Class Imbalance Techniques for High Energy Physics

Christopher W. Murphy[1*]

1 Insight Data Science, San Francisco, CA 94107, USA
* chrismurphybnl@gmail.com

November 24, 2019

Abstract

A common problem in a high energy physics experiment is extracting a signal from a much larger background. Posed as a classification task, there is said to be an imbalance in the number of samples belonging to the signal class versus the number of samples from the background class. In this work we provide a brief overview of class imbalance techniques in a high energy physics setting. Two case studies are presented: (1) the measurement of the longitudinal polarization fraction in same-sign WW scattering, and (2) the decay of the Higgs boson to charm-quark pairs.

Contents

1 Overview

The Large Hadron Collider (LHC) has been an incredibly successful experiment. To date it has discovered the Higgs boson, and measured hundreds, if not thousands, of other processes to be consistent with the predictions of the Standard Model (SM) [1]. A common problem in making these measurements is extracting a signal from a much larger background. Occasionally in this situation there is a single feature that is powerful enough to discriminate the signal from the large background. An example of this the Higgs boson decaying to two photons where the invariant mass of the photon pair is the discriminating observable [2, 3]. More often however a multi-variate analysis of many features needs to be performed. Machine learning (ML) and deep learning (DL) are well suited for such tasks. Therefore it is not surprising that ML and DL have become, and will likely continue to be, an important part of the success of the LHC program. See Refs. [4, 5, 6, 7] for some recent reviews.

If one treats the extraction of a signal from a much larger background as a classification problem there is an imbalance in the number of sample belonging to the signal class versus the number of events from the background class. In the machine learning community techniques for learning from imbalanced data are well established. There is now even a software package, $\texttt{imbalanced}-\texttt{learn}$ [8], dedicated to this task. In high energy physics there do not appear to be many cases where imbalanced learning techniques were explicitly used. However the measurement of the time-integrated CP asymmetry in $D^0 \to K^0_S K^0_S$ decays by LHCb [9] is one such example. In particular LHCb classified the D^0 decay signal from its background using the analysis methods developed in Refs. [10, 11]. An alternative approach to classification with imbalance techniques is using an anomaly detection framework. There are several examples of this in high energy physics [12, 13, 14, 15].

Given the lack of examples where imbalanced learning techniques were used in high energy physics, the purpose of this note is two-fold. Firstly, in Section 2, we aim to provide a brief overview of modern class imbalance techniques in a high energy physics setting, introducing novel loss functions and a data resampling technique. Secondly, we provide two case studies of how class imbalance techniques can be used in high energy physics settings. The first case, presented in Sec. 3, is the measurement of the longitudinal polarization fraction in same-sign WW scattering. We find a modest improvement in the performance of both the classical machine learning models and the deep learning models used in the longitudinal WW study. The second study is the decay of the Higgs boson to charm-quark pairs, which follows in Sec. 4. Our Higgs-to-charm tagger gives a 14% improvement in the background rejection rate. Another application of these techniques is training directly on experimental data [16, 17, 18]. Conclusions are then given in Sec. 5. Much of the code for this project is available at [19].

2 Class Imbalance Techniques

There is no definitive answer to the question: What should one do when dealing with imbalanced data? The answer will depend on the data in question, see [20] for a study of benchmark datasets. In this Section we present a few approaches one might try to improve performance on an unbalanced dataset.

Using the accuracy of a classifier as a metric can be misleading. (See Table. 3 for a glossary of model evaluation terms used in this work.) Consider a model that predicts that every sample to be background. The accuracy of this model is $A = 1 - r$, where r is the ratio of the number of signal events to the total number of events. Although this model would be highly accurate if the data were sufficiently imbalanced, it would not be useful as it says nothing about the signal, which is what we were interested in to begin with. For this reason accuracy is not a recommended metric in this setting. The ROC curve is a good general purpose metric, providing information about the true and false positive rates across a range of thresholds, and the area under the ROC curve (AUC) is a good general purpose, single number metric. However, when dealing with imbalanced data, we argue in what follows that the precision-recall curve is the preferred metric to use on imbalanced data. If one instead prefers a single number metric, average precision is approximately the area under the precision-recall curve in analogy with AUC for the ROC curve.

The ROC curve describes the false positive (background rejection) rate as a function of the true positive rate (signal efficiency), whereas the precision-recall curve, true to its name, gives precision as a function of recall. Recall is equivalent to the true positive rate, but precision does not correspond to the false positive rate. Recall or the true positive rate is a measure of how many true signal events have actually been identified as signal. Similarly the false positive rate is a measure of how many of the true background events have been identified as background. Precision, on the other hand, quantifies how likely an event is to truly be signal when a classifier has predicted it to be signal. A classifier's prediction will vary as the baseline probability of the positive class varies. As such, precision depends on how rare the signal is. This motivates using the precision-recall curve when the positive class samples are rare compared to the negative class examples. When this is not an issue the ROC curve is the metric to use as it does not care about the baseline probability of the positive class.

One might also try to balance the training set either by under-sampling [21, 22, 23, 24, 25] the majority class, oversampling the minority class [26, 27], or a combination of over- and under-sampling [28, 29]. Oversampling runs the risk of overfitting, and training with oversampling takes longer because of the additional data. For these reasons we will focus on under-sampling in this work. In particular, we will use random under-sampling to create a balanced random forest [30, 31]. Analogous procedures exist for creating a balanced boosted decision trees [32] and making balanced batches to feed into a neural network. The algorithm for how the balanced random forest makes classifications is as follows: (1) take bootstrap samples from the original dataset, (2) balance each sample by downsampling randomly, (3) learn a decision tree from each sample, (4) make predictions based on a majority vote. It is the second step of this process that is absent in a standard random forest. Even if this does not lead to a gain in performance training is faster with this approach because less data is used.

Lastly, one might consider making changes to the algorithms being used [33, 34, 35]. A simple example of this is if a metric such as precision, recall, or F_1 score is being used, its decision threshold can be optimized to maximize performance. One approach along these lines is to add hyperparameters to the loss function, creating a relatively larger penalty for misclassifying an example. To start consider the standard cross entropy loss function used for binary classification

$$BCE = -y\log(p) - (1-y)\log(1-p), \tag{1}$$

where y is the ground-truth class with $y = 1$ for the signal class, and p is the model's estimated probability that a given event belong to the signal class. Following Ref. [36] we introduce the following compact notation[1]

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise} \end{cases} \tag{2}$$

With this definition Eq. (1) becomes

$$BCE = -\log(p_t). \tag{3}$$

When there is class imbalance it is common to add a weighting hyperparameter, α, to loss function. Weighting the loss function can be implemented as follows

$$CE = -\alpha\, y \log(p) - (1-\alpha)(1-y)\log(1-p) \equiv -\alpha_t \log(p_t), \tag{4}$$

where α_t is defined analogously to p_t in Eq. (2). With this normalization α takes values between 0 and 1. Often α is taken to be proportional to the inverse class frequency, $\alpha \propto r^{-1}$. The weighting hyperparameter balances the importance of signal and background events in the loss function. However α does not do anything to differentiate between easy- and hard-to-classify examples. In particular, easy-to-classify background examples may come to overwhelm the loss function even though they are individually negligible if the class imbalance is extreme enough.

This issue was rectified in Ref. [36], which introduced the focal loss function

$$FL = -(1-p_t)^\gamma \log(p_t), \tag{5}$$

where the modulating parameter, γ, puts the focus on hard-to-classify examples. In particular, when a sample is misclassified and p_t is small the modulating factor is approximately one, and the loss is unaffected. However, as p_t approaches one the modulating factor approaches zero, down-weighting the loss function for well-classified examples. When $\gamma = 0$ focal loss is equivalent to cross entropy, and as γ is increased the rate at which easy-to-classify samples are down weighted also increases. Focal Loss is an optimal classifier just as cross entropy or mean square error are. One way to see this is Focal Loss produces a concave ROC curve (given sufficiently large statistics), which is equivalent to being optimized by the likelihood ratio [37].

