Identification of Hadronically-Decaying $W$ Bosons and Top Quarks Using High-Level Features as Input to Boosted Decision Trees and Deep Neural Networks in ATLAS at $\sqrt{s} = 13$ TeV

The ATLAS Collaboration

12th April 2017

The application of boosted decision trees and deep neural networks to the identification of hadronically-decaying $W$ bosons and top quarks using high-level jet observables as inputs is investigated using Monte Carlo simulations. In the case of both boosted decision trees and deep neural networks, the use of machine learning techniques is found to improve the background rejection with respect to simple reference single jet substructure and mass taggers. Linear correlations between the resulting classifiers and the substructure variables are also presented.

© 2017 CERN for the benefit of the ATLAS Collaboration. Reproduction of this article or parts of it is allowed as specified in the CC-BY-4.0 license.
1 Introduction

At the Large Hadron Collider (LHC) [1], the identification of the original particle that initiated a hadronic jet can be used as an effective tool to reject events produced by background processes and improve the sensitivity in searches for physics beyond the Standard Model. In the case of the ATLAS b-tagging algorithm, this is done through the identification of multiple independently motivated observables which are subsequently combined in a boosted decision tree (BDT) to better exploit the full discrimination power of the ensemble of observables [2]. For W-boson and top-quark jet identification, the physically motivated observables concern the radiation pattern of the jet and have been used to effectively tag large radius jets [3–7]. It has been shown that the combination of these observables to create a multivariate W-boson or top-quark classifier provides higher discrimination, albeit to differing degrees [8, 9].

The aim of this study is to assess the ability of BDT and deep neural network (DNN) algorithms to discriminate samples of W-boson or top-quark jets from the dominant gluon or other (non-top) quark jet background. The Monte Carlo samples used and the event selections applied are described in Sections 3 and 4, respectively. Boosted decision tree and deep neural network classifiers are implemented and studied in parallel, as described in Section 5, allowing for comparisons of the performance of discriminants built from different algorithms in a similar way as studied in Reference [10]. For comparability with the simple taggers similar to those used in many recent ATLAS publications1 [11–14], which use a transverse momentum-dependent selection defined by the jet mass \( m^{\text{calo}} \) and a second jet substructure variable, the jet mass is excluded from training. However, the BDT and DNN algorithms are trained using variables that are correlated to each other and the jet mass (such as \( ECF_2 \) and \( \sqrt{d_{12}} \) as described in Section 5) and the correlations are quantified in Section 6.2.

Likewise, although the jet transverse momentum \( p_T \) is not included in the training, a similar variable \( ECF_1 \), which is used in the construction of several substructure variables, is included in the training. In Section 6 after a common jet mass selection is applied, the performance of each of the discriminants (BDT, DNN, simple reference tagger) is compared. The correlations of these classifiers with individual jet substructure observables and jet mass are studied to assess the overall utility of the resultant taggers.

2 ATLAS detector

The ATLAS detector [15] at the LHC covers nearly the entire solid angle around the collision point. It consists of an inner tracking detector surrounded by a thin superconducting solenoid, electromagnetic and hadronic calorimetry, and a muon spectrometer composed of three large superconducting toroid magnets. For this study, the most important subsystem is the calorimetry, which covers the pseudorapidity range \( |\eta| < 4.9 \). Within the region \( |\eta| < 3.2 \), electromagnetic calorimetry is provided by barrel and endcap high-granularity lead/liquid-argon (LAr) electromagnetic calorimeters, with an additional thin LAr presampler covering \( |\eta| < 1.8 \), to correct for energy loss in material upstream of the calorimeters. Hadronic calorimetry is provided by a steel/scintillating-tile calorimeter, segmented into three barrel structures within \( |\eta| < 1.7 \), and two copper/LAr hadronic endcap calorimeters. The forward region \( 3.2 < |\eta| < 4.9 \) is instrumented with forward copper/LAr and tungsten/LAr calorimeter modules.

---

1 The primary difference between the simple taggers used as a baseline here and those commonly used in ATLAS is that here, the selection on the jet mass is fixed rather than tuned as a function of jet \( p_T \).
3 Samples

Three Monte Carlo (MC) samples for both signal and background processes are used for this study. The dijet process is used to simulate jets from gluons and other (non-top) quarks and this process is simulated using the PYTHIA8 [16] generator with the NNPDF2.3LO [17] parton distribution function (PDF) set and the A14 tune [18]. Events are generated in slices of jet $p_T$ to sufficiently populate the kinematic region of interest (between 200 and 2000 GeV). Event-by-event weights are applied to correct for this generation and to produce the expected smoothly falling $p_T$ distribution. The signal samples, containing either high-$p_T$ top-quark or $W$-boson jets are obtained from two physics processes modelling phenomena beyond the Standard Model. For the $W$-boson sample, the high-mass $W' \rightarrow WZ \rightarrow qqqq$ is used. For the top-quark sample, high-mass $Z' \rightarrow t\bar{t}$ events are used as a source of signal jets.

