invisible branching ratio (BR ($h^0 \to$ inv)) of the Higgs boson.

The latest report from ATLAS collaboration [82] puts an upper limit on BR(h^0 →inv) at 95% confidence level (CL) to 0.13, from an integrated luminosity of 139 fb^{-1} at the LHC. The CMS analysis also puts an upper limit at 95% CL to 0.19 for combined data set of 7, 8 and 13 TeV for 4.9 fb^{-1}, 19.7 fb^{-1} and 38.2 fb^{-1} integrated luminosities respectively [83]. These bounds still allow the significant presence of BSM physics. Our principal aim, therefore, is to study the viability of CNNs to improve these results using low-level variables in the form of the entire calorimeter image, as well as to compare its performance to DNN/ANN architecture with high-level variables as input. We find that the bounds on the BR(h^0 →inv) can indeed be significantly improved using these networks.

The rest of this paper is organized as follows. In Sect. 2, we discuss the Higgs production mechanism via the VBF channel and different SM backgrounds contributing to this process. We also discuss the generation of simulated data consistent with the VBF search strategy. In Sect. 3, we describe the details of the data representation used in the present study. Here, different classes of high-level variables are also defined. Preprocessing methods of feature spaces are addressed in Sect. 4. We discuss the seven different neural network architecture and its performance in Sect. 5. The results, interpreted in terms of expected bounds on the invisible branching ratio, for all the architectures are presented in Sect. 6. There, we also discuss the impact of pileup on the result of our analysis. Finally, we close our discussion with the summary and conclusion in the last section.

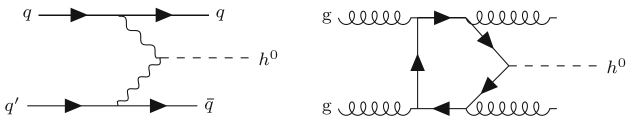


Fig. 3 Representative diagrams for production of Higgs signal through (left) electroweak VBF channel and (right) a higher-order QCD process in gluon fusion where two QCD jets can be detected along with a sizable missing transverse-energy from invisible Higgs decay

2 Vector boson fusion production of Higgs and analysis set-up

VBF production of the SM Higgs has the second-highest production cross-section after gluon-fusion at the LHC. Loop induced Higgs production and decay depend on the presence of contributing particles and different modifiers in fermions and gauge boson coupling with the scalar. Hence, both production cross-section and decay branching ratios are modified in the presence of new physics. In this present work, we consider the production of SM like Higgs boson and constrain its invisible decay width. Such constraint is essential in many new physics scenarios, such as Higgs portal dark matter [70–74], where new particles do not modify their couplings with SM particles.

The electroweak production of Higgs is dominated by the fusion of two massive vector bosons, which are radiated off two initial (anti-)quarks, as represented in Fig. 3 (left plot). This exchange of color singlet state between two quarks ensures no color connection between two final jets, typically produced in a forward (backward) region of the opposite hemisphere. The central region – between these two jets remain color quiet, lacking any jet activity even after radiation and fragmentation of the two scattered quarks while looking at the hadronic final states. As we have already discussed, an agnostic viewpoint requires a serious re-examination after the inclusion of all other processes, such as non-VBF Higgs signal from gluon fusion. One such sample diagram is shown in Fig. 3 (right plot). Additional radiation from gluons can provide a typical VBF type signal, once again, in the absence of the key attributes like color-quiet central region, etc.

Another interesting feature of VBF Higgs production is that the corresponding cross-section has very modest correction under higher-order QCD, which has been known for a long time [84,85]. Integrated and differential cross sections for VBF Higgs production have now been calculated up to very high levels of accuracy. QCD corrections are known up to N3LO [86], reducing the scale-uncertainty up to 2%, while Electroweak corrections are known up to NLO [87]. Moreover, non-factorizable contributions have also been calculated for the first time [88], and show up to percent level corrections compared to the leading order (LO) distributions.

At hadron colliders, traditional searches [89–92] of non-hadronically decaying color-singlet particles in the VBF production channel focus on rejecting the large QCD backgrounds from $Z + jets$, and $W + jets$ background via a *central jet-veto*, after a hard cut on the separation of the two forward jets in pseudorapidity $|\Delta\eta_{jj}|$, and the dijet invariant mass m_{jj}. This opens up the possibility of using inclusive event-shape variables like N-jettiness [93], to improve the selection efficiency [94]. In this study, we explore the feasibility of using deep-learning techniques instead of event-shape variables. We study the invisible decay of the Higgs boson as a prototype channel for gauging the power of deep-learning methods in VBF since there is no contamination on the radiation patterns between the two forward jets from the decay products. We closely follow the shape-based analysis performed by the CMS experiment at CERN-LHC [83].[1] As already commented, the central jet veto played a critical role in the usual searches of VBF to control the vast QCD background. The role of additional information from QCD radiation between the tagging jets and within the jet itself was explored in references [96,97]. It was found that relaxation of the minimum p_T requirement of the central jet improved the sensitivity, and the inclusion of subjet level information resulted in further suppression of backgrounds. However, the present analysis does not rely on a central jet veto, as the main aim is to study the VBF topology with the low-level data, made possible with modern deep-learning algorithms. Therefore, with the relaxed selection requirements on $|\Delta\eta_{jj}|$ and m_{jj}, the selected signal gets a significant contribution from the gluon-fusion production of Higgs on top of VBF processes. Due to the relaxed selection criteria, we also get a substantial contribution from QCD backgrounds.

2.1 Signal topology

The present study relies on all dominant contributions to Higgs coming both from electroweak VBF processes and also higher-order in QCD gluon fusion processes. Here at least two jets should be reconstructed along with sizable missing transverse-energy from invisible decay of Higgs. Hence, we classify the full signal contribution in two channels:

- S_{QCD}: Gluon-fusion production of Higgs with two hard jets, where the Higgs decays invisibly.
- S_{EW}: Vector-Boson fusion production of Higgs decaying invisibly.

The subscript $EW(QCD)$ denotes the absence (presence) of strong coupling α_S, at leading order(LO) for the interested topology. This also segregates the channels with absence or presence of color exchange between the two incoming partons at LO. Figure 3 shows a representative Feynman diagram of the signal channels in each class.

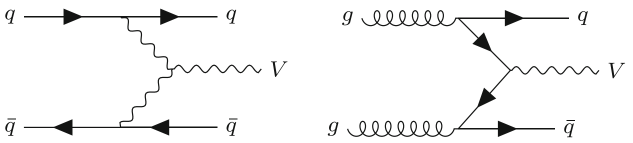


Fig. 4 Representative diagrams for dominant background processes through (left) VBF type weak production and (right) QCD production of massive vector-bosons V, such as W or Z which decay invisibly by producing undetected lepton or neutrinos

2.2 Backgrounds

The major backgrounds contributing to the invisible Higgs VBF signature can come from the different standard model processes. Among them, VBF type electroweak, and QCD production of massive vector-bosons (W or Z) contribute copiously. All these processes ensure a pair of reconstructed jets along with considerable missing transverse energy from invisible decay of these gauge bosons. A substantial fraction of W and Z can produce neutrinos or a lepton which remain undetected at the detector. We consider the following backgrounds in all our analyses:

- Z_{QCD}: $Z(\nu\bar{\nu}) + jets$ process contributes as the major SM background due to high cross section.
- W_{QCD}: $W^{\pm}(l^{\pm}\nu) + jets$ process also contribute to the SM background when the lepton is not identified.
- Z_{EW}: Electroweak production of Z decaying invisibly along with two hard jets is topologically identical with the electroweak signal and contributes significantly to the background.
- W_{EW}: Electroweak production of $W^{\pm}$ with two hard jets can also produce an identical signal when the lepton does not satisfy the identification criteria.

Similar to the signal processes, the subscript $EW(QCD)$ denotes the absence (presence) of strong coupling α_S, at LO for the interested topology having at least two reconstructed jets in the final state. Figure 4 shows representative Feynman diagrams of the background channels divided into four different classes.

There are also other background processes like top-quark production, diboson processes, and QCD multijet backgrounds whose contribution would be highly suppressed compared to these four backgrounds. The top and diboson backgrounds would contribute to leptonic decay channels where charged leptons, if present, are not identified, while the QCD multijet background would contribute when there is severe mismeasurement of the jet energies.

[1] For ATLAS analysis with similar data, see Ref. [95].

2.3 Simulation details

We used MadGraph5_aMC@NLO(v2.6.5) [98] to generate parton-level events for all processes at 13 TeV LHC. These events are then showered and hadronized with Pythia(v8.243) [99]. Delphes (v3.4.1) [100] is used for fast-detector simulation of the CMS working conditions. Jets are clustered using the FastJet(v3.2.1) [101] package. The signal processes are generated using a modified version of the Higgs Effective Field Theory (HEFT) model [102–104], where the Higgs boson can decay to a pair of scalar dark matter particle at tree level. We are interested in probing high transverse momentum of Higgs, where the finite mass of top quark in gluon fusion becomes essential. Hence, we have taken into account such effect by reweighting the missing transverse energy (MET) distribution of the events with recommendations from reference [105]. The parton level cross-sections of Z_{QCD} and W_{QCD} were also matched up to four and two jets, respectively, via the MLM procedure [106]. Since the $W^{\pm}$ backgrounds contribute when the leptons are missed within the range of tracker or when they are not reconstructed at the detector, the parton level cuts on the generated leptons are removed to cover the whole range in pseudorapidity (η).

For a consistent comparison with current experimental results, we repeat the shape-analysis of reference [83] with our simulated dataset. The MET cut for the deep-learning study is relaxed from 250 to 200 GeV.

Baseline selection criteria: We apply the following pre-selections:

- **Jet** p_T: At least two jets with leading (sub-leading) jet having minimum transverse momentum $p_T > 80$ (40) GeV.
- **Lepton-veto:** We veto events with the reconstructed electron (muon) with minimum transverse momentum $p_T > 10$ GeV in the central region, *i.e.* $|\eta| < 2.5$ (2.4). This rejects leptonic decay of single $W^{\pm}$, and semi-leptonic $t\bar{t}$ backgrounds.
- **Photon-veto:** Events having any photon with $p_T > 15$ GeV in the central region, $|\eta| < 2.5$ are discarded.
- τ **and b-veto**: No tau-tagged jets in $|\eta| < 2.3$ with $p_T > 18$ GeV, and no b-tagged jets in $|\eta| < 2.5$ with $p_T > 20$ GeV are allowed. This rejects leptonic decay of single $W^{\pm}$, semi-leptonic $t\bar{t}$ and single top backgrounds.
- **MET**: Total missing transverse momentum for the event must satisfy MET > 200 GeV for all our deep-learning study, whereas we compared CMS shape-analysis consistent with MET > 250 GEV.
- **Alignment of MET with respect to jet directions**: Azimuthal angle separation between the reconstructed jet with the missing transverse momentum to satisfy $\min(\Delta\phi(\mathbf{p}_T^{\text{MET}}, \mathbf{p}_T^{j})) > 0.5$ for up to four leading jets with $p_T > 30$ GeV and $|\eta| < 4.7$. QCD multi-jet background that arises due to severe mismeasurement is reduced significantly via this requirement.
- **Jet rapidity**: We require both jets to have produced with $|\eta_j| < 4.7$, and at least one of the jets to have $|\eta_{j_i}| < 3$, since the L1 triggers at CMS do not use the information from the forward regions.
- **Jets in opposite hemisphere**: Those events which have the two leading jets reside in the opposite hemisphere in η are selected. This is done by imposing the condition $\eta_{j_1} \times \eta_{j_2} < 0$.
- **Azimuthal angle separation between jets**: Events with $|\Delta\phi_{jj}| < 1.5$ are selected. This helps in reducing all non-VBF backgrounds.
- **Jet rapidity gap**: Events having minimum rapidity gap between two leading jets $|\Delta\eta_{jj}| > 1$ are selected.
- **Di-jet invariant mass**: We required a minimum invariant mass of two leading jets, $m_{jj} > 200$ GeV. Note that, this along with the previous selection requirements are relatively loose compared to traditional selection criteria of VBF topologies, which result in significant enhancement of the signal from S_{QCD}, although at the cost of increased QCD backgrounds (Z_{QCD} and W_{QCD}).

Interestingly, one can notice that a relaxed selection requirement may give rise to additional contamination from Higgs-strahlung type topologies to the S_{EW} channel, which is included in our EW generation of events. However, these events are not expected to survive a selection of di-jet invariant masses of more than 200 GeV. After extracting the events passing the above selection requirements and the respective selection efficiency (calculated from the weights) for S_{QCD}, the pre-selected events are unweighted again so that we get equal weights for individual events.[2] The background and signal classes are formed by mixing the channels with the expected proportions using appropriate k-factors, cross-sections, and the baseline selection efficiencies. We use cross-sections quoted in reference [105] for both signal processes. For instance, the S_{QCD} is calculated up to NNLL +NNLO accuracy [107], while for S_{EW} it is calculated up to NNLO [108] in QCD and NLO in electroweak. We use the LO distributions with their overall normalizations increased to accommodate the total cross-section at higher perturbative accuracies without accounting for the possible change in shape. Similarly, all background cross sections are calculated by scaling the LO result with global NLO k-factors [109,110]. We generated 200,000 training and 50,000 validation balanced dataset of events for the deep-learning classifier. The signal class consists of 44.8% S_{EW} and the 55.2% S_{QCD} channels; while the background class consists of 51.221%

[2] See Appendix A, for distribution of the important kinematic-variables and details of the re-weighting and unweighting of events.

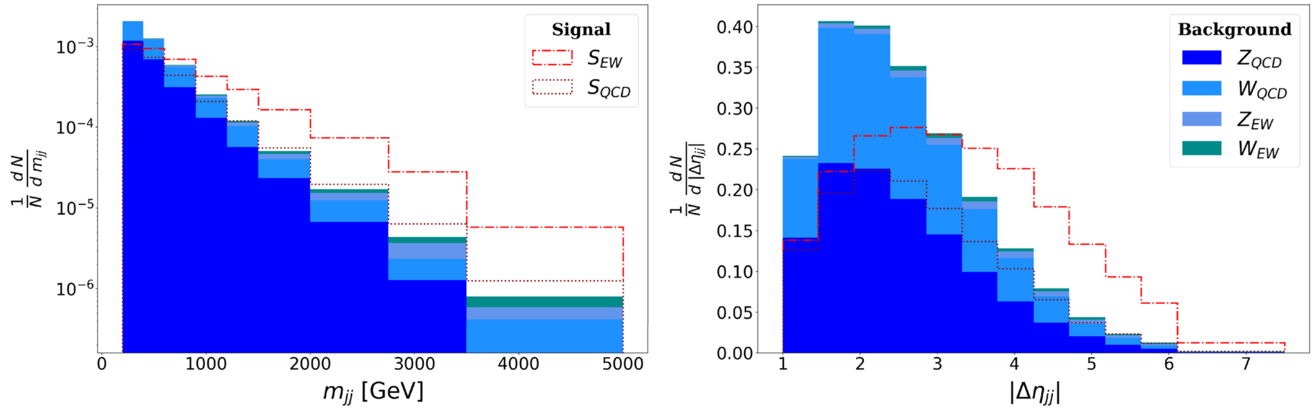


Fig. 5 Distribution of (left) m_{jj} and (right) $\Delta\eta_{jj}$ of events passed after the passing the tighter selection requirement (MET $>$ 250 GeV). The contribution of each channel to its parent class has been weighted by their cross-sections and the baseline efficiency at 13 TeV. The signal and backgrounds are then individually normalized, and the lines/color show the contribution of each channel to its parent class

Z_{QCD}, 44.896% W_{QCD}, 2.295% Z_{EW} and 1.587% W_{EW} channels.

We also extract event sample for all channels with the harder selection requirement on missing transverse momentum (MET $>$ 250 GeV), the value used in reference [83], from the same set of generated events used for the deep-learning analysis. The extracted dataset contains: 39% S_{EW} and the 61% S_{QCD} channels for the signal class; and 54.43% Z_{QCD}, 40.92% W_{QCD}, 3.05% Z_{EW} and 1.58% W_{EW} channels for the background class. The bin-wise stacked histogram of all channels for m_{jj} and $|\Delta\eta_{jj}|$ are shown in Fig. 5. The properties of the EW and the QCD subsets are evident from these distributions: EW contribute more at higher m_{jj} and $|\Delta\eta_{jj}|$, while the opposite is true for QCD.