In this work we will use the weighted variation of focal loss

$$FL = -\alpha_t(1-p_t)^\gamma \log(p_t), \tag{6}$$

with default values for the hyperparameters, $\alpha = 0.25$ and $\gamma = 2$.

Another generalization of focal loss is from binary classification to multi-class classification. Here the compact p_t notation does not work, so to set the stage we define the categorical cross entropy loss for classification with K classes

$$CCE = -\sum_{i=1}^{K} y_i \log(p_i). \tag{7}$$

where p_i the probability that an example belongs to class i, and is given by the softmax function

$$p_i = \frac{e^{s_i}}{\sum_{j=1}^{K} e^{s_j}}. \tag{8}$$

[1] A model's estimated probability of an event belong to a class, p_t, is not to be confused with the transverse momentum of a particle, p_T.

with s_i being the score for the ith class for an example. The vector y is a one-hot representation of the classes with one component equal to one and the remain $K-1$ components equal to zero. When $K=2$ Eq. (8) reduces to Eq. (3). With all of the setup in place, we can now write the categorical focal loss for multi-class classification

$$CFL = -\sum_{i=1}^{K} y_i (1-p_i)^{\gamma} \log(p_i). \tag{9}$$

3 Longitudinal Polarization Fraction in Same-Sign WW Production

3.1 Introduction

Same-sign WW production at the LHC is the vector boson scattering (VBS) process with the largest ratio of electroweak-to-QCD production. As such it provides a great opportunity to study whether the discovered Higgs boson leads to unitary longitudinal VBS, and to search for physics beyond the SM (BSM) [38, 39]. The ATLAS and CMS experiments have observed electroweak same-sign WW production in the two jet, two same-sign lepton final state in 13 TeV pp collisions with significances of 6.9σ [40] and 5.5σ [41], respectively. Confirming or refuting the unitarity of VBS requires not just a measurement of $pp \to jjW^{\pm}W^{\pm}$, but of the fraction of these events where both Ws are longitudinally polarized (LL fraction).

Prospects for the extraction of the longitudinal component of $W^{\pm}W^{\pm}$ scattering during the High-Luminosity phase of the LHC (HL-LHC) were studied in Refs. [42, 43, 44]. The fraction of longitudinally polarized events is predicted to be only $r \sim 0.07$ in the SM at large dijet invariant mass (m_{jj}) [43] making this a challenging measurement. Using the difference in the azimuthal angle of the two jets

$$\Delta\phi_{jj} = \min(|\phi_{j_1} - \phi_{j_2}|, 2\pi - |\phi_{j_1} - \phi_{j_2}|), \tag{10}$$

as a discriminant, the significance for the observation of the LL fraction is expected to be up to 2.7σ with 3000 fb^{-1} of integrated luminosity [43].

The observation significance can be improved through the use of deep learning [45, 46]. Ref. [45] regressed on the angles between the charged leptons in their parent boson's rest frame and the W boson's direction of motion, whereas Ref. [46] treated this as a binary classification problem distinguishing between events where both Ws were longitudinally polarized versus when one or none of the Ws were polarized. In the classification setting it is important to keep in mind that the predicted LL fraction is small, and thus there is an imbalance in the number of events belonging to the class $N(W_L) = 2$ versus the class $N(W_L) < 2$ (LL class vs. $TL + TT$ class). We proceed treating this as a classification problem with imbalanced classes.

3.2 Data

`MadGraph5` v2.6.6 [47] is used to simulate events for the leading order electroweak, $\mathcal{O}(\alpha^4)$, contribution to process $pp \to jjW^{\pm}W^{\pm}$ at center of mass energy $\sqrt{s} = 14$ TeV. The fraction of events where both Ws are longitudinally polarization is $r \approx 7.5\%$. Additionally, `MadSpin` [48] is used to include spin correlation effects in the decays of the W bosons such that the final process under consideration is $pp \to jj\ell^{\pm}\nu\ell^{\pm}\nu$ with $\ell = \{e, \mu\}$. Representative Feynman diagrams are given in Figure 1. Note that in this case study, unlike the one that follows it, the "jets" are

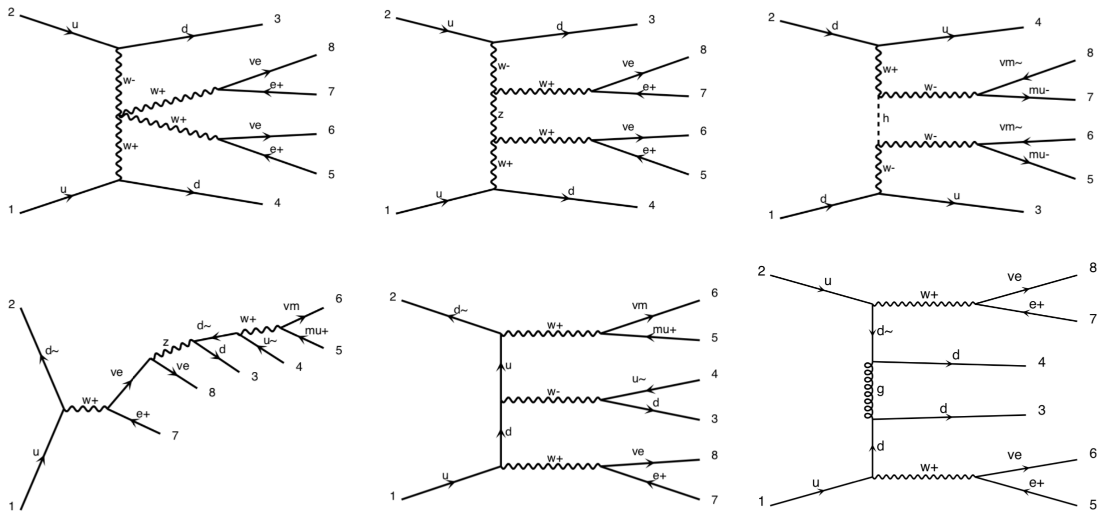


Figure 1: Representative leading order Feynman diagrams for $pp \to jj\ell^{\pm}\nu\ell^{\pm}\nu$. Top row: Diagrams contributing to the signal, $pp \to jjW^{\pm}W^{\pm} \to jj\ell^{\pm}\nu\ell^{\pm}\nu$ with $\sigma \propto \alpha^6$. Bottom row: Diagrams considered irreducible background in this work. The diagrams are drawn with `MadGraph5` [47].

partons from the hard scattering process and are not showered or hadronized. We comment on the impact this choice has in the results subsection of this case study. The cuts are chosen to match those of Ref. [46]. We require two jets with transverse momentum, $p_T > 50$ GeV, and pseudorapidity, $|\eta| < 4.7$. The jet pair must also have an absolute difference in pseudorapidity $\Delta\eta_{jj} > 2.5$, consistent with VBS, and have an invariant mass $m_{jj} > 850$ GeV to suppress non-prompt and WZ backgrounds [41]. Additionally we select for two same-sign charged leptons with $p_T > 20$ GeV and $|\eta| < 2.4$. A total of approximately $1.7 \cdot 10^5$ events pass these cuts.