Both the $W$ bosons and top quarks are required to decay hadronically. The two signal processes are simulated using the PYTHIA8 generator with the NNPDF2.3LO PDF set and A14 tune for multiple values of the resonance ($W'$ or $Z'$ boson) mass between 400 and 5000 GeV in order to populate the entire $p_T$ range from 200 to 2000 GeV and to ensure that the calculated signal efficiencies have small statistical uncertainties.

4 Baseline event and object selection

In this MC-based study, the event selection is aimed at isolating ensembles of jets which are representative of those originating from either $W$ bosons or top quarks (signal) and gluon or other (non-top) quarks (background). Initially, events which contain a reconstructed primary vertex with at least two tracks are selected. Jets are reconstructed from topological clusters of calorimeter cells calibrated using the local cell weighting procedure [19] and are formed using the anti-$k_t$ algorithm [20] with a distance parameter of $R = 1.0$. To reduce the effects of pileup and underlying-event contamination, jets are groomed using a trimming algorithm with $R_{\text{sub}} = 0.2$ and $f_{\text{cut}} = 5\%$ [21]. All substructure observables are calculated using the set of constituents resulting from this procedure. In each event the two highest-$p_T$ jets are retained if they satisfy $|\eta| < 2$ and have a $p_T$ between 200 and 2000 GeV in the case of the $W$-boson or background quark and gluon jets, and 350 and 2000 GeV for top-quark jets. The selected jets are then calibrated in a two-step procedure that first corrects the jet-energy scale and then the jet-mass scale [22]. For the purpose of identifying the flavor of the jet at truth level, a second set of jets is formed from truth particles with lifetimes greater than 10 picoseconds, except for muons and neutrinos, which are not included. These jets are reconstructed with the anti-$k_t$ algorithm with a distance parameter of $R = 1.0$ but are not modified with the trimming algorithm. These jets are referred to as truth jets.

In this study, signal jets are defined as hadronically-decaying $W$ bosons or top quarks when all partonic decay products are fully contained within the region of interest of the reconstructed jet in a three-step process. First, reconstructed jets are matched to truth jets. Next, those truth jets are matched to the truth $W$-boson and top-quark particles ($W, \text{top}$). Lastly their partonic decay products ($q_1, q_2, b$) are matched to the initial reconstructed jet. All stages of this matching procedure are performed with simple matching in $(\eta, \phi)$ with $\Delta R(J_{\text{reco}}, J_{\text{truth}}) < 0.75^2$. The truth matching procedure is summarised in Table 1. The resulting signal and background jets are then used in the following as inputs for the simple reference tagger and the

---

$^2$The requirement $\Delta R(J_{\text{reco}}, J_{\text{truth}}) < 0.75$ is found to give a good separation between fully and partially contained jets for medium jet $p_T < 500$ GeV where the angular separation of the top-quark/$W$-boson decay products tends to be large.
multivariate taggers. For the latter, they are randomly split into training and testing sets in order to evaluate the performance of the BDT and DNN based taggers independent of the set which they were trained on. Object selection for the training and testing sets and the resulting number of jets are summarised in Table 1 and in Table 2, respectively.

**4.1 Testing set and testing event weights**

When evaluating the performance of the taggers, 30% of each signal and background sample (testing set) are used. The simulated signal samples (separately for $W$ bosons and top quarks) are combined and weighted such that the truth $p_T$ distribution of the ensemble of signal jets matches that of the dijet background to remove any bias on the tagging performance due to the difference in the $p_T$ spectrum of the signal and background jet samples.

**4.2 Training set and training event weights**

The training of each tagging algorithm is performed using a training set consisting of 70% of the signal samples. Two separate background training samples are created to match the number of jets in each of the signal training sets. The requirement of equal numbers of signal and background jets avoids using very large training weights which would not be optimal for the training and also prevents learning one flavour more than the other because of the limited sample size. To ensure that all used jet substructure features are well defined for the training jets, two additional selection criteria are applied on the jet mass ($m^{\text{calo}} > 40$ GeV) and number of constituents ($N^{\text{const}} > 2$) as summarised in Table 1. The jets which fail the training criteria are not used in the training.