3 Data representation for the network

Neural network architectures for deep-learning are mostly designed with two blocks. The first stage generally consists of locally-connected layers (with or without weight sharing) with some particular domain level specifications which extract the features. The second stage consists typically of densely connected layers, whose function is to find a direction in the learned feature-space, which optimally satisfies the particular target of the network locally by learning its projections in different representations at each subsequent layer. For instance, in classification problems, it finds the decision boundary between different classes. At the same time, in an unsupervised clustering, it compresses the feature-space so that the modes become localized in a smaller volume. A synergy between the representation of data and the network architecture is a must for efficient feature extraction. This is evident from the fact that convolutional neural networks perform best with data structures that have an underlying Euclidean structure, while recurrent networks work best with sequential data structures. In the context of classifying boosted heavy particles like W, Higgs, top quark or heavy scalars decaying to large-radius jets from QCD background, a lot of efforts [24,25,27–29,111] went into representing the data like an image in the (η, ϕ) plane to use convolutional layers for feature extraction, while some others [112,113], use physics-motivated architectures. Convolutional architectures work in these cases because the differences between the signal jet and the background (QCD) follows a Euclidean structure.[3] The Minkowski structure of space-time prohibits a direct use of convolutional architectures. Although geometric approaches [114] exist to counter the non-Euclidean nature, the number of dimensions makes it computationally expensive. Graph neural networks [115–118] provide a possible workaround which is computationally less intensive, for feature learning in non-Euclidean domains.

In the present work, we want to study the difference in radiation patterns between the two forward jets for signal and background events; hence, we primarily choose a convolutional architecture for automatic feature extraction. Therefore, the low-level feature space we prefer is the *tower-image*, in the (η, ϕ)-plane, with the transverse energy E_T, as the pixel values. One can take into account the different resolutions in the central and forward regions of calorimeter towers in LHC detectors. For simplicity, and also to demonstrate the resolution dependence, we construct two images - a high-resolution image with bin size 0.08×0.08, and a low-resolution image with bin size 0.17×0.17, in the full range of the tower, $[-5, 5]$ for η and while $[-\pi, \pi)$ for ϕ. Convolutional neural-networks, in general, look at global differences, and increasing the resolution does not play as important a

[3] Most high-level variables designed from QCD knowledge are functions of $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$.

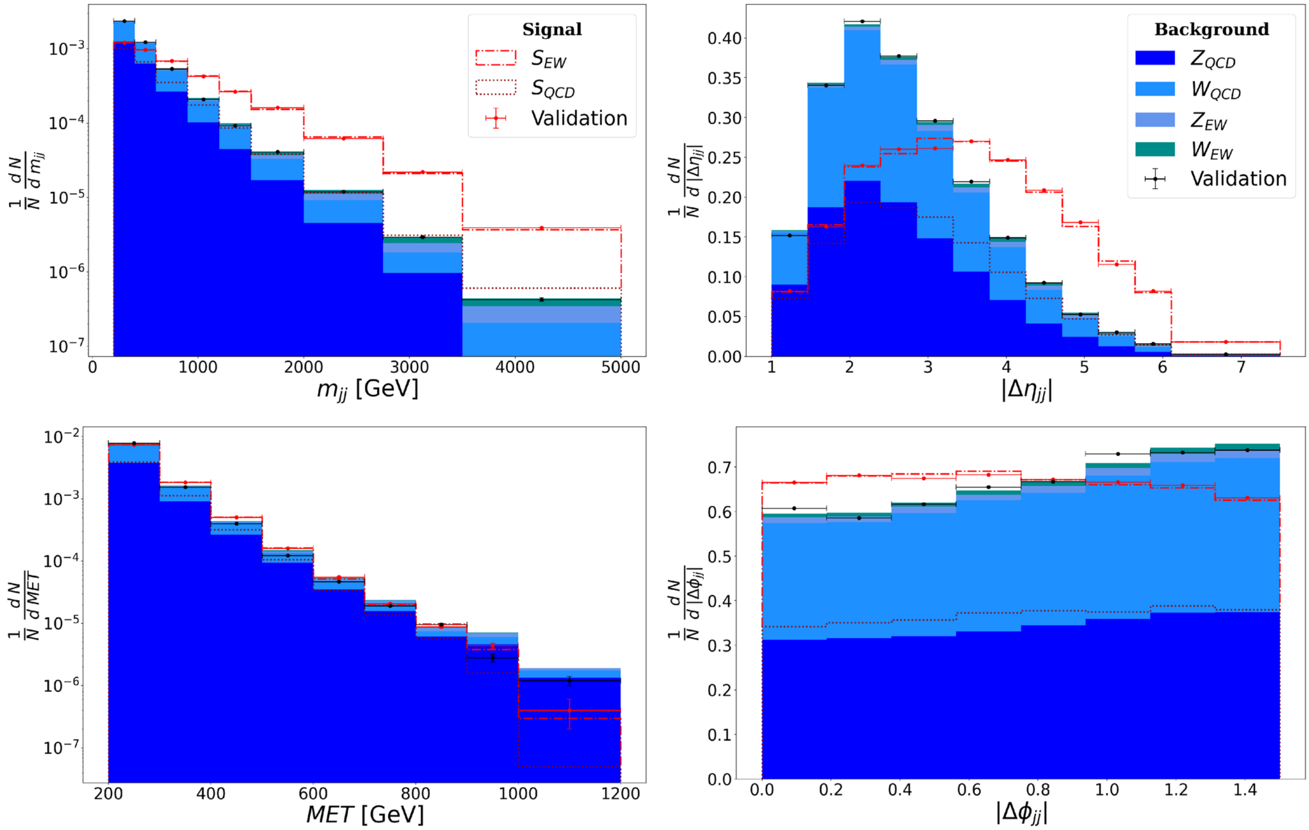


Fig. 6 Similar to Fig. 5, some of the basic input high-level kinematic variables used for our analysis (MET > 200 GeV) are shown for signal and background

role. We examine CNNs in these two different resolutions to inspect this for our particular case. The procedure of forming a tower-image does not naturally take the periodicity of the ϕ axis into account. In order to let the network know this inherently, we expand the image obtained after binning, in the ϕ axis such that the connectivity between the two edges is not broken. This is done by taking a predetermined number of ϕ-rows from each edge of the original image and forming a new image where these rows are padded [46,49] in their corresponding opposite sides, thereby mimicking the periodicity. This is similar to cutting the cylinder at two different points in ϕ for each edge, such that there is an overlapping region in the final image. Taking the jet radius $R = 0.5$, which have a regular geometry since they are clustered with anti-k_t algorithm [119], we choose the number of rows to be 4 (8) for the low (high)-resolution images, with one bin as a buffer. This gives a low-resolution (LR) image of 59 × 45 and a high-resolution (HR) one of 125 × 95.

A significant difference between low-level and high-level feature spaces is that the modes of the data in low-level representations are not distinct. Although this is marginally enhanced by preprocessing, high-level features derived from the said low-level features have distinctly localized modes in their distribution. An exemplary ability of deep-learning algorithms is to by-pass this step and learn their own representations which perform better than the high-level variables developed by domain-specific methods. To analyze the relative performance of physics-motivated variables derived from the calorimeter deposits, we consider two classes of high-level variables. The first one consists of the following kinematic variables:

$$\mathcal{K} \equiv (\,|\Delta\eta_{jj}|,\ |\Delta\phi_{jj}|\,,\ m_{jj}\ ,\ \text{MET}\ ,\ \phi_{\text{MET}}\ ,\ \Delta\phi^{j_1}_{\text{MET}}\ ,\ \Delta\phi^{j_2}_{\text{MET}}\ ,\ \Delta\phi^{j_1+j_2}_{\text{MET}}\ ) \qquad (1)$$

ϕ_{MET} is the azimuthal direction of MET in the lab-frame. $\Delta\phi^{j_1}_{\text{MET}}$, $\Delta\phi^{j_2}_{\text{MET}}$ and $\Delta\phi^{j_1+j_2}_{\text{MET}}$ are the azimuthal separation of MET with the direction of the leading, sub-leading and the vector sum of these two jets, respectively. Clearly, these do not contain any information about the radiation pattern between the tagging jets. The second class of variables: the sum of E_T of the tower constituents in the interval $[-\eta_C, \eta_C]$, incorporates this information:

$$\mathcal{R} \equiv (H_T^{\eta_C} | \eta_C \in \mathcal{E}) \quad , \quad H_T^{\eta_C} = \sum_{\eta < |\eta_C|} E_T \quad . \qquad (2)$$

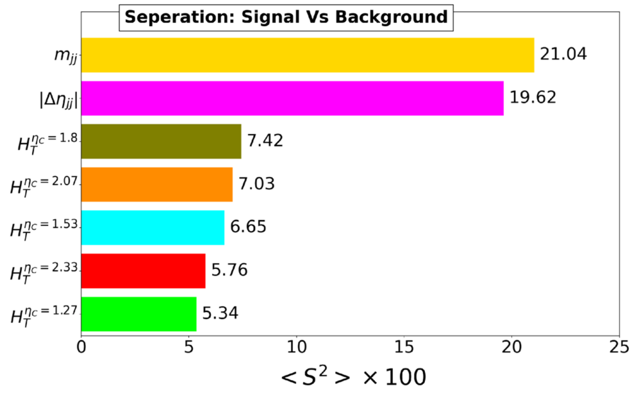


Fig. 7 The separation of the 7 highest performing variables (given in percentage)

$\mathcal{E}$ denotes the set of chosen η_C's. We vary η_C uniformly in the interval [1,5]:

$$\mathcal{E} = \{1, 1.27, 1.53, 1.8, 2.07, 2.33, 2.6, 2.87, 3.13, 3.4, 3.67, 3.93, 4.2, 4.47, 4.73, 5\}, \tag{3}$$

to get 16 such variables. Their inclusion helps us to provide a thorough comparison of the high-level and low-level feature spaces. Figure 6 shows the signal vs background distribution of some important kinematic-variables. The channel-wise contributions to the parent class are also stacked with different colors/lines. We see that the characteristics of the m_{jj} and $|\Delta\eta_{jj}|$ are the same with Fig. 5, with the electroweak processes contributing more at higher values. A feature seen for $|\Delta\phi_{jj}|$ is the shape of the signal and background distributions. Clearly, the difference is due to the S_{EW} contribution since S_{QCD} has a very similar shape as that of the background. This is another characteristic of VBF processes that the leading jets, originating from electroweak vertices, have lower separation in ϕ compared to those originating from QCD. Similar plots for the remaining four kinematic variables and the $\mathcal{R}$ set of variables are shown in Figs. 19 and 17 in Appendix B. A brief discussion of the two feature spaces (mainly $\mathcal{R}$) is also presented. We denote the combined high-level feature-space as $\mathcal{H}$, which is 24-dimensional.

In order to gauge the discriminating power of each feature x, we determine the separation [120] defined as,

$$\langle S^2 \rangle = \frac{1}{2}\int \frac{(p_S(x) - p_B(x))^2}{p_S(x) + p_B(x)}\, dx \quad . \tag{4}$$

$p_S(x)$ and $p_B(x)$ denote the normalized probability distribution of the signal and background classes. It gives a classifier-independent discrimination power of the feature x. A value of zero (one) denotes identical (non-overlapping) distributions. We plot the separation (in percentage) of the seven highest important variables out of the 24 features in Fig. 7. It is interesting to note that out of these, there are five variables from $\mathcal{R}$, even though the first and the second are from $\mathcal{K}$, and they are much greater in magnitude.

4 Preprocessing of feature space

Preprocessing of features is indispensable for shallow machine learning as it helps maximize the statistical output from smaller data sizes. In deep-learning applications, it helps in faster convergence of the training and in approaching optimal accuracy with a lesser amount of data using simpler architectures. Even though the primary aim of our model is to learn the differing QCD radiation patterns, we can only devise preprocessing operations that preserve the Lorentz symmetries of the event. The spatial orientation of the events, in general, can be regularized by the following procedure:

1. **Identify principal directions:** Choose three final-state directions $\{\hat{n}_1, \hat{n}_2, \hat{n}_3\}$. These can be any three final state objects, which are the interest of our studies like photons, leptons, and jets, or they can be chosen to be generic directions in the lab frame.
2. **First Rotation:** Rotate the event such that:

   $$\hat{n}_1 \to \hat{n}_1' = (0,0,1) \equiv \hat{n}^a \quad , \quad \hat{n}_2 \to \hat{n}_2' \quad , \quad \hat{n}_3 \to \hat{n}_3' \quad .$$

   After this operation, the orientation of $\hat{n}_1$ is the same for all events.
3. **Second Rotation:** Rotate the event along $\hat{n}^a$ such that:

   $$\hat{n}_2' \to \hat{n}_2'' = (0, n_y^b, n_z^b) \equiv \hat{n}^b \quad , \quad \hat{n}_3' \to \hat{n}_3'' \quad .$$

   The plane formed by $\hat{n}_1$ and $\hat{n}_2$ has the same orientation for all events after this operation.
4. **Reflection:** Reflect along yz-plane such that:

   $$n_3'' \to (|n_x^c|, n_y^c, n_z^c) \equiv \hat{n}^c \quad .$$

   The half-space containing $\hat{n}_3$ becomes the same for all events after this step.

These are passive operations which affect the orientation of the reference frame without changing the physics. For most event topologies, we can see that there will be better feature regularisation when $\hat{n}_2$ and $\hat{n}_3$ are equal. In hadron colliders, due to the unknown partonic center-of-mass energy $\sqrt{\hat{s}}$, we set the z-axis as $\hat{n}_1$, preserving the transverse momentum of all final state particles. We choose two different instances of $\hat{n}_2 \in \{\hat{n}_{\text{MET}}, \hat{n}_{j_1}\}$. For our choice of $\hat{n}_1$, the z-direction of $\hat{n}_2$ does not matter and we can take its value for $\hat{n}_{\text{MET}}$ to be zero. However, the z-direction becomes important for the third operation and we choose $\hat{n}_3 = \hat{n}_{j_1}$. This translates to

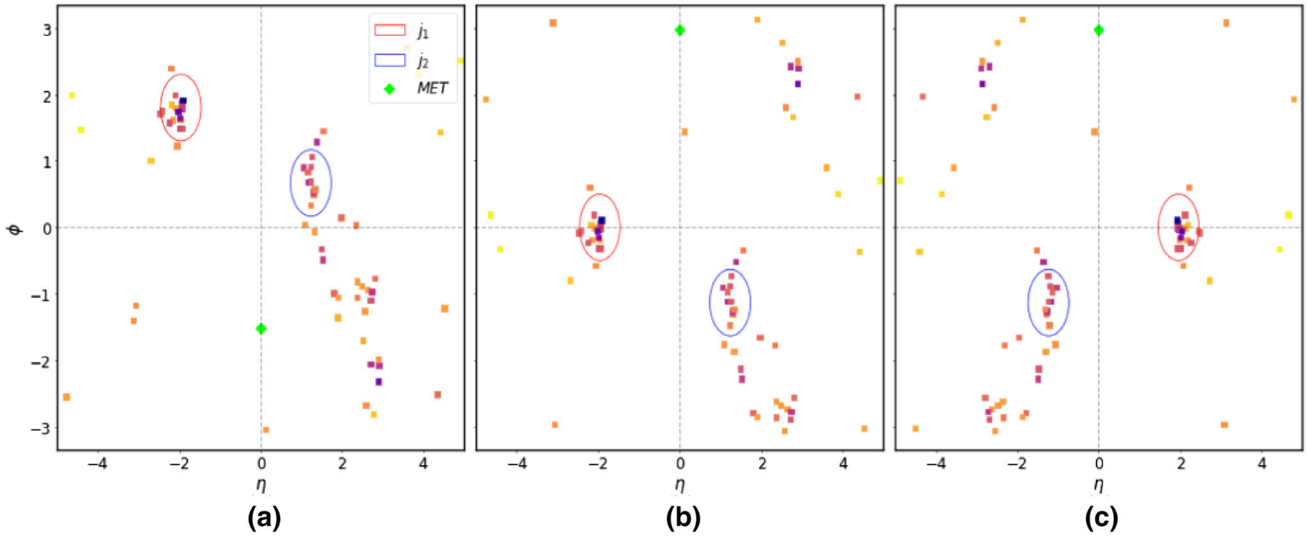


Fig. 8 Scatter plot of tower constituents of an event in the (η, ϕ)-plane showing: (a) the raw event; and the effects of (b) rotation $(\phi_{j_1} = 0)$, and (c) reflection $(\eta_{j_1} > 0)$ operations. The pseudorapidity of MET has been set to zero for illustration. It is important to note that the points here are not binned into pixels and the values are the ones extracted from the Delphes Tower constituents

applying the following operations to the four-momenta of each events:

1. Rotate along z-axis such that $\phi_0 = 0$. We choose two instances of $\phi_0 \in \{\phi_{\text{MET}}, \phi_{j_1}\}$.
2. Reflect along the xy-plane, such that the leading jet's η is always positive.