The feature engineering is also done to match that of Ref. [46] as much as possible. The p_T, η, and ϕ of the two jets and the two leptons are used as features. The subscripts 1 and 2 are used to indicate the jet or lepton with the larger or smaller transverse momentum, *e.g.* $p_T^{j_1} > p_T^{j_2}$. This step that improves the performance of classifiers, and is not done by default in `MadGraph5`. The magnitude and azimuthal angle of the missing transverse energy are included as well. In addition, the following high-level features are added. From the jet system we add the invariant mass, the difference in pseudorapidity, and the difference in the azimuthal angle. We also consider the Zeppenfeld variable [49] for the two charged leptons,

$$z_{\ell_i} = \frac{\eta_{\ell_i} - \bar{\eta}_{jj}}{\Delta\eta_{jj}}, \tag{11}$$

where $\bar{\eta}_{jj}$ is the mean pseudorapidity of the two leading jets. Finally we include the separation of the di-jet and di-lepton systems in the pseudorapidity-azimuthal angle plane, $\Delta R_{jj,\ell\ell}$, bringing the total number of features to 20.

3.3 Models and Training

In addition to using $\Delta\phi_{jj}$ and $p_T^{\ell_1}$ as discriminating observables, we use the following models. For classical machine learning we use a random forest (RF) as a baseline, and look to use a change in performance from weighting or balancing. We use the `imbalanced − learn` [8] implementation of balanced random forest, and use `scikit − learn` [50] for the other random forests. The balanced random forest has no maximum depth, while the other random forests have a maximum depth of 10. Additionally we consider a `LightGBM` [51] (LGBM), which is a gradient boosted decision tree where the trees are grown in a depth first rather than breadth first fashion. The name Light comes from the fact that the training time is often greatly reduced with this construction of the trees. In particular, our LGBM has 10^3 estimators and a learning rate of 0.01. The deep learning models are fully-connected neural networks (DNNs) implemented using the `Keras` API [52] for `TensorFlow` v2.0.0 [53]. Our baseline DNN has a cross entropy loss function, Eq. (3), and the variation we test is a DNN with a focal loss function, Eq. (6). The features are scaled to have zero mean and unit variance before being fed into the neural networks. All of our neural networks have 2 hidden layers each with 150 neurons, He initialization, and ReLU activation functions. Batch normalization is performed to speed up the learning process, dropout is applied at a 50% rate for regularization, and the Adam algorithm is used to optimize the parameters of the DNN.

A five-fold cross validation is performed for each for model. The folds are stratified based on the size of the class imbalance. For the DNNs, a batch size of 50 is used in training. Early stopping is implemented for the DNNs where training runs until there is no decrease in the training loss function for 5 consecutive epochs. Similarly, we grow the Random Forests 10 trees at a time until there is no improvement in the training loss function.

3.4 Results

Table 1 shows the results of the cross validation with performance being reported as the mean ± the standard deviation of the five folds. Both the weighted random forest and the balanced random forest modestly outperform the baseline random forest. Similarly, the DNN with focal loss modestly outperforms its baseline neural network. The uncertainty on the machine learning metrics is statistical in nature; one over the square root of the sample size of a test fold in the cross validation is approximately $5.4 \cdot 10^{-3}$. On the other hand, the uncertainty on the time it takes to fit the models, t_{fit}, does not follow this statistical pattern due to the stochastic nature of the optimization process and the early stopping criteria imposed on training.

The improvement in performance of the balanced RF can be seen visually in Figure 2 where the green curves of the standard random forest are below the red curves of the balanced random forest both precision versus recall (left panel) and the ROC curve (right panel). More strikingly, all of the machine learning models significantly outperform the kinematic variable $p_T^{\ell_1}$. Note that recall is equivalent to signal efficiency, but precision is not related to background rejection.

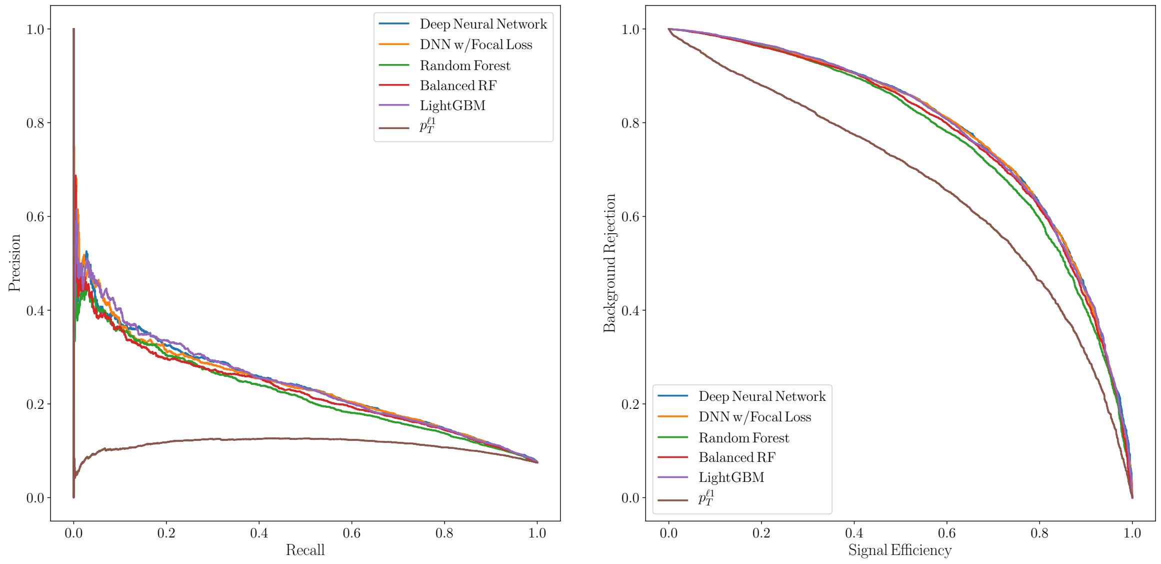
The balanced and weighted random forests also take less time to train. In the case of the balanced RF, t_{fit} does not tell the whole story as it has no maximum depth whereas the standard random forest can only be 10 levels deep. Not to be outdone, the LGBM fits more than an order of magnitude faster than the neural networks and almost an order of magnitude faster than the standard random forest. Its performance is intermediate between the balanced random forest and the baseline the neural network.

Histograms for the probability the event will be predicted to be an LL event are shown in Figure 3 when it is in truth an LL event (red distributions) or when it is actually an $TL + TT$ event (blue distributions). The top row shows the random forest models, and the bottom row shows the DNN models.

The mean predicted probability for a classifier with an unweighted loss function trained on

Model	t_{fit} [s]	Average Precision	AUC
$\Delta\phi_{jj}$	-	0.120 ± 0.003	0.662 ± 0.006
$p_T^{\ell_1}$	-	0.112 ± 0.003	0.663 ± 0.006
Random Forest	84 ± 24	0.223 ± 0.006	0.766 ± 0.006
Weighted RF	30 ± 15	0.227 ± 0.006	0.768 ± 0.006
Balanced RF	63 ± 19	0.228 ± 0.007	0.776 ± 0.005
LightGBM	9.7 ± 0.7	0.241 ± 0.005	0.782 ± 0.005
Deep Neural Network	$(2.8 \pm 0.3) \cdot 10^2$	0.244 ± 0.008	0.789 ± 0.004
DNN w/ Focal Loss	$(3.3 \pm 1.1) \cdot 10^2$	0.246 ± 0.004	0.791 ± 0.005

Table 1: Results of the five-fold cross validation for classifying LL events from TL+TT events in $pp \to jjW^{\pm}W^{\pm} \to jj\ell^{\pm}\nu\ell^{\pm}\nu$. Performance is reported as (the mean ± the standard deviation) of the five folds. t_{fit} is the time it takes to fit the model to a training fold of data. The models utilizing class imbalance techniques show modest improvements in performance with respect to their baselines. See the text more for details.