The resulting signal and background training sets are combined and weighted individually for each topology such that the truth-jet $p_T$ distribution is uniform. By enforcing a uniform truth jet $p_T$ distribution for both signal and background jets, the algorithms are prevented from discriminating using differences in $p_T$ distributions. Each multivariate algorithm is trained inclusive in the truth jet transverse momentum range of $200 < p_T^{\text{true}} < 2000$ GeV due to the limited sample size although it should be noted that the algorithms are still capable of learning $p_T$-dependent features, primarily through the inclusion of other $p_T$-dependent observables such as $ECF_1$.
Table 2: Summary of training and testing sample sizes for the signal and background samples.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Training (Top quark)</th>
<th>Training (W boson)</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of jets</td>
<td>$10^6$</td>
<td>$7 \times 10^5$</td>
<td>$4 \times 10^5$</td>
</tr>
</tbody>
</table>

5 Analysis

Two different machine learning techniques are applied, a boosted decision tree implemented within the TMVA framework [23], and an artificial neural network implemented within the Keras framework [24]. Both techniques are used for W-boson and top-quark tagging and explore the same set of input variables with the primary difference being the final set of input observables chosen by the two methods separately, as shown in Table 3. The jet mass is excluded from the training in order to focus on the standard analysis usage, where a mass cut is followed by a substructure variable cut. The mass selection is then applied separately to start from the same set of signal-enriched events for performance comparisons between the multivariate discriminant and individual substructure variables. Note that although the jet mass is excluded from the training, substructure variables which are highly correlated with the jet mass, such as $E_{T,2}$, are included in the training.

For each multivariate technique, choosing the optimal set of input observables and training hyperparameters is important. A thorough but not exhaustive scan of calorimeter-related jet substructure variables as training inputs to the BDTs and DNNs is performed. The addition of well-known track-related variables which are useful in W-boson and top-quark tagging, such as the number of tracks associated to a jet $N_{trk}$ [12] or the use of b-tagging [25], are beyond the scope of this note. The input variables studied are listed in Table 3. It is desirable to use the fewest variables necessary to achieve maximal performance, as unneeded variables may introduce additional systematic uncertainties. The procedures are slightly different for BDT and DNN and are described in the following.

As described in Section 4 and listed in Table 1, a selection on jet mass and number of constituents is applied on the training jets. Therefore, jets in the testing set that fail the mass criteria are tagged as background jets. Jets that pass the mass requirement but fail the number of constituents requirement are tagged as signal jets. The fraction of jets found to be categorized in this way is less than 1%.

To judge the performance of these techniques, the two primary quantitative figures of merit, the signal efficiency and background rejection, are used and are defined as

$$\text{Signal Efficiency} = \epsilon_{\text{sig}} = \frac{N_{\text{tagged}}}{N_{\text{total}}^{\text{sig}}}$$

$$\text{Background Rejection} = \frac{1}{\epsilon_{\text{bkg}}} = \frac{N_{\text{total}}^{\text{bkg}}}{N_{\text{tagged}}^{\text{bkg}}}.$$  

Hyperparameters are the parameters that define the behavior of the multivariate algorithms, e.g., the number of nodes in a DNN hidden layer.
Table 3: Summary of tagging techniques and resultant variables that have been studied. In the case of the energy correlation observables, the angular exponent $\beta$ is set to 1.0 and for the $N$-subjettiness observables, the winner-take-all [40] configuration is used.

5.1 Boosted Decision Tree Training

To find the best BDT performance, single input variables are sequentially added to the network that give the largest increase in relative performance. This procedure is initiated with the input variable found to be most performant for $W$-boson [3] and top-quark [4] tagging. These observables are $D_2$ and $\tau_{32}$, respectively. The relative performance is evaluated using jets from the testing sample which pass the training criteria and with the training weights described in Section 4.2 in the kinematic range $200 < p_T^{true} < 2000$ GeV. During the input and hyperparameter optimization of the multivariate techniques, the jets which do not pass the jet mass ($m_{calo} > 40$ GeV) and number of constituents ($N_{const} > 2$) training criteria are not included in the relative signal efficiency or background rejection evaluation in order to purely evaluate the performance differences arising from the multivariate classifier. At each step, the variable which gives the greatest increase in relative background rejection at a fixed relative signal efficiency of 50%, when added to the existing set of variables, is retained. The minimum set of variables which reaches the highest relative background rejection within statistical uncertainties are selected. The minimal number of variables is 11 for $W$-boson tagging and 10 for top-quark tagging. The relative background rejections achieved are shown in Figure 1. In addition, the BDT relative background rejection using the inputs found to be optimal by the DNN (see Section 5.2) are shown for comparison, as explained in Section 6. Note that the performance of the BDT is improved when adding the $ECF$ variables from which $D_2$ is calculated, which suggests that more discriminating information is contained in the two-prong tagging variables ($ECF$) than is captured by power-counting alone ($D_2$). A summary of all the variables tested and chosen is shown in Table 10.
5.1.1 Hyperparameter optimization for Boosted Decision Tree

The performance of the BDT depends not only on the inputs, but also on the settings of the internal hyperparameters of the algorithm. These hyperparameters are summarized in Table 4 along with the optimal value determined for this application. To determine these values, a parameter scan of the hyperparameter choices shown in Table 5 is performed. The parameters are scanned in two stages so as to not increase the combinations to an unmanageable level.