After these two steps, the tower-constituents are binned in the resolutions as mentioned earlier and then padded on the ϕ-axis. We denote the feature-spaces obtained after preprocessing with the two instances of ϕ_0 as $\mathcal{P}_{\text{MET}}$ and $\mathcal{P}_J$. Figure 8 shows the different steps of preprocessing steps for an event taking $\phi_0 = \phi_{j_1}$. Averaged low-resolution image of the validation dataset of each class without preprocessing, and for both instances of ϕ_0 are shown in Fig. 9. As emphasized earlier, it is seen that there is a better regularization when $\hat{n}_2 = \hat{n}_3$ $(\phi_{j_1} = 0,\ \eta_{j_1} > 0)$. Clearly, the dominant features are the jets, and while for $\mathcal{P}_J$, these lie in the center; for $\mathcal{P}_{\text{MET}}$ they lie at the ϕ-boundary. Thus, the effect of padding is much more pronounced in $\mathcal{P}_{\text{MET}}$. In analogy, it becomes crucial when the Higgs boson decays in a hadronic channel (say $h^0 \to b\bar{b}$ or even $h^0 \to \tau^+\tau^-$), where we would desire the jets arising from Higgs – be it normal or large-radius, to be at the center of the image. Combining the instances of preprocessing and resolutions, there are four low-level feature spaces, namely: $\mathcal{P}^{LR}_{\text{MET}}$, $\mathcal{P}^{HR}_{\text{MET}}$, $\mathcal{P}^{LR}_{J}$ and $\mathcal{P}^{HR}_{J}$. The superscripts LR and HR denote the low and high-resolutions. We notice that all the high-level variables except ϕ_{MET}, are invariant under the two preprocessing operations, although, for our purpose, we extract them prior to their application. This follows from the usual physical intuition that absolute positions in the lab-frame are of no particular importance, and the useful information comes from the relative position of the different final-states.

We regularize the high-level features by mapping the distribution of each variable to their z-scores. Calculating the mean $\bar{x}^j$, and the standard deviation σ^j for each feature of the whole dataset (training and validation data of both classes together), we perform the following operation on each variable of all events,

$$z_i^j = \frac{x_i^j - \bar{x}^j}{\sigma^j} \quad . \tag{5}$$

The superscript j denotes the feature index, and the subscript i denotes the per-event index. It is particularly useful since the features have very different ranges (for instance, m_{jj} and $|\Delta\eta_{jj}|$), and the operation minimizes this disparity. Furthermore, the features of z^j are now dimensionless. A caveat here is that the values of mean and standard deviations used are calculated from a balanced dataset. In experimental data, the presence of both classes, if at all, there is a positive signal, is never balanced. We justify our choice by their class independence, by virtue of which the relative differences in the shape of the signal and background distributions are conserved, and the same set of values can be used when applying to unknown data with no labels.

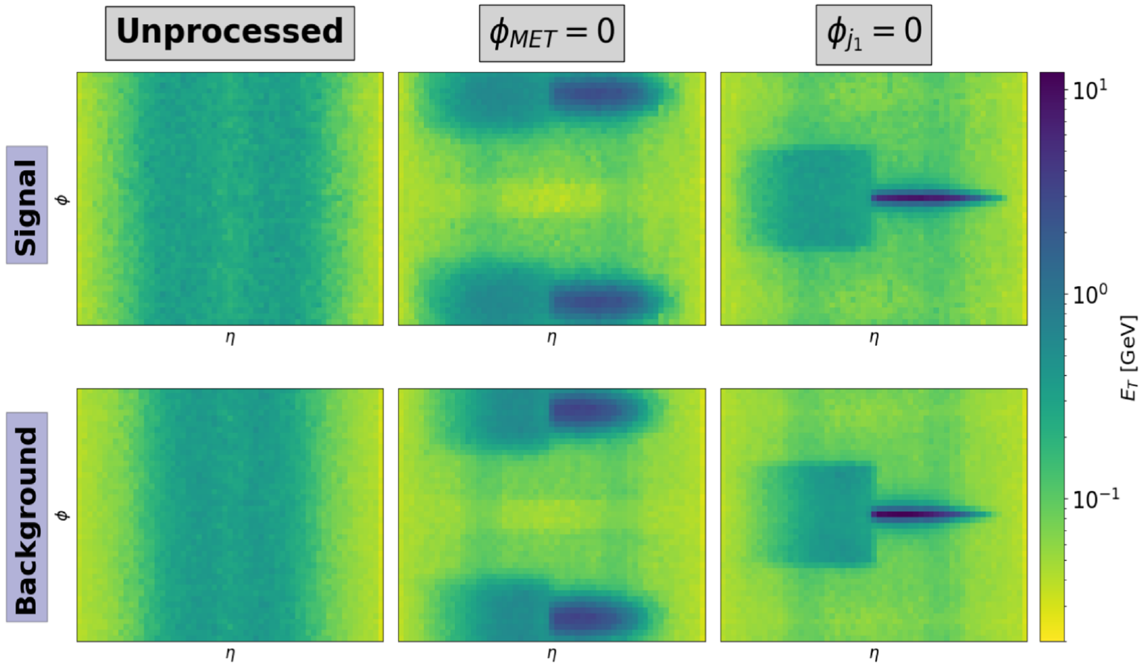


Fig. 9 Average of 25,000 low-resolution tower-images of (left) unprocessed, (center) processed image with $\phi_{\text{MET}} = 0$ and (right) $\phi_J = 0$ for (top panel) signal and (bottom panel) background classes. The images are binned in the full range of the tower: $\eta \in [-5, 5]$ and $\phi \in [-\pi, \pi)$. We can see that as we go from left to right, there is a discernible improvement in regularization of the features. There are no distinctly localized hard regions for the unprocessed case, while there are some for the $\phi_{\text{MET}} = 0$ instance, which becomes harder for $\phi_{j_1} = 0$ case with the hardest region around the leading jet

5 Neural network architecture and performance

In the previous sections, we have defined seven feature spaces, which are broadly grouped into high-level classes comprising of $\mathcal{K}$ (kinematic), $\mathcal{R}$ (QCD-radiative) and $\mathcal{H}$ (a combination of the two previous spaces); while low-level spaces are: $\mathcal{P}^{LR}_{\text{MET}}$, $\mathcal{P}^{HR}_{\text{MET}}$, $\mathcal{P}^{LR}_{J}$ and $\mathcal{P}^{HR}_{J}$. With these as inputs, we train neural-networks for classification. The generic architecture chosen for the high-level feature spaces are dense Artificial Neural Networks (ANNs) while for low-level ones are Convolutional Neural Networks. Hence, we name the 7 networks as: $\mathcal{K}$-ANN, $\mathcal{R}$-ANN, $\mathcal{H}$-ANN, $\mathcal{P}^{LR}_{\text{MET}}$-CNN, $\mathcal{P}^{HR}_{\text{MET}}$-CNN, $\mathcal{P}^{LR}_{J}$-CNN and $\mathcal{P}^{HR}_{J}$-CNN. All networks were executed in Keras(v2.2.4) [121] with TensorFlow(v1.14.1) [122] backend.

5.1 Choice of hyperparameters

The CNN is composed of three modules with each module formed by two convolutional layers followed by an average-pooling layer. Each convolutional layer consists of sixty-four filters with a size 4 × 4, with a single stride in each dimension. We pad all inputs to maintain the size of the outputs after each convolution. The pool-size is set to be 2 × 2 for all three modules with 2 × 2 stride size. The third module's output is flattened and fed into a dense network of three layers having three hundred nodes each, which we pass into the final layer with the two nodes and softmax activation. The convolutional layers and the dense layers before the final layer have ReLu activations. In total, the CNNs for the high-resolution (low-resolution) images have approximately 3.7 (1.2) million trainable parameters. The information bottleneck principle [123] inspires the ANN architectures. It has close connections to coarse-graining of the renormalization-group evolution and was, in fact, priorly pointed out in reference [124]. We choose the number of nodes in the first layer to be equal to the number of input-nodes, which is then successively reduced after two layers of the same dimension.[4] These reductions in successive nodes are chosen to be five for the $\mathcal{R}$-ANN and $\mathcal{H}$-ANN, while for $\mathcal{K}$-ANN, we consider four due to the low-dimensionality of the input. We stop this process when there is no further reduction possible, or after four such reductions. We checked two activation functions: sigmoid and ReLu for the ANNs. We found that sigmoid activation gave the best validation accuracy for $\mathcal{R}$-ANN and $\mathcal{H}$-ANN, while it decreased over ReLu activations for $\mathcal{K}$-ANN. In total, the $\mathcal{K}$-ANN, $\mathcal{R}$-ANN, and the $\mathcal{H}$-ANN have 210, 991, and 2790 trainable parameters, respectively. Since

[4] This provides stability of the representations learned at each dimension.

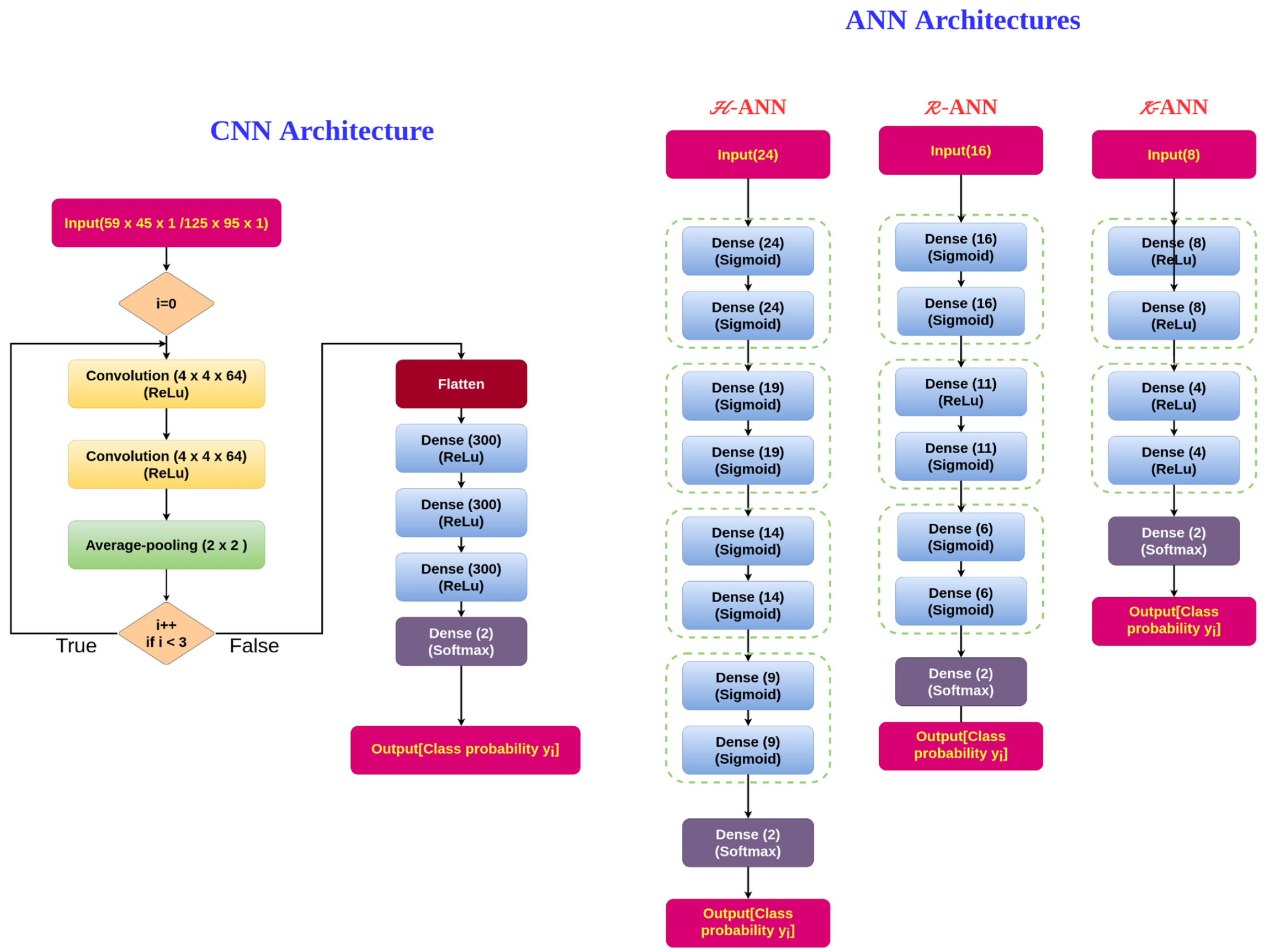


Fig. 10 Simplified architecture of (left) CNNs and (right) ANNs

this is a first exploratory study, we do not optimize the hyper-parameters and use the values specified here for extracting the results. Simplified architecture flowcharts for each of the different networks are given in Fig. 10.

We chose categorical-cross entropy as the loss function. The cross-entropy between two probability distributions y_0 and y_t is defined as,

$$L = -\sum_{\mathbf{x}\in\mathcal{X}} y_t(\mathbf{x})\ \ln(y_0(\mathbf{x})) \quad , \tag{6}$$

where the distributions are functions of the feature-vector $\mathbf{x}$. It is a measure of how well a modeled distribution y_0, corresponding to the network-output, resembles the true distribution of y_t, the true values provided during training. For a fixed true-distribution y_t, minimizing the cross-entropy essentially minimizes the KL-divergence [125],

$$D_{KL}(y_t||y_0) = \sum_{\mathbf{x}\in\mathcal{X}} y_t(\mathbf{x})\ \ln(y_t(\mathbf{x})) - \sum_{\mathbf{x}\in\mathcal{X}} y_t(\mathbf{x})\ \ln(y_0(\mathbf{x})), \tag{7}$$

which is a measure of the similarity between two distributions, and becomes zero iff they are identical. We used Nadam [126] optimizer with a learning rate of 0.001 to minimize the loss function for all neural-networks. The optimizer's adaptive nature: smaller updates for frequently occurring features while larger updates for rare features, helps in better convergence for the sparse image-data that we have, with the added benefits of Nesterov accelerated gradient descent [127]. Moreover, the learning-rate is no longer a hyperparameter. For the CNNs, training does not require more than ten epochs to reach optimal validation accuracy. Nevertheless, we train them five times from random initialization for twenty epochs. The ANNs are trained for more epochs since the relatively fewer parameters make the convergence slower. For the ANNs, ReLu activation networks are trained for two hundred epochs. In comparison, sigmoid activation networks are trained for one thousand epochs due to their relative difference in convergence compounded with fewer parameters.

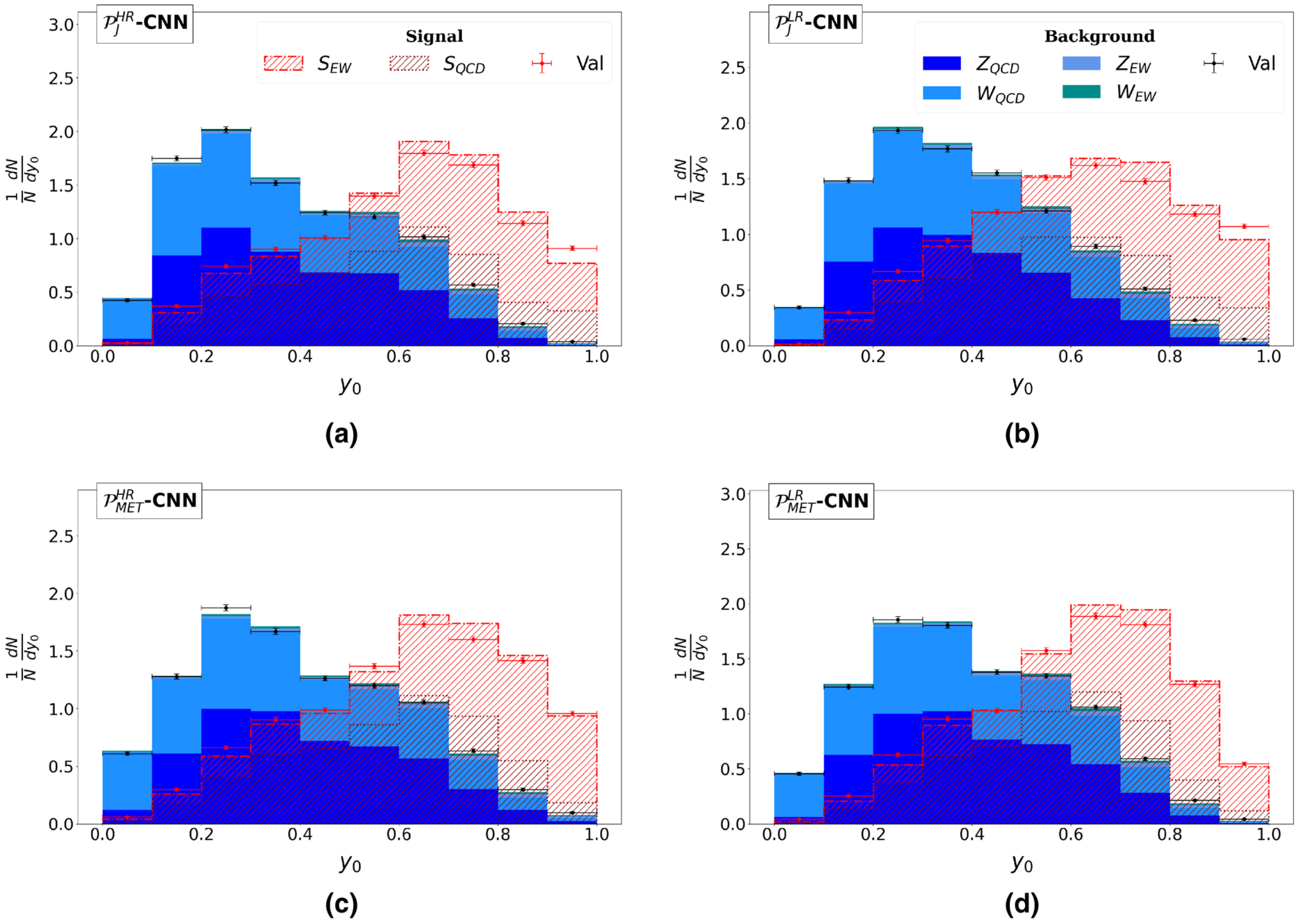


Fig. 11 Binned distribution of the network output for **a** $\mathcal{P}_J^{HR}$-CNN (top-left), **b** $\mathcal{P}_J^{LR}$-CNN (top-right), **c** $\mathcal{P}_{\text{MET}}^{HR}$-CNN (bottom-left) and **d** $\mathcal{P}_{\text{MET}}^{LR}$-CNN (bottom-right)

A batch-size of three hundred was chosen for training all networks. Each model, including all of its parameters, is stored after every epoch in the Keras-provided “hdf5” format during training. Out of these, we use the best performing model with the highest validation accuracy for further analysis.