(a) Precision-Recall Curve

(b) ROC Curve

Figure 2: Multiple parameter performance measures. The precision-recall curve is given in 2a, and the ROC curve is shown in 2b. Visually it is clear that the balanced random forest (red) outperforms its unbalanced counterpart (green). More strikingly, all of the machine learning models significantly outperform the kinematic variable $p_T^{\ell_1}$. Note that recall is equivalent to signal efficiency, but precision is not related to background rejection.

an imbalanced dataset is r, the imbalance ratio. Complete signal-background separation in the training dataset is a sign of overfitting if such behavior is not also observed in the validation dataset, which it's not in this case. Balancing the training set moves the mean value from r to 0.5. This can be seen in the upper right panel of 3 from the balanced random forest. Weighting the loss function with the inverse of the class frequencies also moves the mean value to 0.5. Focal loss is intermediate between these two scenarios, r and 0.5, as can be seen in the bottom right panel of 3.

Finally, this case study would not be complete without a comparison with to Ref. [46]. The most obvious difference between our work and that of [46] is the better performance we find from the kinematic variable $\Delta\phi_{jj}$. However we did not pass our simulated events through a parton shower or hadronize them, which likely would have spoiled some of the correlation between $\Delta\phi_{jj}$ and the polarizations of the W bosons. Beyond that, our results are consistent with those found in Ref. [46]. Specifically, as measured by the AUC, our fully-connected neural network with two hidden layers matches the performance of the neural network with "particle-based" architecture and 10 hidden layers in [46]. Additionally, our balanced random forest matches the performance of the AdaBoost classifier of Ref. [46], where again performance is measured by the AUC. We do not estimate the statistical significance of a non-zero LL fraction from our classifiers for two reasons. Firstly the imbalance ratio r is higher in our simulated dataset than that of Ref. [46], which would make our models appear to significantly outperform those of [46] when based on the comparison of machine learning metrics given above the differences are not so great. Secondly all the machine learning models significantly outperform the kinematic variable $p_T^{\ell_1}$, as can be seen in Fig. 2, so it's safe to assume all of the models tested here would produce a significance similar to 5σ given that the neural network in [46] was able to do so.

4 Higgs Boson Decays to Charm-Quark Pairs

4.1 Introduction

The second application of class imbalance techniques we explore in this note is to the measurement of Higgs boson decays to charm-quark pairs. Searches for the decay of the Higgs boson to charm-quarks have produced only weak limits to date. ATLAS reported an upper limit of 110 times the SM rate for the process $pp \to Zh \to \ell^-\ell^+c\bar{c}$ [54]. LHCb instead considered the associated production of both Ws and Zs in range $2 < \eta < 5$, and set a limit of 6,400 times the SM rate [55]. A result of these weak limits is that direct limits on the charm Yukawa coupling are correspondingly weak. Stronger bounds can be obtained indirectly, *e.g.* through global fits [56, 57, 58, 59, 60, 61, 62, 63, 64], among other methods.[2] However there are assumptions build into any indirect analysis. The limit on the charm Yukawa coupling at HL-LHC is projected to get down to about 2.2 times the SM rate [63] (see also [66]). Based on this projection an observation of $h \to c\bar{c}$ is not expected at HL-LHC motivating ways to improve the analysis, although this projected limit should still be useful in constraining certain BSM physics.

One reason for the weak limits on $h \to c\bar{c}$ is in the SM the rate for $h \to b\bar{b}$ is about 20 times larger ($r \approx 0.05$) than the rate for $h \to c\bar{c}$ [67]. In contrast with $h \to c\bar{c}$, the decay of the Higgs boson to bottom-quarks has been observed by both ATLAS [68] and CMS [69] The analyses of Refs. [54, 55, 68, 69] rely on tagging the flavor of the jets, which involves discriminating charm

[2] The special role in global fits of the Higgs boson coupling to charm-quarks has been known for a long time [65].

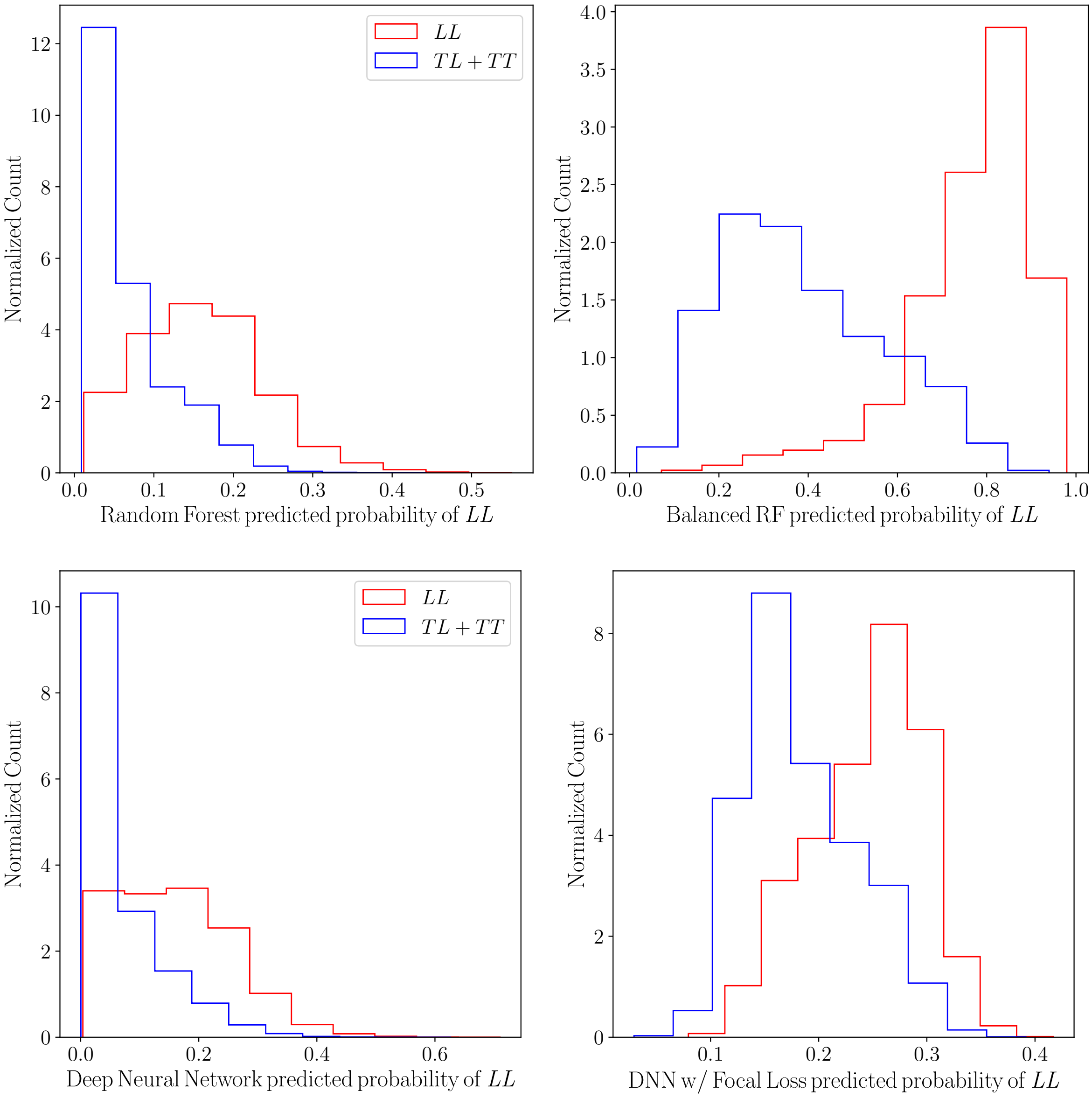


Figure 3: Histograms for the probability the event will be predicted to be an LL event when it is in truth an LL event (red distributions) or when it is actually an $TL + TT$ event (blue distributions). The top row shows the random forest models, and the bottom row shows the DNN models.

initiated jets from bottom jets, or vice versa, and discriminating heavy from light flavored jets.[3] The use of flavor tagging explicitly links the measurements of $h \to b\bar{b}$ and $h \to c\bar{c}$ [71, 72].