In the first stage, the parameters (NTrees, MaxDepth, MinNodeSize), which are found to have a strong influence on the performance, are scanned with both the AdaBoost or GradientBoost settings. Next, using the optimal parameters from the first scan, a scan is performed on (nCuts, Bagged Fraction, Shrinkage). In general the performance increases with increasing complexity of the BDT training, as seen in the example shown in Figure 2 where the relative rejection improves with higher depth and number of trees at the cost of increasing computing time. Therefore, choosing the exact value of these hyperparameters is a trade-off between performance and training time. As similar sets of parameters were found to be optimal for $W$-boson tagging and top-quark tagging, only $W$-boson tagging results are presented in Figure 2 for simplicity. Similar to the BDT input optimization procedure, the relative performance is evaluated using the testing jet sample which pass the training criteria and with the training weights described in Section 4.2 inclusively in $p_T$. The chosen sets of parameters, shown in Table 4, are those that result in a close to optimal BDT performance but that still keep the training time at a reasonable level. Cross validation is carried out, where the impact of swapping the training and testing sets were studied, for the optimum BDTs and no significant overtraining is observed.
<table>
<thead>
<tr>
<th>Setting Name</th>
<th>Description</th>
<th>Chosen Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoostType</td>
<td>Type of boosting technique</td>
<td>GradientBoost</td>
</tr>
<tr>
<td>NTrees</td>
<td>Number of trees in the forest</td>
<td>500</td>
</tr>
<tr>
<td>MaxDepth</td>
<td>Max depth of the decision tree allowed</td>
<td>20</td>
</tr>
<tr>
<td>MinimumNodeSize</td>
<td>Minimum fraction of training events required in a leaf node</td>
<td>1.0%</td>
</tr>
<tr>
<td>Shrinkage</td>
<td>Learning rate for GradientBoost algorithm</td>
<td>0.5</td>
</tr>
<tr>
<td>UseBaggedBoost</td>
<td>Use only a random (bagged) subsample of all events for growing the trees in each iteration</td>
<td>True</td>
</tr>
<tr>
<td>BaggedSampleFraction</td>
<td>Relative size of bagged event sample to original size of the data sample</td>
<td>0.5</td>
</tr>
<tr>
<td>SeparationType</td>
<td>Separation criterion for node splitting</td>
<td>GiniIndex</td>
</tr>
<tr>
<td>nCuts</td>
<td>Number of grid points in variable range used in finding optimal cut in node splitting</td>
<td>500</td>
</tr>
</tbody>
</table>

Table 4: Brief description of the BDT hyperparameters and the chosen configuration.

<table>
<thead>
<tr>
<th>Setting Name</th>
<th>Value Tested</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTrees</td>
<td>[10, 50, 100, 200, 500, 850, 2000]</td>
</tr>
<tr>
<td>MaxDepth</td>
<td>[1, 2, 3, 5, 7, 10, 20, 50, 100]</td>
</tr>
<tr>
<td>MinimumNodeSize</td>
<td>[0.5, 1.0, 2.5, 5.0, 10, 20]</td>
</tr>
<tr>
<td>nCuts</td>
<td>[5, 10, 20, 50, 100, 500]</td>
</tr>
<tr>
<td>Bagged Fraction</td>
<td>[0.1, 0.3, 0.5, 0.7, 0.9]</td>
</tr>
<tr>
<td>Shrinkage</td>
<td>[0.05, 0.1, 0.3, 0.5, 0.7, 0.9]</td>
</tr>
</tbody>
</table>

Table 5: Range of hyperparameter scan for the first BDT grid search.

### 5.2 Deep Neural Network Training

Similar to the BDT training, the DNN is trained on different sets of input variables in order to find the optimum set of input variables which results in the most performant DNN tagger with the training weights described in Section 4.2. The relative performance is evaluated using the jets in the testing set that pass the training criteria. Unlike the BDT, sets of input variables are not defined by successively adding variables but are defined by grouping the inputs variables related to the corresponding signal. The grouping is based on how the variables are constructed and what features of a jet’s substructure they describe, e.g. the pronginess of the hard decay of the object which seeds the jet, the extent to which the constituents of a jet are spread out, the amount of information they provide on the jet-energy and jet-mass scale, and whether or not the observable is defined as a function of the other observables.

The reason behind not adding input variables iteratively to the DNN training is the difficulty and the excessive computation time required to optimize the DNN for a large number of input groups. Additionally, the DNN is trained with the inputs which were found to be optimal by the BDT for comparison. The chosen groups of inputs for $W$-boson tagging and top-quark tagging are listed in Tables 6 and 7, respectively. The relative background rejections achieved with the training weights described in Section 4.2 and inclusively
MaxDepth

W

background rejection at a fixed 50% relative signal efficiency is found to be 18 variables for top-quark tagging. The second-best performing DNN for top-quark tagging is chosen due to the smoothness of the network relative background rejection arising from an artifact. The chosen DNNs are summarised in Table 8.