5.2 Network outputs

We extract the network output y_0, which is the probability of the event being a signal, from the best performing model from each network class. The class-wise binned distribution of y_0, for training and validation datasets of the low-level and high-level feature spaces, are shown in Figs. 11 and 12, respectively. These also show the channel wise contribution to their parent class. The choice of binning is set to the same ones used in extracting the bounds on the invisible branching ratio of the Higgs in Sect. 6. It has been set such that the minimum number of entries of each class for the validation data in the edge bins have enough numbers to reduce the statistical fluctuations to less than 15%. Contributions of the S_{EW} and S_{QCD} components to the signal class follow a definite pattern. As expected, all networks find it difficult to distinguish the S_{QCD} signal from the QCD dominated background. Hence, S_{QCD} contributes more in the bins closer to zero, which is governed by the background class. S_{EW} shows the opposite behavior dominating near one. This same feature, although a little inconspicuous, is present for the background class’s EW subset as well. It may be pointed out that even for traditional analysis methods, there is significant contamination from S_{QCD}. A relevant machine-learning paradigm [36] where mixed samples are used in place of pure ones, could have an interesting application in reducing this S_{QCD} contamination of the signal for precision studies. Another notable feature prominent in the CNN outputs is the relative contribution of the Z_{QCD} and W_{QCD} channels to the background in the first bin, which is dominated by W_{QCD}. This can be apprehended from the fact that some of the leptons from $W^{\pm}$ decay, although not reconstructed, can still make calorimeter deposits on top of the QCD radiation to make itself visible to CNNs.

Receiver operating characteristic (ROC) curves between the signal acceptance ϵ_S, and the background rejection $1/\epsilon_B$;

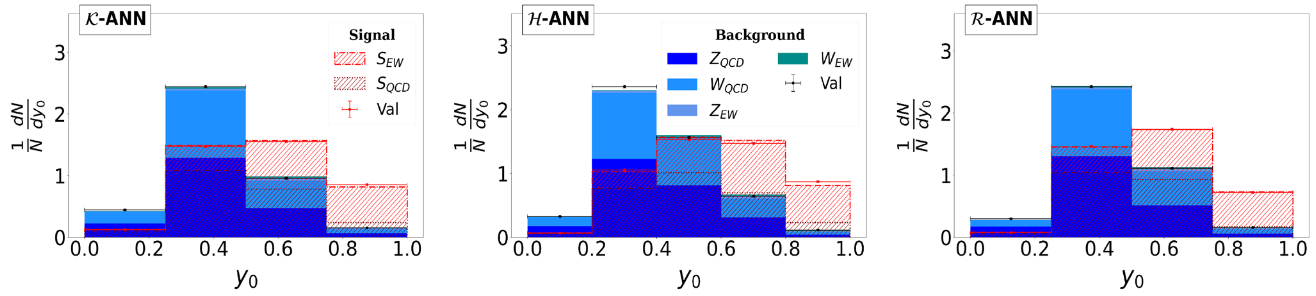


Fig. 12 Binned distribution of the network output for (left) $\mathcal{K}$-ANN, (center) $\mathcal{H}$-ANN, and (right) $\mathcal{R}$-ANN

and also the area under the curve (AUC) for all networks are shown in Fig. 13. The AUCs were calculated using y_0 and the true class labels y_t with the scikit-learn(v0.22) [128] package. It is interesting to see that the so-called QCD-radiative variables ($\mathcal{R}$) perform almost as good as the kinematic-variables ($\mathcal{K}$) with only less than a percent difference in the validation AUCs. It can be understood by recalling that the radiative variables' definition includes the radiation pattern of the event, including the radiation inside the jet in cumulative η bins. This, in principle, has similar information to $|\Delta\eta_{jj}|$, which is one of the kinematic-variables with high separation. We confirm this by observing the correlations (shown in Fig. 14) between the variables $H_T^{\eta_C=2.07}$ and $H_T^{\eta_C=1.8}$ with $|\Delta\eta_{jj}|$ and m_{jj}. They are relatively more correlated with $|\Delta\eta_{jj}|$ than with m_{jj}. The AUC for our combined variable $\mathcal{H}$-ANN shows that the $\mathcal{R}$ variables may contain some extra information on top of what is extracted from the kinematic variables. As emphasized earlier, we get less than 0.1 percent difference in the validation AUCs of the low and high-resolution networks. The difference in AUC between $\mathcal{P}_J$ and $\mathcal{P}_{\text{MET}}$, although small, is still significant. It can be understood by looking at Fig. 9: there is better feature regularization in $\mathcal{P}_J$ due to the choice of ϕ_0 than in $\mathcal{P}_{\text{MET}}$. CNNs, in general, are supposed to be robust to these kinds of differences owing to their properties of translational invariance [114]. In our case, the presence of fully-connected layers and the relatively small training sample hamper the generalization power of the CNNs. Application of global-pooling instead of using fully-connected layers and an increase in data size coupled with proper hyper-parameter optimization should reduce this difference in AUCs. These can be explored in future studies.

The class-wise linear correlation matrix between the network-outputs, along with the four high-level variables possessing the highest separations, are shown in Fig. 14. As expected, the outputs within the respective subset of networks are highly correlated. The outputs of the ANNs and the CNNs are also correlated significantly. A closer look reflects the addition of information in the high-level feature spaces: the correlations increase as we go from $\mathcal{R}/\mathcal{K}$-ANN to $\mathcal{H}$-ANN. In fact, if we extrapolate this argument in conjunction with the relative increase in AUC, we find that the CNNs have extracted the most information from the low-level data, which is not present in any of the high-level variables. A detailed description of the correlation of high-level variables and the ANN outputs are given in Appendix C.

6 Bounds on Higgs invisible branching ratio

In order to quantify our network performance in terms of expected improvements in the invisible Higgs search results at LHC, we obtain expected upper limits on the Higgs to invisible BRs from the distribution of the network output. We use CL$_s$ method [129,130] in the asymptotic approximation [131], to calculate the upper limit on the invisible BR at 95% CL. The method is briefly discussed as follows. In a binned Poisson counting experiment of expected signal s_i and background b_i (which are functions of nuisance parameters jointly denoted by $\boldsymbol{\theta}$) in a bin with observed number n_i of some observable, we can write the likelihood function as:

$$\mathcal{L}(\mu,\boldsymbol{\theta}) = \prod_{i=1}^{N_b} \frac{(\mu\ s_i(\boldsymbol{\theta}) + b_i(\boldsymbol{\theta}))^{n_i}}{n_i!}\ e^{-(\mu\ s_i(\boldsymbol{\theta})+b_i(\boldsymbol{\theta}))} \quad , \tag{8}$$

where N_b is the total number of bins. N_b and the bin-edges for the different variables are chosen as shown in their respective distribution plots (Figs. 5, 6, 11 and 12). The profile-likelihood ratio:

$$\lambda(\mu) = \frac{\mathcal{L}(\mu,\hat{\hat{\boldsymbol{\theta}}})}{\mathcal{L}(\hat{\mu},\hat{\boldsymbol{\theta}})} \quad , \tag{9}$$

where the arguments of the denominator maximizes $\mathcal{L}$, and $\hat{\hat{\boldsymbol{\theta}}}$ conditionally maximizes $\mathcal{L}$ for the particular μ, is used as a test-statistic in the form of log-likelihood,

$$t_\mu = -2\ln(\lambda(\mu)) \quad . \tag{10}$$

The distribution of the test statistic for different values of μ, is required to extract frequentist confidence intervals/limits. Since, we have fixed the total weight of the signal events with

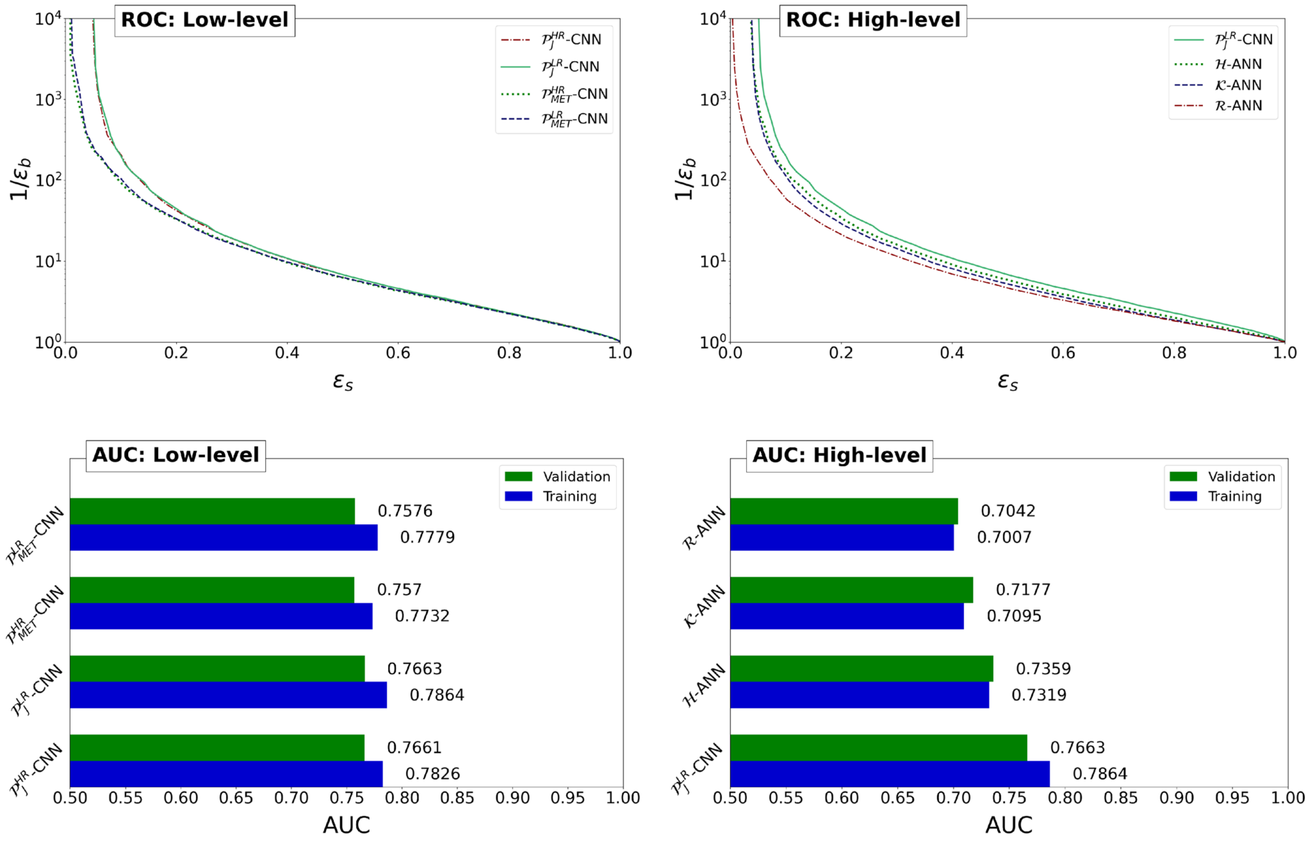
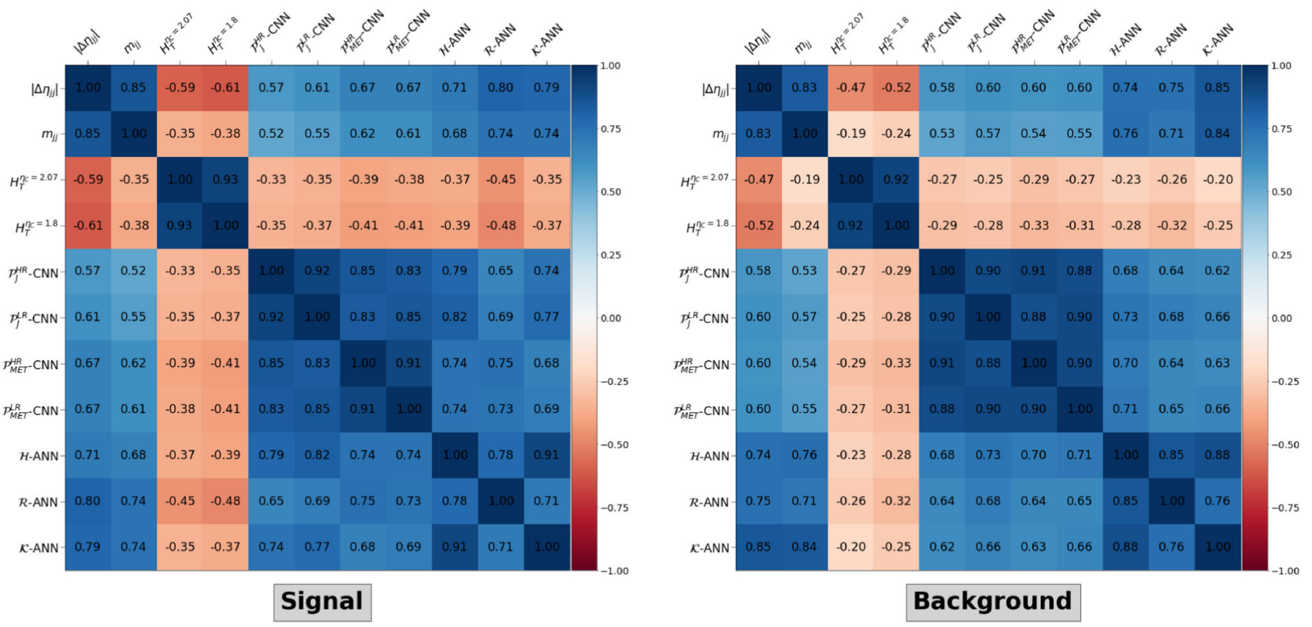


Fig. 13 The validation (top panel) ROC-curves and (bottom panel) training/validation AUC for (left) low-level and (right) high-level feature spaces. In order to compare the feature spaces, the highest performing CNN is added to the plots on the right. The x-axis of the ROC-curve is the signal acceptance ϵ_S, while the y-axis is the inverse of background acceptance ϵ_B



	$\lvert\Delta\eta_{jj}\rvert$	m_{jj}	$H_T^{\eta_c=2.07}$	$H_T^{\eta_c=1.8}$	$\mathcal{P}_j^{HR}$-CNN	$\mathcal{P}_j^{LR}$-CNN	$\mathcal{P}_{MET}^{HR}$-CNN	$\mathcal{P}_{MET}^{LR}$-CNN	$\mathcal{H}$-ANN	$\mathcal{R}$-ANN	$\mathcal{K}$-ANN
$\lvert\Delta\eta_{jj}\rvert$	1.00	0.85	-0.59	-0.61	0.57	0.61	0.67	0.67	0.71	0.80	0.79
m_{jj}	0.85	1.00	-0.35	-0.38	0.52	0.55	0.62	0.61	0.68	0.74	0.74
$H_T^{\eta_c=2.07}$	-0.59	-0.35	1.00	0.93	-0.33	-0.35	-0.39	-0.38	-0.37	-0.45	-0.35
$H_T^{\eta_c=1.8}$	-0.61	-0.38	0.93	1.00	-0.35	-0.37	-0.41	-0.41	-0.39	-0.48	-0.37
$\mathcal{P}_j^{HR}$-CNN	0.57	0.52	-0.33	-0.35	1.00	0.92	0.85	0.83	0.79	0.65	0.74
$\mathcal{P}_j^{LR}$-CNN	0.61	0.55	-0.35	-0.37	0.92	1.00	0.83	0.85	0.82	0.69	0.77
$\mathcal{P}_{MET}^{HR}$-CNN	0.67	0.62	-0.39	-0.41	0.85	0.83	1.00	0.91	0.74	0.75	0.68
$\mathcal{P}_{MET}^{LR}$-CNN	0.67	0.61	-0.38	-0.41	0.83	0.85	0.91	1.00	0.74	0.73	0.69
$\mathcal{H}$-ANN	0.71	0.68	-0.37	-0.39	0.79	0.82	0.74	0.74	1.00	0.78	0.91
$\mathcal{R}$-ANN	0.80	0.74	-0.45	-0.48	0.65	0.69	0.75	0.73	0.78	1.00	0.71
$\mathcal{K}$-ANN	0.79	0.74	-0.35	-0.37	0.74	0.77	0.68	0.69	0.91	0.71	1.00