To perform the flavor tagging LHCb used their standard, state-of-the-art heavy flavor tagger [73], while ATLAS trained boosted decision trees to separate charm from light jets and charm from bottom jets with a procedure analogous to how they train their standard bottom tagger [74, 75]. The use of general purpose flavor tagging algorithms is less then ideal for the specific task of identifying Higgs decays to charms. This was recognized in Ref. [76], which made a dedicated double-charm tagger for $h \to c\bar{c}$. We also advocate making a dedicated $h \to c\bar{c}$ tagger for the following reason. The standard heavy flavor tagging algorithms are not optimized for the imbalance in the expected number of $h \to c\bar{c}$ versus $h \to b\bar{b}$ events. For example, QCD produces roughly equal numbers of bottoms and charms at invariant masses relevant for Higgs physics. Given the statistical nature of heavy flavor tagging, an imbalance in the number of $b\bar{b}$ and $c\bar{c}$ decays will lead to worse performance in identifying the Higgs to charm events. As such this is a well motivated arena for applying class imbalance techniques. Here we are assuming a SM-like rate for $h \to c\bar{c}$. If some BSM physics makes the experimental rate for $h \to c\bar{c}$ much larger than expected this would invalidate our argument (which would be a small price to pay for the discovery of the breakdown of the SM). The rest of this case study delivers proof of principle that it is possible to improve tagging efficiency of $h \to c\bar{c}$ events through the use of the class imbalance techniques.

Looking beyond the proof of principle, a few additional steps to be taken in future work are described in what follows. We are treating this as a binary classification problem of distinguishing Higgs boson decays to charm-quark pairs from bottom-quark pairs. Firstly, extending our approach to also discriminate heavy flavor jets from light flavor jets will make our tagger more like what the experiments are currently doing. A second opportunity area stems from our study of charm-tagging at a lepton collider where experimental tagging might not be based on jets, while it's clear that at hadron colliders jet based analyses are and will continue to be used. Lastly, a direct comparison with the results Ref. [76] is not currently possible given the different background considered in the two works. It would be useful to do a proper comparison of the two tagging methods.

4.2 Data

We consider associated Higgs production at an e^+e^- collider as an observation of $h \to c\bar{c}$ is not expected at HL-LHC. Specifically, the process under consideration is $e^+e^- \to Zh \to \ell^+\ell^- Q\bar{Q}$ with $\ell = e$ and μ, and $Q = b$ or c. A total of $2 \cdot 10^5$ events are simulated with `MadGraph5` [47] with `Pythia6` [77] used for parton showering and hadronization. Half the simulated events are $h \to b\bar{b}$ and the other half are $h \to c\bar{c}$. We focus on the binary classification problem of $h \to c\bar{c}$ versus $h \to b\bar{b}$ as existing tagging algorithms perform well at distinguishing heavy from light flavors, see *e.g.* [73]. The center-of-mass energy of the collisions is $\sqrt{s} = 250$ GeV. Jets are clustered using the `FastJet` [78] implementation of the anti-k_t clustering algorithm [79] with radius parameter $R = 0.4$. We require at least two jets each with $p_T > 10$ GeV. Similarly, we require the leptons to be oppositely charged, and to each have $p_T > 10$ GeV.

The four-vector of each lepton and the two leading jets are used as features. In particular we use the mass, m, of the jet or lepton as a feature. It is unlikely that the mass of a jet could be measured with enough precision in an actual experiment to distinguish a charm initiated jet from a bottom jet. However the mass of the jet is a proxy for the lifetime of the initiating particle

[3] A complementary approach is to exclusively search for charmed-hadrons [70].

of the jet, which is a feature flavor tagging algorithms exploit, see *e.g.* [54]. The four-vectors of the dilepton and dijet systems, which reconstruct the Z and Higgs bosons, respectively, are also included in our feature set. A cut on the invariant mass of the jets is imposed, $95 < m_{jj}/\text{GeV} < 155$, to concentrate on resonant Higgs production. All of the above cuts and requirements reduce the number of simulated events to approximately $8.9 \cdot 10^4$. We include

$$\Delta R = \sqrt{(\eta_{j1} - \eta_{j2})^2 + (\phi_{j1} - \phi_{j2})^2} \tag{12}$$

between the two jets as a feature as well as the rescaled mass drop observable, ISY, and the radius of the dijet system, R_{jj},

$$ISY = \frac{\max(m_{j1}, m_{j2})\Delta R}{m_{jj}}, \quad R_{jj} = \frac{m_{jj}(p_{T,j1} + p_{T,j2})}{p_{T,jj}\sqrt{p_{T,j1}p_{T,j2}}}. \tag{13}$$

Lastly, as bottom- and charm-quarks are oppositely charged, we look at the charge of the jets as defined in [80]

$$\mathcal{Q}^j_\kappa = \frac{1}{(p_{T,j})^\kappa} \sum_{p \in j} Q_p (p_{T,p})^\kappa \tag{14}$$

where the charge, $\mathcal{Q}$, of a jet, j is the p_T weighted sum of the charges, Q, of all the partons, p, in the jet. We use $\kappa = 0.4$ in this work. Of course only the overall magnitude of the jet charges differ between bottom and charm Higgs decays. Therefore, in addition to the charge of each jet, we include the product of the jet charges, the absolute value of the difference of the jet charges, and the charge of the dijet system, bringing our total number of features to 30.

4.3 Models and Training

Our heavy flavor tagging model is a `LightGBM` [51]. In particular, our model combines a mere 50 trees in series, and each tree is allowed to have a maximum depth of 10 with all other hyperparameters fixed to their default values. We take as our baseline heavy flavor tagger a `LightGBM` with an unweighted loss function, and compare its performance against a `LightGBM` with weighting $\alpha = 1 - r$.

For model evaluation we again perform a stratified five-fold cross validation. We test three scenarios. In the first test we assume the rate for $h \to c\bar{c}$ is equivalent to the rate for $h \to b\bar{b}$. Here we use the baseline LGBM with unweighted loss function. In this case there is no class imbalance implying there must be some BSM physics in this scenario. We randomly select $4.0 \cdot 10^4$ bottom and $4.0 \cdot 10^4$ charm events from our full simulated dataset, and perform the cross validation on this sample. For the second test we again use the unweighted, baseline model, but perform the cross validation on dataset with SM-like class imbalance. In particular we randomly select $4.0 \cdot 10^4$ bottom and $2.0 \cdot 10^3$ charm events from our full simulated dataset. For the third and final test we reuse the dataset from the second test, but use our class imbalance optimized LGBM with weighting hyperparameter $\alpha = 1 - r \approx 0.95$.

4.4 Results

The results of our three $h \to c\bar{c}$ tagging tests are given in Table 2 with the rows from top to bottom corresponding to the 1st, 2nd, and 3rd scenarios described in previous subsection. For each scenario we consider two signal efficiency working points, a looser selection of $\epsilon_{h\to c\bar{c}} = TPR = 0.2$ and a tighter selection of $\epsilon_{h\to c\bar{c}} = 0.8$. We report the background rejection rate,

Model	α	r	$\epsilon_{h\to c\bar{c}}$	$1/\epsilon_{h\to b\bar{b}}$	Average Precision	AP/r
LightGBM	$\frac{1}{2}$	1	0.2 0.8	38.8 ± 2.6 1.7 ± 0.0	0.719 ± 0.004	0.7 ± 0.0
LightGBM	$\frac{1}{2}$	0.05	0.2 0.8	30.9 ± 5.2 1.6 ± 0.1	0.166 ± 0.008	3.3 ± 0.2
Weighted LGBM	$1-r$	0.05	0.2 0.8	35.1 ± 8.9 1.5 ± 0.1	0.161 ± 0.011	3.2 ± 0.2

Table 2: The results of our three $h \to c\bar{c}$ tagging tests. We report the background rejection rate, $\epsilon_{h\to b\bar{b}} = FPR$, for two signal efficiency working points, $\epsilon_{h\to c\bar{c}} = TPR = 0.2$(loose), 0.8(tight). There is a 14% increase in $1/\epsilon_{h\to b\bar{b}}$ with loose selection criteria when the class imbalance optimized model is used, 3rd versus 2nd row, demonstrating proof of principle that class imbalance techniques can be used to improve the performance of algorithms used to identify $h \to c\bar{c}$ events. We also report the AP, and AP/r, with the latter given to one decimal place for better readability.