As a final step, the smoothness of the DNN output is checked to avoid any jumps in the performance is evaluated with flat $p_T$ spectra. Chosen parameters and training time are highlighted as stars, and default TMVA settings are highlighted as circles. Intel® Xeon® CPU E5-2680 v3, 2.50 GHz with 4 GB memory per processor is used for training.

in jet $p_T$ are presented in Figure 3. It is confirmed that the performance of the DNN tagger is not only dependent on the number of inputs but also on how much of the necessary information is stored in different groups, based on the better performance of top-quark tagging Groups 4 and 7 (the BDT set) and W-boson tagging Group 7 (the BDT set). The number of variables necessary to achieve the highest relative background rejection at a fixed 50% relative signal efficiency is found to be 18 variables for W-boson tagging and 13 variables for top-quark tagging. A summary of all the variables tested and chosen is shown in Table 10.

5.2.1 Hyperparameter and architecture optimization for the Deep Neural Network

The performance of the DNN tagger is sensitive to the training hyperparameters and the architecture of the neural network. In this analysis, the optimizer algorithm Adam [43] is used as it has been proven to be successful for numerous similar machine learning problems. In order to identify the most performant DNN for each set of input groups, a grid search is carried out. In addition to the hyperparameter grid search of the listed training parameters and architecture types, the number of nodes in the hidden layers are varied and similar to the DNN input optimization the tagger which provides the highest relative background rejection at the 50% relative signal efficiency working point with the training weights described in Section 4.2 is chosen. As a final step, the smoothness of the DNN output is checked to avoid any jumps in the relative background rejection arising from an artifact. The chosen DNNs are summarised in Table 8. The second-best performing DNN for top-quark tagging is chosen due to the smoothness of the network

For a review of DNN terminology, related concepts, and an introduction to deep learning, see for instance References [41, 42].
output. All the DNNs are trained for 200 epochs (number of passes over all the training set jets), with an early-stopping parameter of 50 epochs which stops training when the monitored quantity loss evaluated on the validation set (15% of the general training set used to monitor the training but not used for DNN optimization) has stopped improving for 50 epochs. The list of scanned parameters and architecture types is summarised in Table 9, with more extensive definitions of hyperparameters and layer types found in the References listed in Table 8. The relative background rejections for the various DNN configurations, evaluated with the training weights described in Section 4.2 and inclusively in jet $p_T$, are presented in Figure 4. This evaluation is performed following the same evaluation procedure as described in Section 5.1 and, for simplicity, is only shown for a subset of hyperparameter scan of the DNN chosen inputs for W-boson and top-quark tagging. The performance is observed to depend on learning rate (update step size of layer parameters) and choice of regularizer (regularization term introduced to apply penalties on layer parameters in order to avoid overfitting). However, of particular note is that there is little change in the background rejection when increasing the number of hidden layers, thereby indicating that the use of a deep network is not warranted when combining high-level features such as those studied in this work.

DNNs are prone to overtraining, therefore it is necessary to test the robustness of the obtained DNNs. One way of doing this is to use an independent validation set during training. In this analysis, the DNNs

<table>
<thead>
<tr>
<th>W-Boson Tagging Observable Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observable</td>
</tr>
<tr>
<td>------------</td>
</tr>
<tr>
<td>$E C F_1$</td>
</tr>
<tr>
<td>$E C F_2$</td>
</tr>
<tr>
<td>$E C F_3$</td>
</tr>
<tr>
<td>$C_2$</td>
</tr>
<tr>
<td>$D_2$</td>
</tr>
<tr>
<td>$\tau_1$</td>
</tr>
<tr>
<td>$\tau_2$</td>
</tr>
<tr>
<td>$\tau_{21}$</td>
</tr>
<tr>
<td>$R_W^2$</td>
</tr>
<tr>
<td>$S$</td>
</tr>
<tr>
<td>$\mathcal{P}$</td>
</tr>
<tr>
<td>$\mathcal{D}$</td>
</tr>
<tr>
<td>$a_3$</td>
</tr>
<tr>
<td>$A$</td>
</tr>
<tr>
<td>$T_{\text{MIN}}$</td>
</tr>
<tr>
<td>$T_{\text{MAJ}}$</td>
</tr>
<tr>
<td>$Z_{\text{CUT}}$</td>
</tr>
<tr>
<td>$\mu_{12}$</td>
</tr>
<tr>
<td>$\sqrt{d_{12}}$</td>
</tr>
<tr>
<td>$K_{tDR}$</td>
</tr>
</tbody>
</table>

Table 6: Summary of the various sets of observables that go into each training of the DNN for W-boson tagging.

---

5 Although the activation function of the input and hidden layers is varied during the grid search, the activation function of the output layer chosen as sigmoid in shape.

6 The layer type is chosen to be dense (fully-connected layers), the batch size (number of jets considered before updating the DNN) is chosen to be 200, the activation function is chosen to be rectified linear units and the Glorot uniform function is chosen to initialize the DNN for the presented subset of hyperparameter scan.
is evaluated with flat spectra. Uncertainties are not presented.