	$\lvert\Delta\eta_{jj}\rvert$	m_{jj}	$H_T^{\eta_c=2.07}$	$H_T^{\eta_c=1.8}$	$\mathcal{P}_j^{HR}$-CNN	$\mathcal{P}_j^{LR}$-CNN	$\mathcal{P}_{MET}^{HR}$-CNN	$\mathcal{P}_{MET}^{LR}$-CNN	$\mathcal{H}$-ANN	$\mathcal{R}$-ANN	$\mathcal{K}$-ANN
$\lvert\Delta\eta_{jj}\rvert$	1.00	0.83	-0.47	-0.52	0.58	0.60	0.60	0.60	0.74	0.75	0.85
m_{jj}	0.83	1.00	-0.19	-0.24	0.53	0.57	0.54	0.55	0.76	0.71	0.84
$H_T^{\eta_c=2.07}$	-0.47	-0.19	1.00	0.92	-0.27	-0.25	-0.29	-0.27	-0.23	-0.26	-0.20
$H_T^{\eta_c=1.8}$	-0.52	-0.24	0.92	1.00	-0.29	-0.28	-0.33	-0.31	-0.28	-0.32	-0.25
$\mathcal{P}_j^{HR}$-CNN	0.58	0.53	-0.27	-0.29	1.00	0.90	0.91	0.88	0.68	0.64	0.62
$\mathcal{P}_j^{LR}$-CNN	0.60	0.57	-0.25	-0.28	0.90	1.00	0.88	0.90	0.73	0.68	0.66
$\mathcal{P}_{MET}^{HR}$-CNN	0.60	0.54	-0.29	-0.33	0.91	0.88	1.00	0.90	0.70	0.64	0.63
$\mathcal{P}_{MET}^{LR}$-CNN	0.60	0.55	-0.27	-0.31	0.88	0.90	0.90	1.00	0.71	0.65	0.66
$\mathcal{H}$-ANN	0.74	0.76	-0.23	-0.28	0.68	0.73	0.70	0.71	1.00	0.85	0.88
$\mathcal{R}$-ANN	0.75	0.71	-0.26	-0.32	0.64	0.68	0.64	0.65	0.85	1.00	0.76
$\mathcal{K}$-ANN	0.85	0.84	-0.20	-0.25	0.62	0.66	0.63	0.66	0.88	0.76	1.00

Fig. 14 Pearson's correlation coefficients amongst the first four high-level variables with highest separation and the network-outputs for (left) signal and (right) background. These have been calculated using the validation dataset

respect to the background to correspond to the ones expected with the total expected production cross-section from SM for each channel(S_{EW} and S_{QCD}), μ corresponds to the invisible branching ratio of the Higgs. In the asymptotic method, for one parameter of interest, approximate analytical expressions for the distribution are derived using a result from Wald [132], in the form of a non-central Chi-square distribution. Monte-Carlo simulations required to extract the unknown parameters are by-passed by choosing the best representative data called the Asimov data, by the authors of reference [131]; which is defined as the data when used to estimate the parameters, produces their true values.

We used HistFactory [133] to create the statistical model, and the RooStats [134] package to obtain the expected limits. This provides us with greater ease of handling systematic uncertainties. As stated before, we also redo the shape-based analysis of reference [83] with our dataset only considering a few simpler systematics, to consistently gauge the increased sensitivity of the deep-learning approach. We incorporate three overall-systematics: uncertainty of the total cross-section, statistical uncertainty of Monte Carlo simulated events, and approximate luminosity uncertainties. We do not take into account the possible change in the shape of the distributions due to Monte Carlo simulation effects. The per-bin statistical error is taken into consideration by activating each sample's statistical-error while creating the statistical model in HistFactory. This is essentially a shape-systematics that considers the bin-wise change in shape due to the statistical uncertainties. Its inclusion increases the median expected upper-limit by around three percent in the reproduced shape-analysis. The number of events for the analysis with the higher MET cut is set to the expected number at 36 fb^{-1} for all background channels. This result is also scaled for the other luminosities. For the ones with the lower MET cut, we use the validation data scaled by appropriate weights for the respective luminosities.

The median expected upper limit on the invisible branching ratio of SM Higgs at 95% CL along with the one and two sigma error bands are shown in Fig. 15 for integrated luminosities of 36 fb^{-1} and 140 fb^{-1}. A short description of the datasets used, and the corresponding median-expected upper limits with 95 % CL is tabulated in Table 1. This also contains the projected limits for 300 fb^{-1}, the integrated luminosity expected at the end of LHC Run III. We emphasize that even though we scale to 300 fb^{-1} luminosity, we use the same dataset, and hence, the statistical uncertainties are not reduced. Consequently, our estimation for 300 fb^{-1} is a conservative one. First and foremost, one can notice that the reproduced result of the shape-analysis of reference [83] for an integrated luminosity of 36 fb^{-1} is quite consistent, and the difference can be accounted to the excluded background channels and experimental systematics. We repeat this analysis with the weaker selection criteria and see a modest improvement in the median-expected upper-limit. We also perform similar analyses with $|\Delta\eta_{jj}|$ distributions, and get an improvement of 2.9 % for MET $>$ 200 GeV, and 2.6 % for MET $>$ 250 GeV cuts. The worst (best) performing neural-network $\mathcal{R}$-ANN ($\mathcal{P}_J^{HR}$-CNN) has an improvement of 8.8% (14.6%) from the repeated experimental analysis. This, although, is with different cuts, and for the same cut in MET, we have an improvement of 5.3% (12.1%) for $\mathcal{R}$-ANN ($\mathcal{P}_J^{HR}$-CNN). For an integrated luminosity of 140 fb^{-1}, we get an improvement of 2.2 % and 7.3 % for $\mathcal{R}$-ANN and $\mathcal{P}_J^{HR}$-CNN, respectively. The reduced difference for higher luminosities is, of course, expected since the significance does not scale linearly with an increase in data size. An expected median upper-limit of about 3.5% can be achieved with 300 fb^{-1} of data using the highest performing network, $\mathcal{P}_J^{HR}$-CNN.

The results of the different feature spaces follow the expected trend. For this discussion, we quote the numbers for an integrated luminosity of 36 fb^{-1}. Comparing the performance of high-level feature spaces, we see that $\mathcal{R}$ performs the worst while the combined space $\mathcal{H}$ puts the most stringent bounds. The difference is minimal (0.7 %) with $\mathcal{K}$-ANN, and appreciable (4.4%) with $\mathcal{R}$-ANN. Amongst the image-networks, the difference between the low and high-resolution networks is less than a percent (0.8 % for $\mathcal{P}_J$, and 0.6% for $\mathcal{P}_{\text{MET}}$). Differences in performances of the different preprocessing instances are reflected in this analysis: $\mathcal{P}_J$ puts nominally stricter bounds on the branching ratio (1.4 % for LR, and 1.6 % for HR).

Up to now, we demonstrated the capability of our CNN based low-level networks and also ANN-based networks considering particle level data, including detector effects as well as underlying events during our simulations as discussed in Sect. 2. However, we neglected the effect of simultaneous occurrences of multiple proton-proton interactions (pileup) in our analysis. The amount of pileup is relatively moderate in low luminosity data, but increasingly significant once we move towards high luminosity. We believe that its presence would not alter our primary results substantially from the calorimeter image data. CNN architectures look into the global features of an input image. Calorimeter deposits due to pileup are expected to be similar for different classes since they are independent of the hard scattering processes. The same can be identified as redundant information, as a consequence of the optimization algorithm effectively searching for dissimilarities between the two classes. Optimal pdfs acquired by CNNs remain very similar, whether it is with or without pileup. This issue was analyzed before, where it was shown that unlike high-level methods, deep-learning from calorimeter deposits shows robustness to pileup effects in the classification of jet-image [28]. Although, in these studies, the jets have large transverse boosts and mostly reside in the central regions where its effect is reduced. However, various

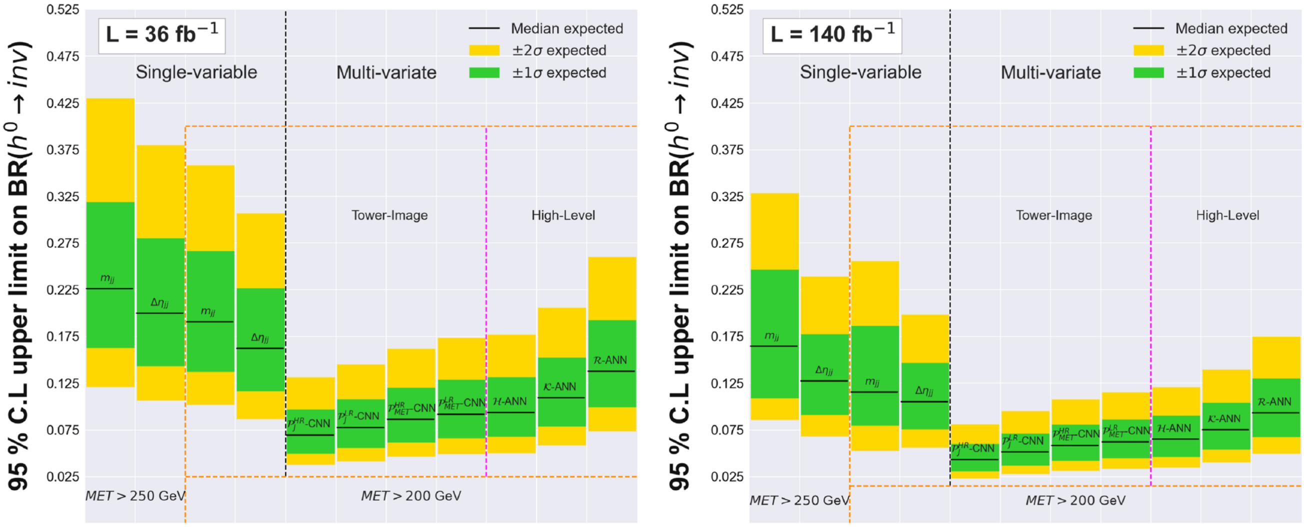


Fig. 15 Expected 95% C.L median upper limit on the invisible branching ratio of SM Higgs with one and two sigma sidebands for (left) 36 fb^{-1} and (right) 140 fb^{-1} integrated luminosities

Table 1 Short description of the different analyses shown in Fig. 15 and the expected median upper-limit on BR($h^0 \rightarrow$ inv) at 95% CL for each integrated luminosities which also include projections for L = 300 fb^{-1}

Sl.No	Name	Description	Expected median upper-limit on BR($h^0 \rightarrow$ inv)		
			L = 36 fb^{-1}	L = 140 fb^{-1}	L = 300 fb^{-1}
1.	m_{jj}(MET > 250 GeV)	Reproduced shape analysis of reference [83]	$0.226^{+0.093}_{-0.063}$	$0.165^{+0.082}_{-0.056}$	$0.130^{+0.089}_{-0.027}$
2.	$\lvert\Delta\eta_{jj}\rvert$(MET > 250 GeV)	$\lvert\Delta\eta_{jj}\rvert$ analysis with shape-cuts of reference [83]	$0.200^{+0.080}_{-0.056}$	$0.128^{+0.050}_{-0.036}$	$0.106^{+0.041}_{-0.025}$
3.	m_{jj}(MET > 200 GeV)	m_{jj} shape analysis with weaker cut	$0.191^{+0.075}_{-0.053}$	$0.116^{+0.071}_{-0.036}$	$0.101^{+0.037}_{-0.045}$
4.	$\lvert\Delta\eta_{jj}\rvert$(MET > 200 GeV)	$\lvert\Delta\eta_{jj}\rvert$ analysis with weaker cut	$0.162^{+0.065}_{-0.045}$	$0.105^{+0.042}_{-0.029}$	$0.087^{+0.034}_{-0.025}$
5.	$\mathcal{P}_J^{LR}$-CNN	Low-Resolution, $\phi_0 = \phi_{j_1}$	$0.078^{+0.030}_{-0.022}$	$0.051^{+0.020}_{-0.014}$	$0.045^{+0.017}_{-0.013}$
6.	$\mathcal{P}_J^{HR}$-CNN	High-Resolution, $\phi_0 = \phi_{j_1}$	$0.070^{+0.027}_{-0.020}$	$0.043^{+0.017}_{-0.012}$	$0.035^{+0.013}_{-0.010}$
7.	$\mathcal{P}_{\text{MET}}^{LR}$-CNN	Low-Resolution, $\phi_0 = \phi_{\text{MET}}$	$0.092^{+0.037}_{-0.025}$	$0.062^{+0.024}_{-0.017}$	$0.053^{+0.023}_{-0.014}$
8.	$\mathcal{P}_{\text{MET}}^{HR}$-CNN	High-Resolution, $\phi_0 = \phi_{\text{MET}}$	$0.086^{+0.035}_{-0.024}$	$0.058^{+0.023}_{-0.016}$	$0.051^{+0.020}_{-0.014}$
9.	$\mathcal{K}$-ANN	8 kinematic-variables	$0.101^{+0.052}_{-0.022}$	$0.075^{+0.029}_{-0.021}$	$0.063^{+0.027}_{-0.017}$
10.	$\mathcal{R}$-ANN	16 radiative $H_T^{\eta C}$ variables	$0.138^{+0.055}_{-0.039}$	$0.094^{+0.036}_{-0.027}$	$0.079^{+0.032}_{-0.022}$
11.	$\mathcal{H}$-ANN	Combination of $\mathcal{K}$ and $\mathcal{R}$ variables	$0.094^{+0.038}_{-0.026}$	$0.065^{+0.026}_{-0.018}$	$0.057^{+0.022}_{-0.015}$

other studies [47,48] have also shown that deep-learning on the full calorimeter information is less prone to pileup effects. These existing results further elucidate our presumption that CNNs would be less affected by higher pileup expected at future runs of LHC. In contrast, the other analyses, including the ANNs trained on high-level feature spaces, can be relatively more affected.

To present our arguments in perspective, we combined each event (tower-image) with an additional N randomly chosen minimum bias event with CMS switch through Pythia8 and Delphes without any pileup subtraction. At the same time, N follows a Poisson distribution with $< N >= 20, 50, 50$ for integrated luminosity 36, 140 and 300 fb^{-1}, respectively. Merged tower-image with pileup is then trained and tested for our high-resolution CNN scenario ($\mathcal{P}_J^{HR}$-CNN, which can be noted from Sl.No (6) in Table 1). We found a very mild depreciation over our estimated median upper-limit at 0.076, 0.059, and 0.045, which all lie within the 1σ error band in the branching ratio constraints. Note that no effort was made to mitigate the effects of the pileup during these estimates, which will not be the case in experimental analyses. In fact, there are extensive studies [135,136] of using powerful machine-learning algorithms specially designed to reduce pileup contamination of events. A new interpretation of collider events in terms of optimal transport [137,138] have also provided promising new techniques for pileup mitigation on top of reinterpretation of existing ones [139,140]. These developments offer further optimism for better mitigation of pileup effects in the future.

To test the robustness of our proposal, we also consider the effect of an important experimental systematic uncertainty. One of the significant experimental systematic uncertainties affecting the result of this analysis can be the uncertainty on the jet energy scale. Therefore, we estimate the effect of uncertainties on the jet energy scale for our main results with calorimeter input data in CNN architecture. We vary the pixel-wise input values (which has already gone through the smearing in Delphes) by 10% in upward and downward directions, [5] and record the variation in the shape of the network output without considering any pileup. This is added as a coherent shape systematics, and we obtain an increased expected median upper-limit of $0.071^{+0.028}_{-0.019}$ for $\mathcal{P}_J^{HR}$-CNN at 140 fb^{-1} integrated luminosity, which is still better by a factor of almost two when compared to the latest result from ATLAS [95].

7 Summary and conclusion

The HEP experimental community is one of the frontrunners in utilizing machine learning algorithms for the last several decades in tagging and characterizing different objects and analyzing the massive data samples with the help of neural-network or boosted decision trees. However, recent developments in deep learning approaches have shown immense prospects in a variety of other applications.

The Large Hadron Collider, after its breakthrough discovery of an SM like Higgs boson, keeps accumulating an enormous amount of data, pinpointing its different properties and also constraining diverse BSM scenarios at the TeV scale. While such high energy data are opening up scope for new analysis techniques filling possible gaps in previous investigations, it is prudent to review the effectiveness of some of the effective machine learning tools.

While proposed as an alternative channel for Higgs search, the vector boson fusion (VBF) mechanism has shown tremendous possibility not only in extracting properties of the Higgs boson but also in many other BSM searches. As a whole, this mechanism reckons upon some of the fundamental features of event shape, vastly used to control the backgrounds.

We choose VBF production of the Higgs boson decaying to invisible particles as a case study for neural networks to learn the entire event topology without any reconstructed objects. We use the compelling capability of Convolutional Neural Networks (CNN) to examine the potential of deep-learning algorithms using low-level variables. Instead of identifying any particular objects, we utilize the entire calorimeter image to study the event topology, which aims to learn the difference in radiation patterns between the two forward jets of the VBF signal. We specifically develop preprocessing steps that preserve the Lorentz symmetry of the events and are essential to maximizing the statistical output of the data.