$\epsilon_{h\to b\bar{b}} = FPR$, for each of these working points. The inverse of the background rejection rate is largest in the scenario without class imbalance. The performance of both tagging models is worse in the presence of class imbalance. However the weighted LGBM outperforms the baseline tagging model in the presence of class imbalance, demonstrating proof of principle that class imbalance techniques can be used to improve the performance of algorithms used to identify $h \to c\bar{c}$ events. In particular, there is a 14% increase in $1/\epsilon_{h\to b\bar{b}}$ with loose selection criteria when the class imbalance optimized model is used.

We also report the average precision, and average precision normalized by the imbalance ratio. The average precision is significantly higher in the scenario without class imbalance. However when the average precision is normalized by the imbalance ratio, which constitutes the naïve expectation for the AP score, higher values are found when the data is imbalanced.

Additionally, Fig. 4 shows the precision-recall curves for our three $h \to c\bar{c}$ tagging tests. The blue, orange, and green curves correspond to the test results in the top, middle, and bottom rows of Table 2, respectively. These curves provide another way of demonstrating that the weighted LGBM outperforms (green) outperforms its unweighted counterpart (orange). Specifically, at lower recall, weighting the loss function to remove class imbalance leads to a gain in performance. Recall is equivalent to true position rate or signal efficiency, $\epsilon_{h\to c\bar{c}}$.

Lastly, we investigate which features are important for the classification. Using the feature importance of the LGBM the charges of the heavy flavor jets and the associated engineered features do not play a significant role in discriminating charm initiated jets from bottom jets. This is in contrast with studies of light flavored jets [81]. A possible explanation for this is the heavy flavored hadrons have more possible decay chains.[4] In particular, a neutral meson may oscillate or there might be a cascade decay that spoils the correlation between the charges of the partons in the jet and the charge of the particle that initiated the jet. Again using the feature importance of the LGBM, we find the the four-vectors of the leptons and the four-vector of the reconstructed Z boson also do not play a major role in discriminating charm initiated jets from bottom jets.

[4] This is one of the main systematic uncertainties in measuring asymmetric heavy quark hadroproduction [82, 83, 84, 85].

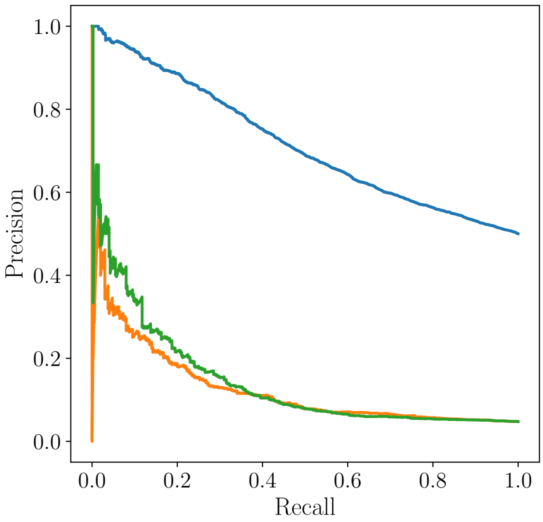


Figure 4: The precision-recall curves for our three $h \to c\bar{c}$ tagging tests. The blue, orange, and green curves correspond to the test results in the top, middle, and bottom rows of Table 2, respectively. These curves provide another way of demonstrating that the weighted LGBM outperforms (green) outperforms it unweighted counterpart (orange). Specifically, at lower recall, weight the loss function to remove class imbalance leads to a gain in performance. Recall is equivalent to true position rate or signal efficiency, $\epsilon_{h\to c\bar{c}}$.

5 Discussion

Extracting a signal from a much larger background is a common problem in high energy physics. Posed as a classification task, there is said to be an imbalance in the number of samples belonging to the signal class versus the number of samples from the background class. Imbalanced learning techniques are not commonly used, explicitly anyways, in high energy physics. Given this lack of use we first provided a brief overview of modern class imbalance techniques in a high energy physics setting, introducing novel loss functions and a data resampling technique. We then presented two case studies illustrating these techniques. The first study is the measurement of the longitudinal polarization fraction in same-sign WW scattering. We found a modest improvement in the performance of both the classic ML models and in the deep learning models tested in the longitudinal WW study. Our neural networks achieves comparable performance to that of Ref. [46] despite having only two hidden layers instead of 10. Given that there are only $\mathcal{O}(10)$ features in this dataset it is not surprising that a very deep network did not continue to improve performance. Having fewer hidden layers with all else being equal results in a reduction in training time. The second case is the decay of the Higgs boson to charm-quark pairs. We delivered proof of principle that it is possible to improve tagging efficiency of $h \to c\bar{c}$ events through the use of the class imbalance techniques. In particular, our Higgs-to-charm tagger with loose selection criteria gave a 14% improvement in the background rejection rate.

Acknowledgements We thank Eder Izaguirre and Brian Shuve for collaborations on a previous incarnation of the $h \to c\bar{c}$ tagging study. We would also like to thank Emmanuel Ameisen, Sally Dawson, Samuel Homiller, and Marc-André Pleier for useful discussions. Finally, we are grateful to Tilman Plehn and the anonymous referee for many helpful comments on the manuscript.

A Glossary

See Table 3 for a glossary of model evaluation terms used in this work.

References

[1] M. Tanabashi *et al.*, *Review of particle physics*, Phys. Rev. D **98**, 030001 (2018), doi:10.1103/PhysRevD.98.030001.

[2] V. Khachatryan *et al.*, *Observation of the diphoton decay of the Higgs boson and measurement of its properties*, Eur. Phys. J. **C74**(10), 3076 (2014), doi:10.1140/epjc/s10052-014-3076-z, 1407.0558.

[3] G. Aad *et al.*, *Measurement of Higgs boson production in the diphoton decay channel in pp collisions at center-of-mass energies of 7 and 8 TeV with the ATLAS detector*, Phys. Rev. **D90**(11), 112015 (2014), doi:10.1103/PhysRevD.90.112015, 1408.7084.

[4] A. J. Larkoski, I. Moult and B. Nachman, *Jet Substructure at the Large Hadron Collider: A Review of Recent Advances in Theory and Machine Learning* (2017), 1709.04464.

Metric	Symbol	Definition
Accuracy	A	$A = (TP + TN)/(FN + FP + TN + TP)$
Area Under the ROC Curve	AUC	$AUC = \int_0^1 d(TPR)\,[1 - FPR(TPR)]$
Average Precision	AP	$AP = \sum_n (R_n - R_{n-1})P_n$
Decision Threshold	n	if $p > n$ for a given event, then that event is predicted to be signal
F1 score	F_1	$F_1 = 2P \cdot R/(P + R)$
False Negative	FN	a signal event that is predicted to be background
False Positive	FP	a background event that is predicted to be signal
False Positive Rate	FPR	$FPR = FP/(FP + TN)$
Ground Truth Class	y	$y = 1$ if the event is truly a signal event, and $y = 0$ if it is background
Precision	P	$P = TP/(FP + TP)$
Probability Estimate	p	a model's estimated probability that a given event belongs to the signal class
Recall	R	$R = TP/(FN + TP)$
True Negative	TN	a background event that is predicted to be background
True Positive	TP	a signal event that is predicted to be signal
True Positive Rate	TPR	$TPR = R$

Table 3: Glossary of model evaluation terms used in this work.