Figure 3: Distributions showing the training with different set of variables and relative improvement in performance for the DNN for W-boson (left) and top-quark tagging (right) at the 50% relative signal efficiency working point. Only jets which pass the training criteria are considered while calculating the relative signal efficiency and relative background rejection, performance is evaluated with flat $p_T^{\text{true}}$ spectra. Uncertainties are not presented.

Figure 4: Relative background rejection at the 50% relative signal efficiency for DNNs trained with different learning rate (lr), regularizer (r) and number of hidden layers. All DNNs are trained on the group of variables determined to be optimal for the DNN W-boson (left) and top-quark (right) tagger as listed in Table 10. All DNNs use dense layers with batch normalization, Glorot uniform initialization and rectified linear unit activation function. Gray, blue, red, green bands separated by vertical lines correspond to DNNs with 3, 4, 5, 6 hidden layers, respectively for W-boson tagging. Gray, blue, red, green bands separated by vertical lines correspond to DNNs with 3, 4, 5, 6 hidden layers, respectively for top-quark tagging. If 50% signal efficiency working point could not be found due to the spikes in the discriminant distribution, the corresponding DNN bin is left empty. Only jets which pass the training criteria are considered when calculating the relative signal efficiency and relative background rejection. The performance is evaluated with flat $p_T^{\text{true}}$ spectra.
are updated during training to minimize the training loss defined by the cross-entropy value for binary classification problems. In order to control the overtraining while training, the loss is calculated in 2 different subsets of the general training set referred to as the training set and validation set in Figure 5. The training set is the set which the DNN uses to optimize the classifier whereas the validation set is an independent set on which the training loss is calculated after each epoch \(^7\). The training and validation set losses are compared as a function of the number of epochs in Figure 5. The loss of the validation set being less than or equal to the loss of the training set indicates that there is no overtraining. The loss of the validation set is found to be systematically less than the loss of the training set. This is expected as the training loss is calculated continuously during training whereas the loss of the validation set is calculated after each epoch and after the loss has been regularized.

\(^7\) Note that here what is considered as the training set for the DNN is a subset of the general training set (used throughout the rest of the note) that is used to minimize the loss. Thus, the general training set is the union of both (the training and the validation) sets.
### Table 9: Hyperparameter and architecture scan for the DNN grid search.

<table>
<thead>
<tr>
<th>Layer type</th>
<th>W-Boson Tagging</th>
<th>Top-Quark Tagging</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of hidden layers</td>
<td>3, 4, 5, 6</td>
<td>3, 4, 5, 6</td>
</tr>
<tr>
<td>Activation function</td>
<td>relu, tanh</td>
<td>relu, tanh</td>
</tr>
<tr>
<td>Learning rate</td>
<td>$10^{-3}$, $10^{-4}$, $10^{-3}$</td>
<td>$10^{-5}$, $10^{-4}$, $10^{-3}$</td>
</tr>
<tr>
<td>L1 Regularizer</td>
<td>$10^{-3}$, $10^{-2}$</td>
<td>$10^{-3}$, $10^{-2}$</td>
</tr>
<tr>
<td>NN weight initialization</td>
<td>Glorot uniform, He normal</td>
<td>Glorot uniform, He normal</td>
</tr>
<tr>
<td>Batch size</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>Batch normalization</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Number of Maxout layers</td>
<td>-</td>
<td>5, 10, 15, 20, 25</td>
</tr>
<tr>
<td>Training groups</td>
<td>7 groups</td>
<td>7 groups</td>
</tr>
</tbody>
</table>

Figure 5: Training and validation loss distributions for $W$-boson tagging (left) and top-quark tagging (right) shows the robustness of DNNs with respect to overtraining.
### 6 Results

The BDT and the DNN algorithms result in a single discriminant that allows for the classification of a jet as either a top-quark or gluon/other (non-top) quark jet and a W-boson or gluon/other (non-top) quark jet. The distributions of the discriminants are shown in Figure 6.

Furthermore, the performance of the two new taggers is studied more quantitatively in two ways. The discrimination power of the trained BDTs and DNNS are compared with respect to the simple reference taggers to estimate the gain expected by adding more variables and taking advantage of non-linear correlations. Second, the performance of BDTs and DNNs are compared with each other to determine if one algorithm is able to extract more information than the other.