Apart from low-level variables as calorimeter images for CNN, we also consider two sets of high-level features. One such set is based on the kinematics of the VBF, whereas the other set of variables are designed to portray the radiation pattern H_T calculated in different η ranges of the calorimeter. For a comprehensive analysis, we constructed several neural network architectures and demonstrated the comparative performance of CNN and ANN using different feature spaces. All these networks achieved excellent separation between signal and background. However, we found that CNN based low-level $\mathcal{P}_J^{HR}$-CNN performs the best among all the networks, which is based on the high-resolution images, although the dependence on image resolution is relatively insignificant. We also note that deep-learning on the full calorimeter information is less prone to pileup effects as well. Without relying on any exclusive event reconstruction, this novel technique can provide the most stringent bounds on the invisible branching ratio of the SM-like Higgs boson, which can be expected to be constrained up to 4.3% (3.5%) using a dataset corresponding to an integrated luminosity of 140 fb^{-1} (300 fb^{-1}). These limits can severely constrain many BSM scenarios, especially in the context of (Higgs-portal) dark matter models. The techniques presented in this work can easily be extended to a more complex event topology.

Acknowledgements The work of AB, PK, and VSN is supported by the Physical Research Laboratory (PRL), Department of Space, Government of India. Computational work for this work was performed using the HPC resources (Vikram-100 HPC) and TDP project at PRL. Authors gratefully acknowledge WHEPP'19, where part of this work was initiated. We also thank Regional Centre for Accelerator-based Particle Physics (RECAPP), Harish-Chandra Research Institute, for offering forums for such workshops and stimulating discussions with S. Bhattacharya, R. K. Singh, S. Rai, and B. Mukhopadhyaya.

Data Availability Statement This manuscript has no associated data or the data will not be deposited. [Authors' comment: The data used in the work is very large owing to the nature of the work (Deep-learning), hence it was impractical to submit the data.]

[5] Reference [141] reports jet energy scale uncertainty for various observables, which lie well within 5%. However, since such uncertainties are significantly controlled in jets reconstructed with the particle-flow (PF) method, we take a relatively conservative measure for the pixel-wise uncertainty of the measured energies.

Funded by SCOAP^3.

Appendix A: Incorporating finite mass effect of top quark in gluon-fusion events

We generate the gluon-fusion production of the Higgs boson by using the Higgs Effective Field Theory (HEFT) model, where the interaction of the Higgs boson with gluons is approximated by an effective vertex calculated by taking the top-quark mass to infinity. This is a reasonable approximation only when all relevant scales in the physical process are less than 2 m_t. The distribution of p_T of the Higgs boson (equivalently MET with detector effects introduced via Delphes) has a significant portion of events in regions where the approximation is not valid. We remove this inconsistency by reweighting the MET distribution of the events obtained after Delphes. We extract weights (ratio of the full SM results to HEFT) and bins in p_T of the Higgs for the present final state topology from figure 30 on reference [105]. Each event is then assigned the corresponding weight of the bin of its MET. After reweighting the events, we apply the preselection-cuts and extract the cut efficiency using the weights.

Since we need unweighted events for the neural network training, the passed events are again unweighted. This is done in the following steps. We divide all events into sets with unique weights. This is nothing but grouping the events into the extracted bins in MET. We get mutually exclusive subsets of events $\mathcal{S}_i$, with i being the bin-index. The per-bin weights are divided by their maximum value. We get a weight $w_i \in (0, 1]$ for each $\mathcal{S}_i$. From each set $\mathcal{S}_i$, we randomly choose w_i proportion of events rounded to the closest integer. We show in Fig. 16, the distribution of some kinematic-variables of the three datasets: unweighted events generated with HEFT model, weighted events with finite-top mass effects, and unweighted events used in neural network training. The effect of rounding to the nearest integer is seen in the later bins in MET, where the statistics are weaker due to fewer events.

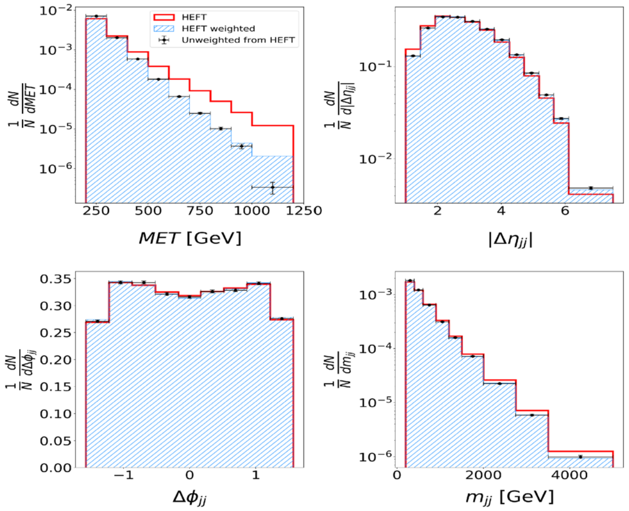


Fig. 16 Comparative distribution of kinematic variables for HEFT, weighted with finite-top mass effects and unweighted distributions for passed events used in deep-learning training and validation

Appendix B: Characteristics of high-level variables

In this section, we take a closer look at the high-level variables, especially the $\mathcal{R}$ variables defined in eq. 2. A key element in the extraction of variables belonging to the two spaces $\mathcal{K}$ and $\mathcal{R}$ is that the $\mathcal{K}$ variables are functions of four-momenta of reconstructed objects while the $\mathcal{R}$ variables are functions of four-momenta of tower-constituents (in our case from the Tower class of Delphes). The $\mathcal{R}$ variables do not take into account the tower-resolutions in the strict sense. This may point to a further reduction in the performance of ANNs compared to CNNs, where the tower-resolutions are better modeled.

We show the signal vs background distribution of all $\mathcal{R}$ variables in Fig. 17. The contribution of S_{EW} and S_{QCD} to the total signal is stacked. The separation, as defined in eq. 4, are shown for these variables for the total signal (also, S_{EW}) and background in Fig. 18. We can see that the trends in the distribution are in accordance with their respective values of separation. The shape of S_{QCD} and the background distributions are similar for all values of η_C, and the overall differences, if any, comes from the contribution of S_{EW}. The separation is minimal and remains constant for $\eta_C > 4$. This can be attributed to the fact that above these values, almost all of the calorimeter hits contribute to $H_T^{\eta_C}$. It increases continuously up to $\eta_C = 1.8$ and then decreases till $\eta_C = 1.0$. The increase is expected from the VBF topology, while the decrease can be attributed to the smallness of the region $[-\eta_C, \eta_C]$.

In Fig. 19, we show the remaining kinematic variables not shown in Fig. 6. As can be seen, there is not much discriminatory information in any of these variables: ϕ_{MET} is uniform for all channels since the beams are unpolarized, while $\Delta\phi^J_{\text{MET}}$ $(J \in \{j_1, j_2, j_1 + j_2\})$ has most contributions around ±π, due to the imposed separation of two jets $\Delta\phi_{jj}$ and momentum conservation in the recoil of quarks/gluons against heavy bosons ($W^\pm$, Z^0 and h^0).

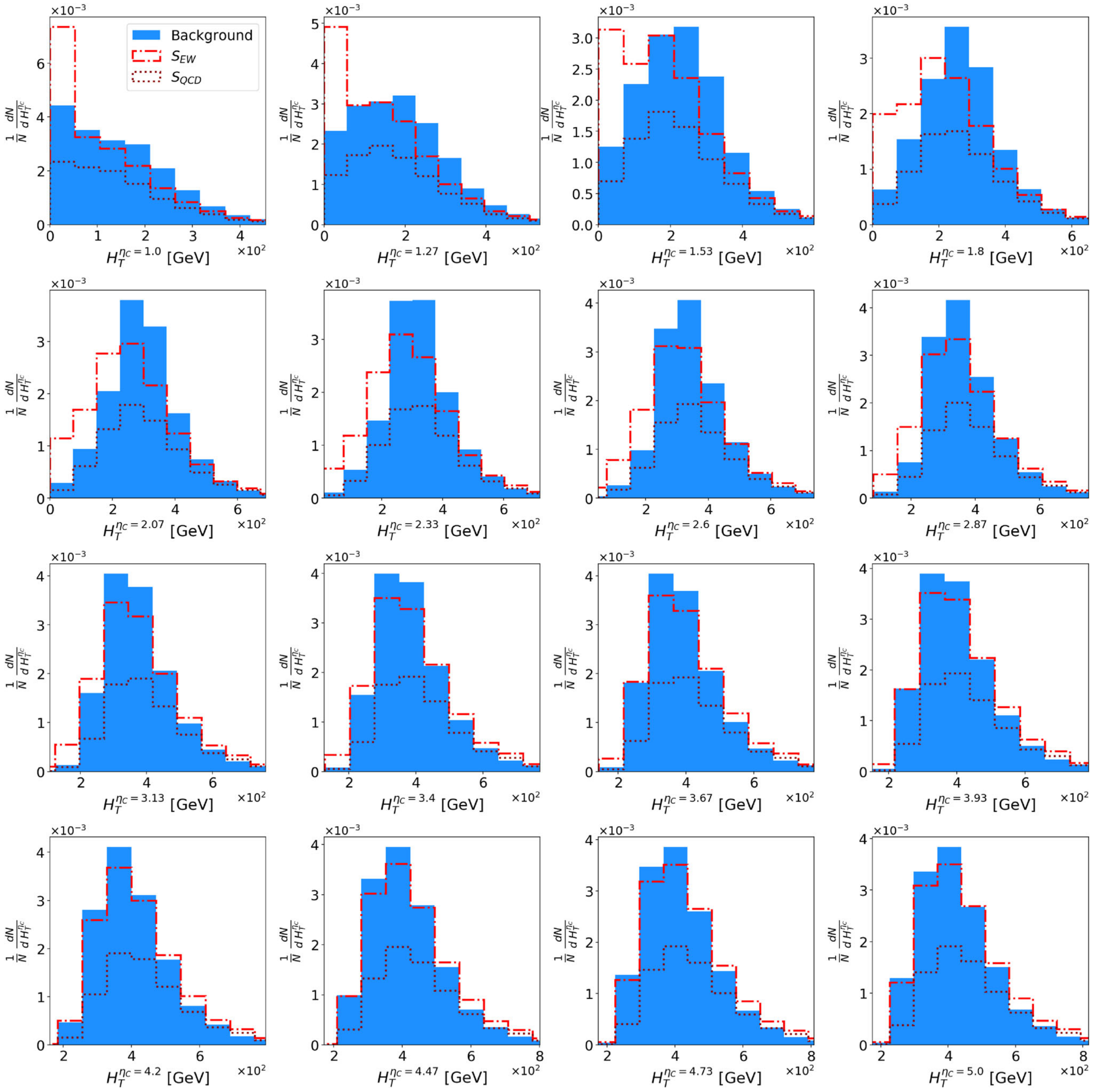


Fig. 17 Signal vs Background distribution for all $H_T^{\eta_C}$ variables. We can see that for higher values of η_C the signal and background are not that different and the difference grows as we approach the cut value of η cut

Appendix C: Correlation between high-level variables and network-oustputs

Salient features of the correlation of important variables with all neural network outputs have been given in the main text (Fig. 14). We examine the correlation of the ANNs with their inputs in this section. All correlations have been calculated using the inbuilt function in NumPy(v1.17.2)[142].

In Fig. 20 we show the correlations amongst the $\mathcal{K}$ variables including the $\mathcal{K}$-ANN network output for each class. As expected, the $\mathcal{K}$-ANN output is highly correlated with the two most discriminating variables $|\Delta\eta_{jj}|$ and m_{jj}. The next highest correlation with $\mathcal{K}$-ANN is found to be with MET for background and $|\Delta\phi_{jj}|$ for signal. Except for $|\Delta\phi_{jj}|$, all other ϕ variables are almost uncorrelated with $\mathcal{K}$-ANN for both classes. The uniformity of ϕ_{MET} results in its negligible correlation with all other variables. In the correlation

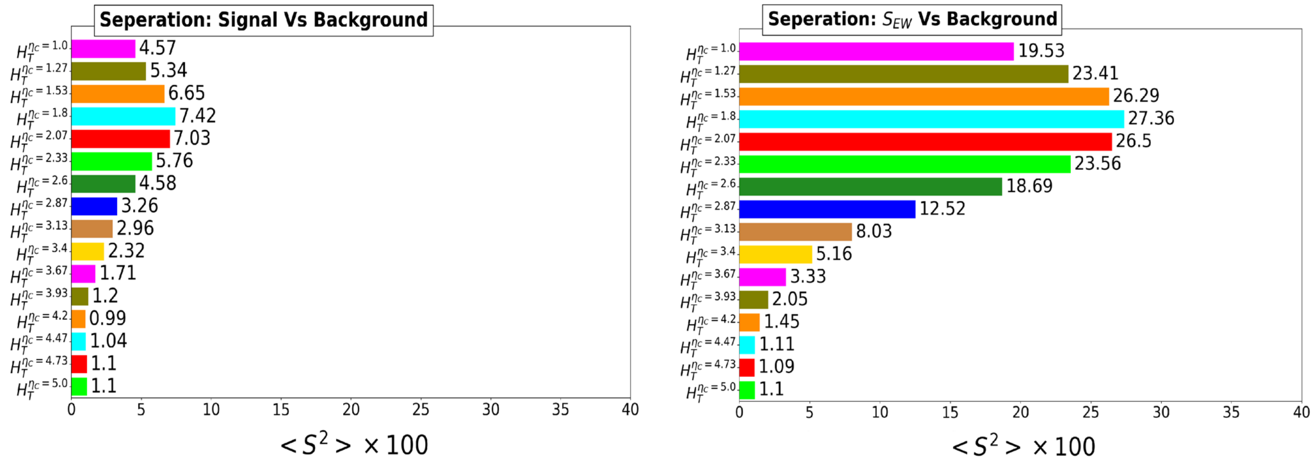


Fig. 18 Seperation of all $H_T^{\eta C}$ variables for (left) signal vs background and (right) S_{EW} vs background. These have been calculated with 25000 events for each of the three datasets with the same binning. We can see that the presence of S_{QCD} significantly reduces the discriminating power of $H_T^{\eta C}$ variables on the left

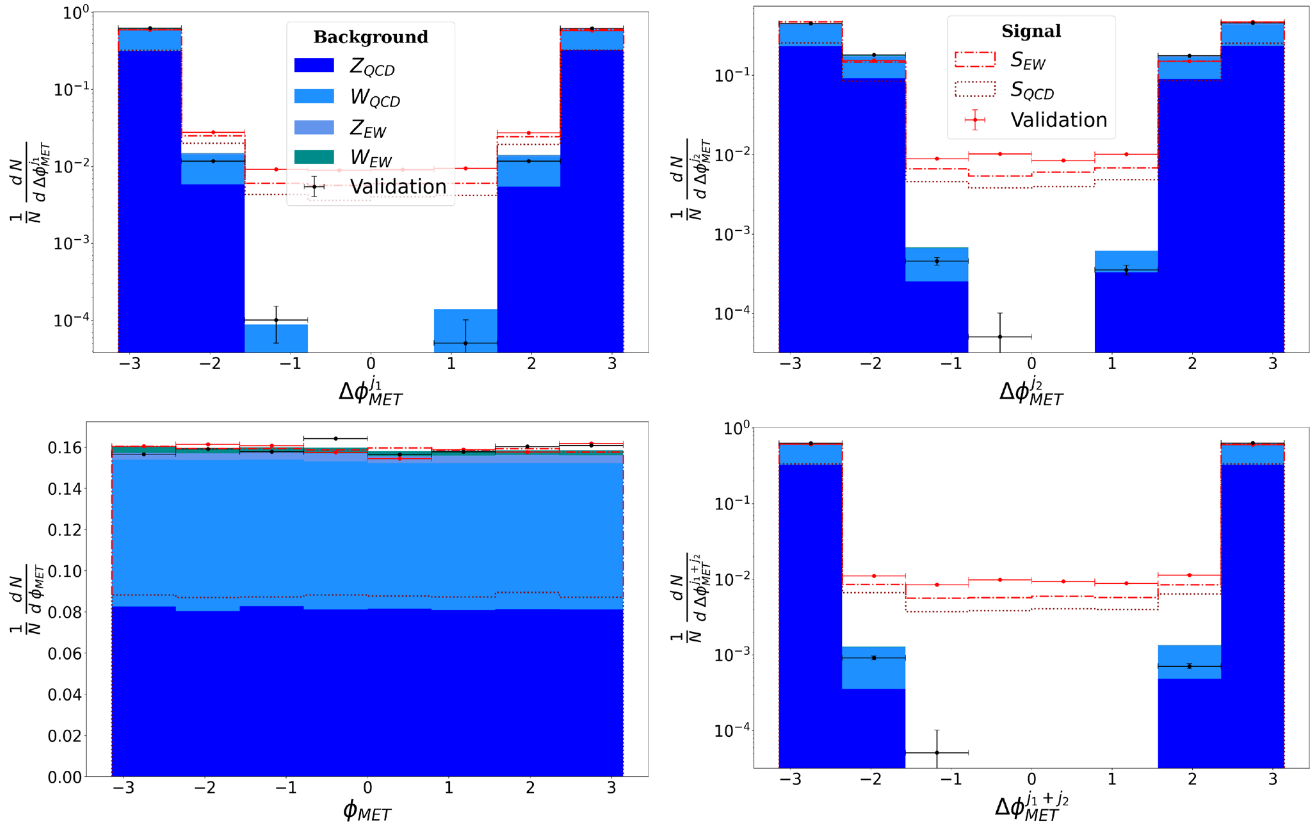


Fig. 19 Signal vs Background distribution of the high-level kinematic variables excluded in Fig. 6

among $\mathcal{K}$ variables, we can see two distinct sets of variables with comparatively moderate to high correlations formed amongst $\{|\Delta\eta_{jj}|, m_{jj}, \text{MET}\}$ and $\{\Delta\phi_{\text{MET}}^{j_1}, \Delta\phi_{\text{MET}}^{j_2}, \Delta\phi_{\text{MET}}^{j_1+j_2}\}$. In the first set, $|\Delta\eta_{jj}|$ and m_{jj} are almost completely correlated since, the angular opening between two four vectors $p_{j_1}^{\mu}$ and $p_{j_2}^{\mu}$, determine the invariant mass $m_{jj} = (p_{j_1}^{\mu} + p_{j_2}^{\mu})^2$. The MET shows a moderate correlation with both $|\Delta\eta_{jj}|$ and m_{jj} as— momentum conservation forces $|\mathbf{p}_{j_1} + \mathbf{p}_{j_2}|$ to be higher for higher MET. The correlation amongst the second subset can also be explained by transverse momentum conservation in the collision, with contamination from subsidiary QCD radiation and detector effects.