[5] D. Guest, K. Cranmer and D. Whiteson, *Deep Learning and its Application to LHC Physics*, Ann. Rev. Nucl. Part. Sci. **68**, 161 (2018), doi:10.1146/annurev-nucl-101917-021019, 1806.11484.

[6] K. Albertsson *et al.*, *Machine Learning in High Energy Physics Community White Paper*, J. Phys. Conf. Ser. **1085**(2), 022008 (2018), doi:10.1088/1742-6596/1085/2/022008, 1807.02876.

[7] A. Radovic, M. Williams, D. Rousseau, M. Kagan, D. Bonacorsi, A. Himmel, A. Aurisano, K. Terao and T. Wongjirad, *Machine learning at the energy and intensity frontiers of particle physics*, Nature **560**(7716), 41 (2018), doi:10.1038/s41586-018-0361-2.

[8] G. Lemaître, F. Nogueira and C. K. Aridas, *Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning*, Journal of Machine Learning Research **18**(17), 1 (2017).

[9] R. Aaij *et al.*, *Measurement of the time-integrated CP asymmetry in $D^0 \to K^0_S K^0_S$ decays*, JHEP **10**, 055 (2015), doi:10.1007/JHEP10(2015)055, 1508.06087.

[10] M. Britsch, N. Gagunashvili and M. Schmelling, *Application of the rule-growing algorithm RIPPER to particle physics analysis*, PoS **ACAT08**, 086 (2008), doi:10.22323/1.070.0086, 0910.1729.

[11] M. Britsch, N. Gagunashvili and M. Schmelling, *Classifying extremely imbalanced data sets*, PoS **ACAT2010**, 047 (2010), doi:10.22323/1.093.0047, 1011.6224.

[12] J. H. Collins, K. Howe and B. Nachman, *Anomaly Detection for Resonant New Physics with Machine Learning*, Phys. Rev. Lett. **121**(24), 241803 (2018), doi:10.1103/PhysRevLett.121.241803, 1805.02664.

[13] M. Farina, Y. Nakai and D. Shih, *Searching for New Physics with Deep Autoencoders* (2018), 1808.08992.

[14] T. Heimel, G. Kasieczka, T. Plehn and J. M. Thompson, *QCD or What?*, SciPost Phys. **6**(3), 030 (2019), doi:10.21468/SciPostPhys.6.3.030, 1808.08979.

[15] J. H. Collins, K. Howe and B. Nachman, *Extending the search for new resonances with machine learning*, Phys. Rev. **D99**(1), 014038 (2019), doi:10.1103/PhysRevD.99.014038, 1902.02634.

[16] L. M. Dery, B. Nachman, F. Rubbo and A. Schwartzman, *Weakly Supervised Classification in High Energy Physics*, JHEP **05**, 145 (2017), doi:10.1007/JHEP05(2017)145, 1702.00414.

[17] T. Cohen, M. Freytsis and B. Ostdiek, *(Machine) Learning to Do More with Less*, JHEP **02**, 034 (2018), doi:10.1007/JHEP02(2018)034, 1706.09451.

[18] E. M. Metodiev, B. Nachman and J. Thaler, *Classification without labels: Learning from mixed samples in high energy physics*, JHEP **10**, 174 (2017), doi:10.1007/JHEP10(2017)174, 1708.02949.

[19] https://github.com/christopher-w-murphy/Class-Imbalance-in-WW-Polarization.

[20] Z. Ding, *Diversified ensemble classifiers for highly imbalanced data learning and their application in bioinformatics* (2011).

[21] M. Kubat, S. Matwin *et al.*, *Addressing the curse of imbalanced training sets: one-sided selection*, In *Icml*, vol. 97, pp. 179–186. Nashville, USA (1997).

[22] J. Laurikkala, *Improving identification of difficult small classes by balancing class distribution*, In S. Quaglini, P. Barahona and S. Andreassen, eds., *Artificial Intelligence in Medicine*, pp. 63–66. Springer Berlin Heidelberg, Berlin, Heidelberg, ISBN 978-3-540-48229-1 (2001).

[23] I. Mani and I. Zhang, *knn approach to unbalanced data distributions: a case study involving information extraction*, In *Proceedings of workshop on learning from imbalanced datasets*, vol. 126 (2003).

[24] B. C. Wallace, K. Small, C. E. Brodley and T. A. Trikalinos, *Class imbalance, redux*, In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining*, ICDM '11, pp. 754–763. IEEE Computer Society, Washington, DC, USA, ISBN 978-0-7695-4408-3, doi:10.1109/ICDM.2011.33 (2011).

[25] M. R. Smith, T. Martinez and C. Giraud-Carrier, *An instance level analysis of data complexity*, Machine Learning **95**(2), 225 (2014), doi:10.1007/s10994-013-5422-z.

[26] H. Han, W.-Y. Wang and B.-H. Mao, *Borderline-smote: A new over-sampling method in imbalanced data sets learning*, In D.-S. Huang, X.-P. Zhang and G.-B. Huang, eds., *Advances in Intelligent Computing*, pp. 878–887. Springer Berlin Heidelberg, Berlin, Heidelberg, ISBN 978-3-540-31902-3 (2005).

[27] H. M. Nguyen, E. W. Cooper and K. Kamei, *Borderline over&#45;sampling for imbalanced data classification*, Int. J. Knowl. Eng. Soft Data Paradigm. **3**(1), 4 (2011), doi:10.1504/IJKESDP.2011.039875.

[28] G. E. Batista, A. L. Bazzan and M. C. Monard, *Balancing training data for automated annotation of keywords: a case study.*, In *WOB*, pp. 10–18 (2003).

[29] G. E. A. P. A. Batista, R. C. Prati and M. C. Monard, *A study of the behavior of several methods for balancing machine learning training data*, SIGKDD Explor. Newsl. **6**(1), 20 (2004), doi:10.1145/1007730.1007735.

[30] C. Chen, A. Liaw, L. Breiman *et al.*, *Using random forest to learn imbalanced data*, University of California, Berkeley **110**, 1 (2004).

[31] X. Liu, J. Wu and Z. Zhou, *Exploratory undersampling for class-imbalance learning*, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **39**(2), 539 (2009), doi:10.1109/TSMCB.2008.2007853.

[32] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse and A. Napolitano, *Rusboost: A hybrid approach to alleviating class imbalance*, IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans **40**(1), 185 (2010), doi:10.1109/TSMCA.2009.2029559.

[33] S. T. Goh and C. Rudin, *Box drawings for learning with imbalanced data*, In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pp. 333–342. ACM, New York, NY, USA, ISBN 978-1-4503-2956-9, doi:10.1145/2623330.2623648 (2014).

[34] F. T. Liu, K. M. Ting and Z.-H. Zhou, *Isolation-based anomaly detection*, ACM Trans. Knowl. Discov. Data **6**(1), 3:1 (2012), doi:10.1145/2133360.2133363.

[35] T. R. Bandaragoda, K. M. Ting, D. Albrecht, F. T. Liu and J. R. Wells, *Efficient anomaly detection by isolation using nearest neighbour ensemble*, In *2014 IEEE International Conference on Data Mining Workshop*, pp. 698–705, doi:10.1109/ICDMW.2014.70 (2014).

[36] T.-Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, *Focal Loss for Dense Object Detection*, arXiv e-prints arXiv:1708.02002 (2017), `1708.02002`.

[37] K. Feng, H. Hong, K. Tang and J. Wang, *Decision making with machine learning and roc curves*, arXiv preprint arXiv:1905.02810 (2019).