In order to compare the performance of the two algorithms using the same set of input variables, the BDT taggers are trained using the variables found optimal for the DNN taggers (DNN Obs., as listed in Table 10), and similarly the DNN taggers are trained using the optimal BDT variable set (BDT Obs.). This comparison shows whether the two algorithms exploit different information found in the variable correlations. For this comparison, the hyperparameter optimization is not redone for the BDT as it is found to not change the performance significantly. However, in the case of the DNN it is found to be necessary

<table>
<thead>
<tr>
<th>Observable</th>
<th>W-Boson Tagging</th>
<th>Top-Quark Tagging</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ECF_1$</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>$ECF_2$</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>$ECF_3$</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>$C_2$</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>$D_2$</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>$\tau_1$</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>$\tau_2$</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>$\tau_3$</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>$\tau_{21}$</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>$\tau_{32}$</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>$R^FW_2$</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>$S$</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>$P$</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>$D$</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>$a_3$</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>$A$</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>$T_{MIN}$</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>$T_{MAJ}$</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>$Z_{CUT}$</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>$\mu_{12}$</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>$\sqrt{d_{12}}$</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>$\sqrt{d_{23}}$</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>$K_{1DR}$</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>$Q_w$</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>

Table 10: Summary of the set of observables that are chosen with respect to Figure 1.3.
to retune the hyperparameters each time the set of input observables is changed and so this optimization is performed again for the set of observables coming from the BDT optimization. The taggers whose performance are compared consist of a varying $p_T^{true}$-dependent cut on the relevant discriminant (BDT, DNN, substructure variable) and a fixed cut on the jet mass. The substructure observables used in the simple reference taggers are chosen as $D_2$ and $\tau_{32}$ for $W$-boson and top-quark tagging, respectively and the fixed mass cuts are defined by a mass window of $[60,100]$ GeV for $W$-boson and a lower mass bound of 120 GeV for top-quark tagging.

Figure 6: BDT $W$-boson tagger (top left) and DNN $W$-boson tagger (top right), BDT top-quark tagger (bottom left), DNN top-quark tagger (bottom right) score distributions for jets with $1000 < p_T^{true} < 1500$ GeV and $m^{calo} > 40$ GeV with one entry per jet and weighted according to the testing event weights described in Section 4.1. Each multivariate tagger is trained with variables optimized by its own method.
6.1 Performance Evaluation

The performance of the taggers is evaluated in two wide jet $p_T^{\text{true}}$ bins in terms of receiver operating characteristic (ROC) curves, as shown in Figures 7 and 8 for the $W$-boson and top-quark taggers, respectively. Furthermore, the background rejection at the fixed 50% signal efficiency working point as a function of $p_T$ are presented in Figures 7 and 8. The simple reference taggers consisting of the combined selections $D_2 + m_{\text{calo}}$ (for $W$ bosons) and $\tau_{32} + m_{\text{calo}}$ (for top quarks) presented here are less optimal than the smoothed working point taggers described in Reference [3] and Reference [4] due to the manner in which the selections on the jet mass and auxiliary tagging observable are defined. In those studies, the selections are chosen to provide the maximum background rejection for a fixed signal efficiency. However, in this study, the selection on jet mass is a fixed to facilitate a more direct comparison of the efficacy of the auxiliary tagging observable when tagging at a fixed signal efficiency. Therefore, this comparison focuses on the relative gain compared with the standard analysis approach to tag a $W$ boson or a top quark: a cut on the jet mass and a substructure variable cut. By retaining the same jet mass cut, the same set of signal-enriched events is used when comparing the performance of the multivariate discriminant to the best single substructure variable. The performance of the BDT and DNN taggers are found to be very similar, which suggests that neither BDT or DNN is learning additional information compared with the other using the same set of input features.

6.2 Discriminant Correlations

It is challenging to intuitively understand what information is extracted by BDTs and DNNs. Although it is expected that these techniques extract non-linear information, quantifying the linear correlations can be useful to study what the algorithms have learned. The $W$-boson and top-quark tagging scores for different signals and QCD background samples are shown on a two dimensional plane in Figure 9 for each classifier. Unlike for the previous performance distributions, top-quark jets which are not fully contained, specifically those that contain the $b$-quark and one of the light quarks from the $W$-boson decay and referred to as $q_{b}$ jets here, are included in this figure as a separate class. For this study the event weights are calculated by taking these jets into account. As the training has been done on only fully contained top quarks and $W$ bosons, each signal and background sample is well separated from each other, as expected. This shows that $W$-boson and top-quark taggers can be used not only to reject QCD jets, but also to discriminate between $W$-boson and top-quark jets.

The effects of the multivariate classifier and substructure requirements on the background jet samples’ mass distribution are presented in Figure 10. The multivariate taggers modify the background jet mass distribution such that it more resemble that of the signal, a feature that may be undesirable if this variable is used to discriminate signal from background contributions. This effect is expected due to usage of training inputs which are highly correlated with mass, as can be seen in Figure 11 which shows the correlation coefficients for input variables and multivariate outputs. The correlation coefficients are beneficial for understanding which variables are most correlated with the multivariate output and how the two multivariate techniques differ. Both multivariate techniques show similar correlations to the input variables, and the BDT and DNN outputs are extremely correlated. It is found that for $W$-boson and top-quark jets, the jet mass is correlated with most of the substructure variables, whereas it is strongly anti-correlated with multivariate $W$-boson classifiers.
Figure 7: Distributions showing comparison of the BDT and DNN taggers performance to the simple W-boson tagger based on a simple fixed cut on the mass and a selection on the $D_2$ observable in a low-$p_T^{\text{true}}$ (top left) and high-$p_T^{\text{true}}$ (top right) bin. The background rejection for a fixed 50% signal efficiency working point is also presented (bottom), where a selection defined by a requirement on the jet mass is followed by a $D_2$, BDT or DNN requirement optimised as a function of the jet $p_T^{\text{true}}$. 