The class-wise correlations amongst the outputs of $\mathcal{R}$-ANN and $\mathcal{H}$-ANN along with six variables from $\mathcal{R}$ with high

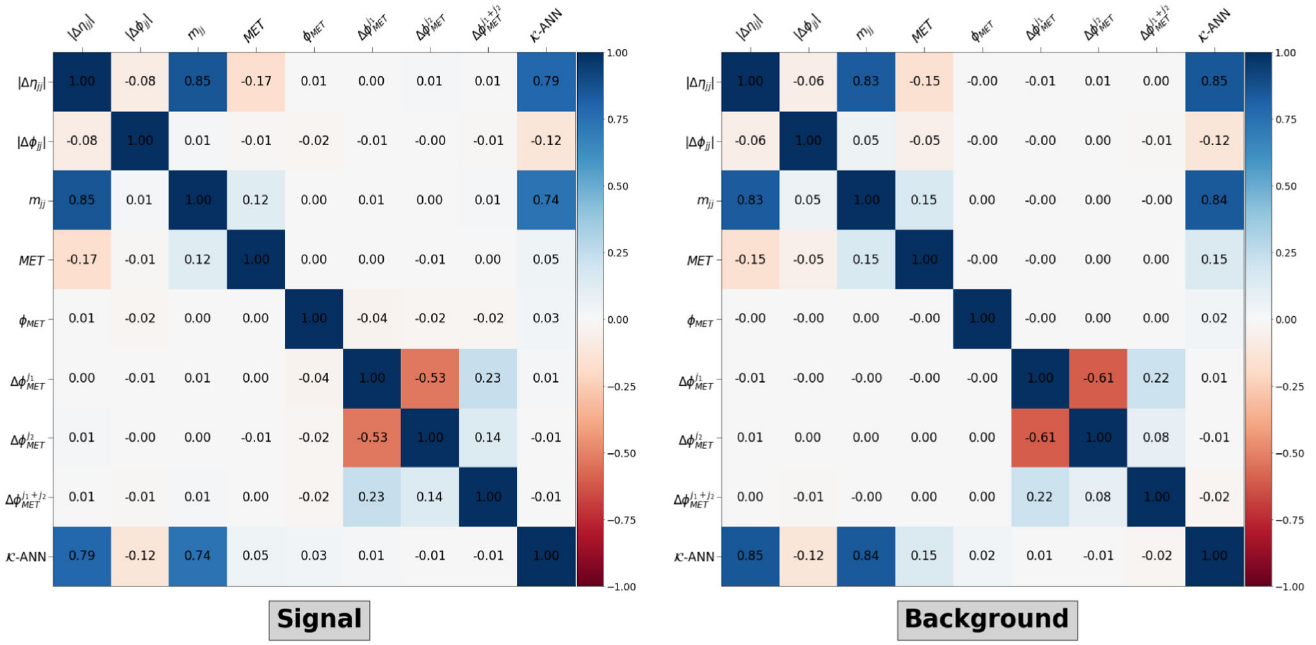


Fig. 20 Correlation between the high-level kinematic variables $\mathcal{K}$ and the network-output of $\mathcal{K}$-ANN for (left) signal and (right) background

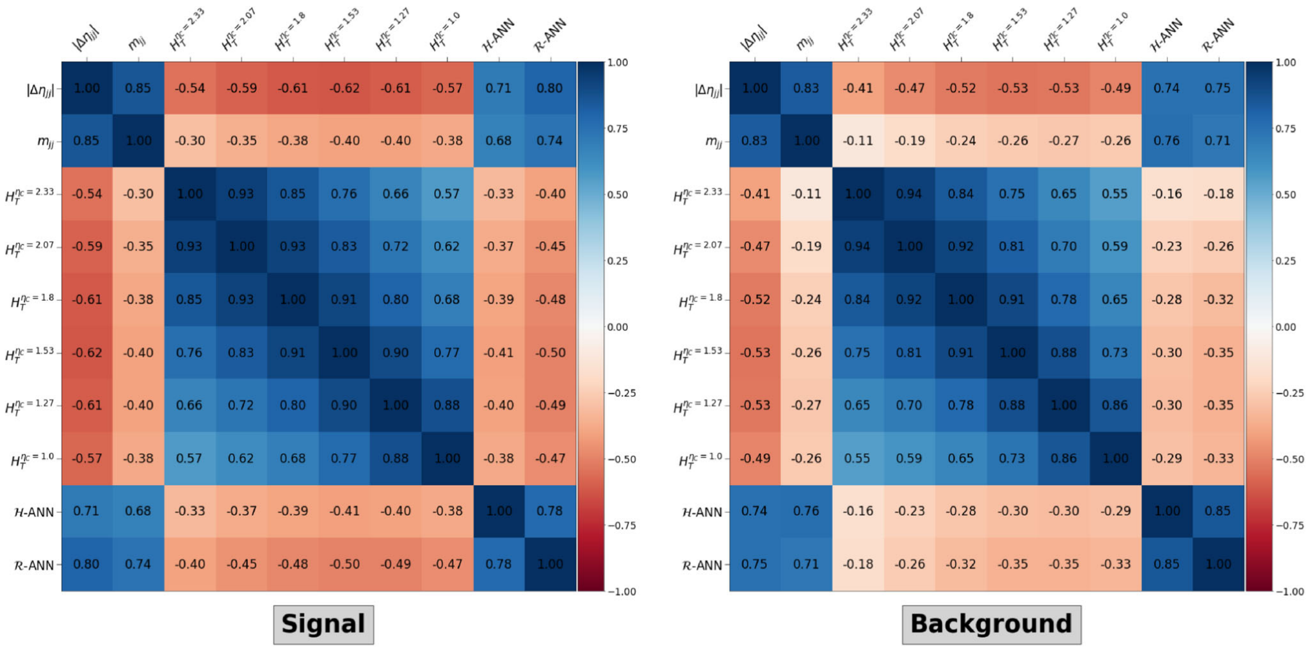


Fig. 21 Correlation between the high-level variables $\mathcal{H}$ and the network-outputs of $\mathcal{R}$-ANN and $\mathcal{H}$-ANN for (left) signal and (right) background. For better representation we have chosen variables with non-negligible correlations with the network outputs

separation, and the two kinematic variables $|\Delta\eta_{jj}|$ and m_{jj} are shown in Fig. 21. As expected, we see that the $\mathcal{R}$ variables are highly correlated with one another, which decreases with increasing distance in η_C. Another highlight is the negative correlation between them and the kinematic variables. It can be understood if we recall that the dominant radiation in the tower comes from the two leading jets, and an increase in $|\Delta\eta_{jj}|$ will decrease the calorimeter hits in the central regions. In the case of correlations between neural-network outputs and their respective inputs, the sign of the correlation is not much relevant for binary classification due to the probabilistic interpretation of the outputs y_i: $y_0 + y_1 = 1$ and $y_i > 0$. On the contrary, the relative difference in sign and magnitude in correlations between the different input features and the output is relevant. In the case of $\mathcal{H}$-ANN, we can see that in terms of both magnitude (importance as plotted in Fig. 7) and sign (as discussed here), the relations amongst $\mathcal{K}$ and $\mathcal{R}$ variables are carried over to their corresponding correlations with the network-output.

References

1. K. Albertsson et al., Machine learning in high energy physics community white paper. J. Phys. Conf. Ser. **1085**(2), 022008 (2018)
2. A. Aurisano, A. Radovic, D. Rocco, A. Himmel, M.D. Messier, E. Niner, G. Pawloski, F. Psihas, A. Sousa, P. Vahle, A convolutional neural network neutrino event classifier. JINST **11**(09), P09001 (2016)
3. E.L. Yates, MicroBooNE Investigation of Low-Energy Excess Using Deep Learning Algorithms, In Meeting of the APS Division of Particles and Fields, 2017
4. A. Radovic, M. Williams, D. Rousseau, M. Kagan, D. Bonacorsi, A. Himmel, A. Aurisano, K. Terao, T. Wongjirad, Machine learning at the energy and intensity frontiers of particle physics. Nature **560**(7716), 41–48 (2018)
5. D. Guest, K. Cranmer, D. Whiteson, Deep learning and its application to LHC physics. Ann. Rev. Nucl. Part. Sci. **68**, 161–181 (2018)
6. D. Bourilkov, Machine and deep learning applications in particle physics. Int. J. Mod. Phys. A **34**(35), 1930019 (2020)
7. J.H. Kim, M. Kim, K. Kong, K.T. Matchev, M. Park, Portraying double Higgs at the large Hadron Collider. JHEP **09**, 047 (2019)
8. J. Amacker et al., Higgs self-coupling measurements using deep learning in the $b\bar{b}b\bar{b}$ final state. **4** (2020)
9. P. Baldi, P. Sadowski, D. Whiteson, Enhanced Higgs Boson to $\tau^+\tau^-$ Search with Deep Learning. Phys. Rev. Lett. **114**(11), 111801 (2015)
10. R. Kogler et al., Jet substructure at the large Hadron Collider: experimental review. Rev. Mod. Phys. **91**(4), 045003 (2019)
11. S.D. Ellis, C.K. Vermilion, J.R. Walsh, Techniques for improved heavy particle searches with jet substructure. Phys. Rev. D **80**, 051501 (2009)
12. J.M. Butterworth, A.R. Davison, M. Rubin, G.P. Salam, Jet substructure as a new Higgs search channel at the LHC. Phys. Rev. Lett. **100**, 242001 (2008)
13. G.P. Salam, Towards jetography. Eur. Phys. J. C **67**, 637–686 (2010)
14. J. Shelton, Jet Substructure. in *Theoretical Advanced Study Institute in Elementary Particle Physics: Searching for New Physics at Small and Large Scales*, pp. 303–340, (2013)
15. S. Marzani, G. Soyez, M. Spannowsky, *Looking inside jets: an introduction to jet substructure and boosted-object phenomenology*, vol. 958 (Springer, New York, 2019)
16. A.M. Sirunyan et al., Search for electroweak production of a vector-like quark decaying to a top quark and a Higgs boson using boosted topologies in fully hadronic final states. JHEP **04**, 136 (2017)
17. A. Das, P. Konar, A. Thalapillil, Jet substructure shedding light on heavy Majorana neutrinos at the LHC. JHEP **02**, 083 (2018)
18. A. Bhardwaj, A. Das, P. Konar, A. Thalapillil, Looking for minimal inverse seesaw scenarios at the LHC with Jet substructure techniques. J. Phys. G **47**(7), 075002 (2020)
19. A. Bhardwaj, P. Konar, T. Mandal, S. Sadhukhan, Probing the inert doublet model using jet substructure with a multivariate analysis. Phys. Rev. D **100**(5), 055040 (2019)
20. R. Patrick, P. Sharma, A.G. Williams, Exploring a heavy charged Higgs using jet substructure in a fully hadronic channel. Nucl. Phys. B **917**, 19–30 (2017)
21. Z. Kang, P. Ko, J. Li, New physics opportunities in the boosted Di-Higgs–Boson plus missing transverse energy signature. Phys. Rev. Lett. **116**(13), 131801 (2016)
22. A. Bhardwaj, J. Dutta, P. Konar, B. Mukhopadhyaya, S.K. Rai, Boosted jet techniques for a supersymmetric scenario with gravitino LSP. **7** (2020)
23. S. Banerjee, C. Englert, R.S. Gupta, M. Spannowsky, Probing electroweak precision physics via boosted Higgs-strahlung at the LHC. Phys. Rev. D **98**(9), 095012 (2018)
24. J. Cogan, M. Kagan, E. Strauss, A. Schwarztman, Jet-images: computer vision inspired techniques for jet tagging. JHEP **02**, 118 (2015)
25. L. de Oliveira, M. Kagan, L. Mackey, B. Nachman, A. Schwartzman, Jet-images—deep learning edition. JHEP **07**, 069 (2016)
26. J. Barnard, E.N. Dawe, M.J. Dolan, N. Rajcic, Parton shower uncertainties in jet substructure analyses with deep neural networks. Phys. Rev. D **95**(1), 014018 (2017)
27. P.T. Komiske, E.M. Metodiev, M.D. Schwartz, Deep learning in color: towards automated quark/gluon jet discrimination. JHEP **01**, 110 (2017)
28. P. Baldi, K. Bauer, C. Eng, P. Sadowski, D. Whiteson, Jet substructure classification in high-energy physics with deep neural networks. Phys. Rev. D **93**(9), 094034 (2016)
29. G. Kasieczka, T. Plehn, M. Russell, T. Schell, Deep-learning top taggers or the end of QCD? JHEP **05**, 006 (2017)
30. S. Macaluso, D. Shih, Pulling out all the tops with computer vision and deep learning. JHEP **10**, 121 (2018)
31. T.S. Roy, A.H. Vijay, A robust anomaly finder based on autoencoders. **3** (2019)
32. K. Datta, A. Larkoski, How much information is in a jet? JHEP **06**, 073 (2017)
33. S.H. Lim, M.M. Nojiri, Spectral analysis of jet substructure with neural networks: boosted Higgs case. JHEP **10**, 181 (2018)
34. A. Chakraborty, S.H. Lim, M.M. Nojiri, Interpretable deep learning for two-prong jet classification with jet spectra. JHEP **19**, 135 (2020)
35. G. Louppe, K. Cho, C. Becot, K. Cranmer, QCD-aware recursive neural networks for jet physics. JHEP **01**, 057 (2019)
36. E.M. Metodiev, B. Nachman, J. Thaler, Classification without labels: learning from mixed samples in high energy physics. JHEP **10**, 174 (2017)
37. J. Guo, J. Li, T. Li, X. Fangzhou, W. Zhang, Deep learning for R-parity violating supersymmetry searches at the LHC. Phys. Rev. D **98**(7), 076017 (2018)
38. A. De Simone, T. Jacques, Guiding new physics searches with unsupervised learning. Eur. Phys. J. C **79**(4), 289 (2019)
39. J. Hajer, Y.-Y. Li, T. Liu, H. Wang, Novelty detection meets Collider physics. Phys. Rev. D **101**(7), 076015 (2020)
40. B. Bhattacherjee, S. Mukherjee, R. Sengupta, Study of energy deposition patterns in hadron calorimeter for prompt and displaced jets using convolutional neural network. JHEP **19**, 156 (2020)
41. A. Blance, M. Spannowsky, P. Waite, Adversarially-trained autoencoders for robust unsupervised new physics searches. JHEP **10**, 047 (2019)
42. S. Jung, D. Lee, K.-P. Xie, Beyond $M_{t\bar{t}}$: learning to search for a broad $t\bar{t}$ resonance at the LHC. Eur. Phys. J. C **80**(2), 105 (2020)
43. P. Windischhofer, M. Zgubi, D. Bortoletto, Preserving physically important variables in optimal event selections: a case study in Higgs physics. JHEP **07**, 001 (2020)
44. J. Ren, W. Lei, J.M. Yang, J. Zhao, Exploring supersymmetry with machine learning. Nucl. Phys. B **943**, 114613 (2019)

45. M.t Abdughani, D. Wang, L. Wu, J.M. Yang, J. Zhao, Probing triple Higgs coupling with machine learning at the LHC. **5** (2020)
46. S. Diefenbacher, H. Frost, G. Kasieczka, T. Plehn, J.M. Thompson, CapsNets Continuing the Convolutional Quest. (2019)
47. W. Bhimji, S.A. Farrell, T. Kurth, M. Paganini, E. Racah, Deep neural networks for physics analysis on low-level whole-detector data at the LHC. J. Phys. Conf. Ser. **1085**(4), 042034 (2018)
48. M. Andrews, M. Paulini, S. Gleyzer, B. Poczos, Exploring end-to-end deep learning applications for event classification at CMS. EPJ Web Conf. **214**, 06031 (2019)
49. J. Lin, M. Freytsis, I. Moult, B. Nachman, Boosting $H \to b\bar{b}$ with Machine Learning. JHEP **10**, 101 (2018)
50. A.J. Larkoski, I. Moult, B. Nachman, Jet substructure at the large Hadron Collider: a review of recent advances in theory and machine learning. Phys. Rep. **841**, 1–63 (2020)
51. S. Carrazza, Machine learning challenges in theoretical HEP. J. Phys. Conf. Ser. **1085**(2), 022003 (2018)
52. M. Abdughani, J. Ren, W. Lei, J.M. Yang, J. Zhao, Supervised deep learning in high energy phenomenology: a mini review. Commun. Theor. Phys. **71**(8), 955 (2019)
53. S. Troyan, Y. Dokshitzer, V. Khoze, i Proceedings of the international conference. *Physics in Collision VI, page 365, Chicago, Illinois*. World Scientific, Singapore (1986)
54. J.D. Bjorken, Rapidity gaps and jets as a new-physics signature in very-high-energy hadron-hadron collisions. Phys. Rev. D **47**, 101–113 (1993)
55. R.S. Fletcher, T. Stelzer, Rapidity gap signals in higgs-boson production at the ssc. Phys. Rev. D **48**, 5162–5167 (1993)
56. V. Khachatryan et al., Search for the standard model Higgs boson produced through vector boson fusion and decaying to $b\bar{b}$. Phys. Rev. D **92**(3), 032008 (2015)
57. R.N. Cahn, S. Dawson, Production of very massive higgs bosons. Phys. Lett. B **136**(3), 196–200 (1984)
58. D.L. Rainwater, D. Zeppenfeld, Searching for $H \to \gamma\gamma$ in weak boson fusion at the LHC. JHEP **12**, 005 (1997)
59. D.L. Rainwater, D. Zeppenfeld, Observing $H \to W^*W^* \to e^{\pm}\mu \mp \neg p_T$ in weak boson fusion with dual forward jet tagging at the CERN LHC. Phys. Rev. D **60**, 113004 (1999). [Erratum: Phys.Rev.D 61, 099901 (2000)]
60. D.L. Rainwater, D. Zeppenfeld, K. Hagiwara, Searching for $H \to \tau^+\tau^-$ in weak boson fusion at the CERN LHC. Phys. Rev. D **59**, 014037 (1998)
61. T. Plehn, D.L. Rainwater, D. Zeppenfeld, Determining the structure of Higgs Couplings at the LHC. Phys. Rev. Lett. **88**, 051801 (2002)
62. V. Hankele, G. Klamke, D. Zeppenfeld, T. Figy, Anomalous Higgs boson couplings in vector boson fusion at the CERN LHC. Phys. Rev. D **74**, 095001 (2006)
63. T. Han, S. Mukhopadhyay, B. Mukhopadhyaya, W. Yongcheng, Measuring the CP property of Higgs coupling to tau leptons in the VBF channel at the LHC. JHEP **05**, 128 (2017)
64. D. Zanzi, ATLAS, CMS Collaborations. Measurement of the Higgs Boson Couplings and CP Structure Using Tau Leptons at the LHC. *Nuclear and Particle Physics Proceedings*, 287–288:115–118, (2017)
65. O.J.P. Eboli, D. Zeppenfeld, Observing an invisible Higgs boson. Phys. Lett. B **495**, 147–154 (2000)
66. A. Datta, P. Konar, B. Mukhopadhyaya, Signals of neutralinos and charginos from gauge boson fusion at the Large Hadron Collider. Phys. Rev. D **65**, 055008 (2002)
67. D. Choudhury, A. Datta, K. Huitu, P. Konar, S. Moretti, B. Mukhopadhyaya, Slepton production from gauge boson fusion. Phys. Rev. D **68**, 075007 (2003)
68. P. Konar, D. Zeppenfeld, Next-to-leading order QCD corrections to slepton pair production via vector-boson fusion. Phys. Lett. B **647**, 460–465 (2007)
69. R.E. Shrock, M. Suzuki, Invisible decays of Higgs Bosons. Phys. Lett. B **110**, 250 (1982)
70. G. Arcadi, A. Djouadi, M. Raidal, Dark Matter through the Higgs portal. Phys. Rep. **842**, 1–180 (2020)
71. A. Djouadi, O. Lebedev, Y. Mambrini, J. Quevillon, Implications of LHC searches for Higgs-portal dark matter. Phys. Lett. B **709**, 65–69 (2012)
72. A. Djouadi, A. Falkowski, Y. Mambrini, J. Quevillon, Direct detection of Higgs-Portal dark matter at the LHC. Eur. Phys. J. C **73**(6), 2455 (2013)
73. H. Han, J.M. Yang, Y. Zhang, S. Zheng, Collider signatures of Higgs-portal scalar dark matter. Phys. Lett. B **756**, 109–112 (2016)
74. P. Seungwon Baek, W.-I.P. Ko, E. Senaha, Higgs Portal vector dark matter: revisited. JHEP **05**, 036 (2013)
75. K. Belotsky, M. Daniele Fargion, R.K. Khlopov, K. Shibaev, Invisible Higgs boson decay into massive neutrinos of fourth generation. Phys. Rev. D **68**, 054027 (2003)
76. G. Bambhaniya, S. Goswami, S. Khan, P. Konar, T. Mondal, Looking for hints of a reconstructible seesaw model at the Large Hadron Collider. Phys. Rev. D **91**, 075007 (2015)
77. G. Bélanger, F. Boudjema, A. Cottrant, R.M. Godbole, A. Semenov, The mssm invisible higgs in the light of dark matter and g-2. Phys. Lett. B **519**(1–2), 93–102 (2001)
78. A. Datta, P. Konar, B. Mukhopadhyaya, Invisible charginos and neutralinos from gauge boson fusion: a way to explore anomaly mediation? Phys. Rev. Lett. **88**, 181802 (2002)
79. G.F. Giudice, R. Rattazzi, J.D. Wells, Graviscalars from higher-dimensional metrics and curvature-higgs mixing. Nuclear Physics B **595**(1–2), 250–276 (2001)
80. K. Hagiwara, P. Konar, Q. Li, K. Mawatari, D. Zeppenfeld, Graviton production with 2 jets at the LHC in large extra dimensions. JHEP **04**, 019 (2008)
81. G. Klamke, D. Zeppenfeld, Higgs plus two jet production via gluon fusion as a signal at the CERN LHC. JHEP **04**, 052 (2007)
82. Search for invisible Higgs boson decays with vector boson fusion signatures with the ATLAS detector using an integrated luminosity of 139 fb^{-1}. Technical Report ATLAS-CONF-2020-008, CERN, Geneva, (2020)
83. A.M. Sirunyan et al., Search for invisible decays of a Higgs boson produced through vector boson fusion in proton-proton collisions at $\sqrt{s} = 13$ TeV. Phys. Lett. B **793**, 520–551 (2019)
84. G. Tao, H. Valencia, S. Willenbrock, Structure function approach to vector boson scattering in p p collisions. Phys. Rev. Lett. **69**, 3274–3277 (1992)
85. T. Figy, C. Oleari, D. Zeppenfeld, Next-to-leading order jet distributions for Higgs boson production via weak boson fusion. Phys. Rev. D **68**, 073005 (2003)
86. F.A. Dreyer, A. Karlberg, Vector-Boson fusion Higgs production at three loops in QCD. Phys. Rev. Lett. **117**(7), 072001 (2016)
87. M. Ciccolini, A. Denner, S. Dittmaier, Strong and electroweak corrections to the production of Higgs + 2jets via weak interactions at the LHC. Phys. Rev. Lett. **99**, 161803 (2007)
88. T. Liu, K. Melnikov, A.A. Penin, Nonfactorizable QCD effects in Higgs Boson production via vector Boson fusion. Phys. Rev. Lett. **123**(12), 122002 (2019)
89. A. Datta, P. Konar, B. Mukhopadhyaya, New Higgs signals from vector boson fusion in R-parity violating supersymmetry. Phys. Rev. D **63**, 095009 (2001)
90. P. Konar, B. Mukhopadhyaya, Gauge boson fusion as a probe of inverted hierarchies in supersymmetry. Phys. Rev. D **70**, 115011 (2004)
91. A.G. Delannoy et al., Probing dark matter at the LHC using vector Boson fusion processes. Phys. Rev. Lett. **111**, 061801 (2013)
92. A. Berlin, T. Lin, M. Low, L.-T. Wang, Neutralinos in vector Boson fusion at high energy colliders. Phys. Rev. D **91**(11), 115002 (2015)

93. I.W. Stewart, F.J. Tackmann, W.J. Waalewijn, N-Jettiness: an inclusive event shape to veto jets. Phys. Rev. Lett. **105**, 092002 (2010)
94. F. Braren, Selection of Vector Boson Fusion Events in the $H \rightarrow \gamma\gamma$ Decay Channel Using an Inclusive Event Shape. Master's thesis, Hamburg U., **4** (2015)
95. M. Aaboud et al., Search for invisible Higgs boson decays in vector boson fusion at $\sqrt{s} = 13$ TeV with the ATLAS detector. Phys. Lett. B **793**, 499–519 (2019)
96. C. Bernaciak, T. Plehn, P. Schichtel, J. Tattersall, Spying an invisible Higgs boson. Phys. Rev. D **91**, 035024 (2015)
97. A. Biekötter, F. Keilbach, R. Moutafis, T. Plehn, J. Thompson, Tagging jets in invisible Higgs searches. SciPost Phys. **4**(6), 035 (2018)
98. J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H.S. Shao, T. Stelzer, P. Torrielli, M. Zaro, The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. JHEP **07**, 079 (2014)
99. T. Sjöstrand, S. Ask, J.R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C.O. Rasmussen, P.Z. Skands, An introduction to PYTHIA 8.2. Comput. Phys. Commun. **191**, 159–177 (2015)
100. J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, M. Selvaggi, DELPHES 3, A modular framework for fast simulation of a generic collider experiment. JHEP **02**, 057 (2014)
101. M. Cacciari, G.P. Salam, G. Soyez, FastJet user manual. Eur. Phys. J. C **72**, 1896 (2012)
102. A. Alloul, N.D. Christensen, C. Degrande, C. Duhr, B. Fuks, FeynRules 2.0—a complete toolbox for tree-level phenomenology. Comput. Phys. Commun. **185**, 2250–2300 (2014)
103. T.G. Rizzo, Gluon final states in higgs-boson decay. Phys. Rev. D **22**, 178–183 (1980)
104. R.P. Kauffman, S.V. Desai, Production of a higgs pseudoscalar plus two jets in hadronic collisions. Phys. Rev. D **59**, 057504 (1999)
105. D. de Florian et al, Handbook of LHC Higgs cross sections: 4. Deciphering the Nature of the Higgs Sector. 2/2017, 10 (2016)
106. S. Hoeche, F. Krauss, N. Lavesson, L. Lonnblad, M. Mangano, A. Schalicke, S. Schumann, Matching parton showers and matrix elements. In *HERA and the LHC: A Workshop on the Implications of HERA for LHC Physics: CERN - DESY Workshop 2004/2005 (Midterm Meeting, CERN, 11-13 October 2004; Final Meeting, DESY, 17-21 January 2005)*, pages 288–289, (2005)
107. I.W. Stewart, F.J. Tackmann, J.R. Walsh, S. Zuberi, Jet p_T resummation in Higgs production at $NNLL' + NNLO$. Phys. Rev. D **89**(5), 054001 (2014)
108. M. Cacciari, F.A. Dreyer, A. Karlberg, G.P. Salam, G. Zanderighi, Fully differential vector-boson-fusion Higgs production at next-to-next-to-leading order. Phys. Rev. Lett. **115**(8), 082002 (2015). [Erratum: Phys.Rev.Lett. 120, 139901 (2018)]
109. J.M. Lindert et al., Precise predictions for $V+$ jets dark matter backgrounds. Eur. Phys. J. C **77**(12), 829 (2017)
110. C. Oleari, D. Zeppenfeld, QCD corrections to electroweak nu(l) j j and l+ l- j j production. Phys. Rev. D **69**, 093004 (2004)
111. L.G. Almeida, M. Backović, M. Cliche, S.J. Lee, M. Perelstein, Playing tag with ANN: boosted top identification with pattern recognition. JHEP **07**, 086 (2015)
112. M. Erdmann, E. Geiser, Y. Rath, M. Rieger, Lorentz boost networks: autonomous physics-inspired feature engineering. JINST **14**(06), P06006 (2019)
113. A. Butter, G. Kasieczka, T. Plehn, M. Russell, Deep-learned top tagging with a Lorentz layer. SciPost Phys. **5**(3), 028 (2018)
114. M.M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, P. Vandergheynst, Geometric deep learning: Going beyond euclidean data. IEEE Signal Processing Magazine **34**(4), 18–42 (2017)
115. J. Ren, W. Lei, J.M. Yang, Unveiling CP property of top-Higgs coupling with graph neural networks at the LHC. Phys. Lett. B **802**, 135198 (2020)
116. M. Abdughani, J. Ren, W. Lei, J.M. Yang, Probing stop pair production at the LHC with graph neural networks. JHEP **08**, 055 (2019)
117. A. Mullin, H.y Pacey, M.l Parker, M. White, S. Williams, Does SUSY have friends? A new approach for LHC event analysis. (2019)
118. J. Shlomi, P.Battaglia, J.-R. Vlimant, Graph Neural Networks in Particle Physics. **7** (2020)
119. M. Cacciari, G.P. Salam, G. Soyez, The anti-k_t jet clustering algorithm. JHEP **04**, 063 (2008)
120. D. Boutigny et al. *The BABAR physics book: Physics at an asymmetric B factory*. **10** (1998)
121. F. Chollet et al. Keras. https://keras.io, (2015)
122. M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, (2015). Software available from tensorflow.org
123. N. Tishby, N. Zaslavsky, Deep learning and the information bottleneck principle. in *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, (2015)
124. P. Mehta, D.J. Schwab, An exact mapping between the variational renormalization group and deep learning. (2014)
125. S. Kullback, R.A. Leibler, On information and sufficiency. Ann. Math. Stat. **22**(1), 79–86 (1951)
126. T. Dozat, Incorporating nesterov momentum into adam. in *ICLR 2016 Workshop*, (2016)
127. Y.E. Nesterov, A method for solving the convex programming problem with convergence rate $o(1/k^2)$. Dokl. Akad. Nauk SSSR **269**, 543–547 (1983)
128. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
129. T. Junk, Confidence level computation for combining searches with small statistics. Nucl. Instrum. Methods A **434**, 435–443 (1999)
130. A.L. Read, Presentation of search results: the CLs technique. J. Phys. Nucl. Part. Phys. **28**(10), 2693–2704 (2002)
131. G. Cowan, K. Cranmer, E. Gross, O. Vitells, Asymptotic formulae for likelihood-based tests of new physics. Eur. Phys. J. C **71**, 1554 (2011). [Erratum: Eur.Phys.J.C 73, 2501 (2013)]
132. A. Wald, Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans. Am. Math. Soc. **54**(3), 426–482 (1943)
133. K. Cranmer, G. Lewis, L. Moneta, A. Shibata, W. Verkerke, HistFactory: A tool for creating statistical models for use with RooFit and RooStats. **6** (2012)
134. L. Moneta, K. Belasco, K.S. Cranmer, S. Kreiss, A. Lazzaro, D. Piparo, Gregory Schott, Wouter Verkerke, Matthias Wolf, The RooStats Project. *PoS*, ACAT2010:057, (2010)
135. P.T. Komiske, E.M. Metodiev, B. Nachman, M.D. Schwartz, Pileup Mitigation with Machine Learning (PUMML). JHEP **12**, 051 (2017)

136. J. Arjona Martínez, O. Cerri, M. Pierini, M. Spiropulu, J.-R. Vlimant, Pileup mitigation at the Large Hadron Collider with graph neural networks. Eur. Phys. J. Plus **134**(7), 333 (2019)
137. P.T. Komiske, E.M. Metodiev, J. Thaler, Metric space of Collider events. Phys. Rev. Lett. **123**(4), 041801 (2019)
138. P.T. Komiske, E.M. Metodiev, J. Thaler, The hidden geometry of particle collisions. JHEP **07**, 006 (2020)
139. M. Cacciari, G.P. Salam, Pileup subtraction using jet areas. Phys. Lett. B **659**, 119–126 (2008)
140. P. Berta, M. Spousta, D.W. Miller, R. Leitner, Particle-level pileup subtraction for jets and jet shapes. JHEP **06**, 092 (2014)
141. V. Khachatryan et al., Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV. JINST **12**(02), P02014 (2017)
142. S. van der Walt, S.C. Colbert, G. Varoquaux, The numpy array: a structure for efficient numerical computation. Comput. Sci. Eng. **13**(2), 22–30 (2011)