- **DNN with DNN Obs.**: DNN trained with W-boson tagging DNN inputs,
- **DNN with BDT Obs.**: DNN trained with W-boson tagging BDT inputs,
- **BDT with BDT Obs.**: BDT trained with W-boson tagging BDT inputs,
- **BDT with DNN Obs.**: BDT trained with W-boson tagging DNN inputs. BDT and DNN inputs are summarized in Table 10.
Figure 8: Distributions showing comparison of the BDT and DNN performance to the simple top-quark tagger based on a simple fixed cut on the mass and a selection on $\tau_{32}$ observable in a low-$p_T^{true}$ (left) and high-$p_T^{true}$ (right) bin. The background rejection for a fixed 50% signal efficiency working point is also presented (bottom), where a selection defined by a requirement on the jet mass is followed by a $\tau_{32}$, BDT or DNN requirement optimised as a function of the jet $p_T^{true}$. \textit{DNN with DNN Obs.}: DNN trained with top-quark tagging DNN inputs, \textit{DNN with BDT Obs.}: DNN trained with top-quark tagging BDT inputs, \textit{BDT with BDT Obs.}: BDT trained with top-quark tagging BDT inputs, \textit{BDT with DNN Obs.}: BDT trained with top-quark tagging DNN inputs. BDT and DNN inputs are summarized in Table 10.
Figure 9: Two dimensional distribution of the $W$-boson tagger vs. the top-quark tagger scores for the BDT (left) and DNN (right) for jets with $1000 < p_T^{\text{true}} < 1500$ GeV and $m_{\text{calo}} > 40$ GeV with one entry per jet and weighted according to the testing event weights described in Section 4.1. $q\bar{b}$ jets represent the top-quark jets which are not fully contained and missing one of the light quarks from the $W$-boson decay. Each multivariate tagger is trained with variables optimized by their own method.

Figure 10: Jet mass distributions showing the signal before selection and the background both before selection, with a selection only on $D_2$ ($W$-boson tagging) or $\tau_{32}$ (top-quark tagging), with a selection on the BDT discriminant, and the DNN discriminant in the case of $W$-boson tagging (left) and top-quark tagging (right) for jets with $1000 < p_T^{\text{true}} < 1500$ GeV.
Figure 11: The four linear correlation coefficient matrices for QCD (upper left), $W \to qq$ (upper right), $top \to qqb$ (bottom left), and $top \to qb$ (bottom right) jets, which represent the not fully contained top-quark jets and missing one of the light quarks from the $W$ decay, with training weights described in Section 4.2. Each multivariate tagger is trained with variables optimized by their own method. The entries where 0 is reported are correlation coefficients less than 0.5%.
7 Conclusion

In this note a preliminary study of the use of BDTs and DNNs in ATLAS as applied to the identification of hadronically-decaying $W$-bosons or top-quarks is provided. Input variables and hyperparameters of the multivariate algorithms are optimized and trained with MC simulation. Simple reference taggers which use a selection using a fixed jet mass requirement and cut on second jet substructure variable optimised as a function of the jet $p_T^{\text{true}}$ are used as a reference for comparisons to the BDTs and DNNs. The performance of the BDT and DNN taggers is rather similar. However, due to a combination of variables and the ability to exploit non-linear correlations, the BDTs and DNNs outperform the respective simple reference tagger. Given that variables which are strongly correlated to the jet mass are used in the training, the application of the machine learning discriminant leads to a strong shaping of the jet mass distribution. Investigations into data modelling and systematic uncertainties are outside the scope of this note.
References


[34] Jonathan M Butterworth et al.,
Jet Substructure as a New Higgs-Search Channel at the Large Hadron Collider,

[35] ATLAS Collaboration,
Measurement of kT splitting scales in W → lν events at √s TeV with the ATLAS detector,


[37] Anson Hook et.al., Jet Dipolarity: Top Tagging with Color Flow, JHEP 04 (2012) 007,
arXiv: 1102.1012 [hep-ph].

[38] A. Collaboration, ATLAS measurements of the properties of jets for boosted particle searches,
Phys. Rev. D 86 (7 2012) 072006,

[39] S Catani et al., Longitudinally-invariant k⊥-clustering algorithms for hadron-hadron collisions,


[41] I. Goodfellow, Y. Bengio and A. Courville, Deep Learning,


[44] X. Glorot and Y. Bengio,
‘Understanding the difficulty of training deep feedforward neural networks’,
Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics,
AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010, 249,

[45] S. Ioffe and C. Szegedy,
Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,