[38] J. F. Gunion, H. E. Haber and J. Wudka, *Sum rules for Higgs bosons*, Phys. Rev. **D43**, 904 (1991), doi:10.1103/PhysRevD.43.904.

[39] B. Grinstein, C. W. Murphy, D. Pirtskhalava and P. Uttayarat, *Theoretical Constraints on Additional Higgs Bosons in Light of the 126 GeV Higgs*, JHEP **05**, 083 (2014), doi:10.1007/JHEP05(2014)083, `1401.0070`.

[40] *Observation of electroweak production of a same-sign W boson pair in association with two jets in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, Tech. Rep. ATLAS-CONF-2018-030, CERN, Geneva (2018).

[41] A. M. Sirunyan *et al.*, *Observation of electroweak production of same-sign W boson pairs in the two jet and two same-sign lepton final state in proton-proton collisions at $\sqrt{s}$ = 13 TeV*, Phys. Rev. Lett. **120**(8), 081801 (2018), doi:10.1103/PhysRevLett.120.081801, `1709.05822`.

[42] *Prospects for the measurement of the $W^\pm W^\pm$ scattering cross section and extraction of the longitudinal scattering component in pp collisions at the High-Luminosity LHC with the ATLAS experiment*, Tech. Rep. ATL-PHYS-PUB-2018-052, CERN, Geneva (2018).

[43] *Study of $W^\pm W^\pm$ production via vector boson scattering at the HL-LHC with the upgraded CMS detector*, Tech. Rep. CMS-PAS-FTR-18-005, CERN, Geneva (2018).

[44] P. Azzi *et al.*, *Standard Model Physics at the HL-LHC and HE-LHC* (2019), `1902.04070`.

[45] J. Searcy, L. Huang, M.-A. Pleier and J. Zhu, *Determination of the WW polarization fractions in $pp \to W^\pm W^\pm jj$ using a deep machine learning technique*, Phys. Rev. **D93**(9), 094033 (2016), doi:10.1103/PhysRevD.93.094033, `1510.01691`.

[46] J. Lee, N. Chanon, A. Levin, J. Li, M. Lu, Q. Li and Y. Mao, *Polarization fraction measurement in same-sign WW scattering using deep learning*, Phys. Rev. **D99**(3), 033004 (2019), doi:10.1103/PhysRevD.99.033004, `1812.07591`.

[47] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli and M. Zaro, *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, JHEP **07**, 079 (2014), doi:10.1007/JHEP07(2014)079, `1405.0301`.

[48] P. Artoisenet, R. Frederix, O. Mattelaer and R. Rietkerk, *Automatic spin-entangled decays of heavy resonances in Monte Carlo simulations*, JHEP **03**, 015 (2013), doi:10.1007/JHEP03(2013)015, `1212.3460`.

[49] D. L. Rainwater, R. Szalapski and D. Zeppenfeld, *Probing color singlet exchange in Z + two jet events at the CERN LHC*, Phys. Rev. **D54**, 6680 (1996), doi:10.1103/PhysRevD.54.6680, `hep-ph/9605444`.

[50] F. Pedregosa *et al.*, *Scikit-learn: Machine learning in Python*, Journal of Machine Learning Research **12**, 2825 (2011).

[51] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu, *Lightgbm: A highly efficient gradient boosting decision tree*, In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, eds., *Advances in Neural Information Processing Systems 30*, pp. 3146–3154. Curran Associates, Inc. (2017).

[52] F. Chollet *et al.*, *Keras*, `https://keras.io` (2015).

[53] M. Abadi *et al.*, *TensorFlow: Large-scale machine learning on heterogeneous systems*, Software available from tensorflow.org (2015).

[54] M. Aaboud *et al.*, *Search for the Decay of the Higgs Boson to Charm Quarks with the ATLAS Experiment*, Phys. Rev. Lett. **120**(21), 211802 (2018), doi:10.1103/PhysRevLett.120.211802, `1802.04329`.

[55] *Search for $H^0 \to b\bar{b}$ or $c\bar{c}$ in association with a W or Z boson in the forward region of pp collisions* (2016).

[56] J. de Blas, M. Ciuchini, E. Franco, S. Mishima, M. Pierini, L. Reina and L. Silvestrini, *The Global Electroweak and Higgs Fits in the LHC era*, PoS **EPS-HEP2017**, 467 (2017), doi:10.22323/1.314.0467, 1710.05402.

[57] J. de Blas, O. Eberhardt and C. Krause, *Current and Future Constraints on Higgs Couplings in the Nonlinear Effective Theory*, JHEP **07**, 048 (2018), doi:10.1007/JHEP07(2018)048, 1803.00939.

[58] J. Ellis, C. W. Murphy, V. Sanz and T. You, *Updated Global SMEFT Fit to Higgs, Diboson and Electroweak Data*, JHEP **06**, 146 (2018), doi:10.1007/JHEP06(2018)146, 1803.03252.

[59] J. Aebischer, J. Kumar, P. Stangl and D. M. Straub, *A Global Likelihood for Precision Constraints and Flavour Anomalies*, Eur. Phys. J. **C79**(6), 509 (2019), doi:10.1140/epjc/s10052-019-6977-z, 1810.07698.

[60] E. da Silva Almeida, A. Alves, N. Rosa Agostinho, O. J. P. Eboli and M. C. Gonzalez-Garcia, *Electroweak Sector Under Scrutiny: A Combined Analysis of LHC and Electroweak Precision Data*, Phys. Rev. **D99**(3), 033001 (2019), doi:10.1103/PhysRevD.99.033001, 1812.01009.

[61] A. Biekotter, T. Corbett and T. Plehn, *The Gauge-Higgs Legacy of the LHC Run II*, SciPost Phys. **6**, 064 (2019), doi:10.21468/SciPostPhys.6.6.064, 1812.07587.

[62] N. P. Hartland, F. Maltoni, E. R. Nocera, J. Rojo, E. Slade, E. Vryonidou and C. Zhang, *A Monte Carlo global analysis of the Standard Model Effective Field Theory: the top quark sector*, JHEP **04**, 100 (2019), doi:10.1007/JHEP04(2019)100, 1901.05965.

[63] M. Cepeda *et al.*, *Higgs Physics at the HL-LHC and HE-LHC* (2019), 1902.00134.

[64] I. Brivio, S. Bruggisser, F. Maltoni, R. Moutafis, T. Plehn, E. Vryonidou, S. Westhoff and C. Zhang, *O new physics, where art thou? A global search in the top sector* (2019), 1910.03606.

[65] R. Lafaye, T. Plehn, M. Rauch, D. Zerwas and M. Duhrssen, *Measuring the Higgs Sector*, JHEP **08**, 009 (2009), doi:10.1088/1126-6708/2009/08/009, 0904.3866.

[66] *Prospects for $H \to c\bar{c}$ using Charm Tagging with the ATLAS Experiment at the HL-LHC*, Tech. Rep. ATL-PHYS-PUB-2018-016, CERN, Geneva (2018).

[67] J. R. Andersen *et al.*, *Handbook of LHC Higgs Cross Sections: 3. Higgs Properties* (2013), doi:10.5170/CERN-2013-004, 1307.1347.

[68] M. Aaboud *et al.*, *Observation of $H \to b\bar{b}$ decays and VH production with the ATLAS detector*, Phys. Lett. **B786**, 59 (2018), doi:10.1016/j.physletb.2018.09.013, 1808.08238.

[69] A. M. Sirunyan *et al.*, *Observation of Higgs boson decay to bottom quarks*, Phys. Rev. Lett. **121**(12), 121801 (2018), doi:10.1103/PhysRevLett.121.121801, 1808.08242.

[70] M. Aaboud *et al.*, *Searches for exclusive Higgs and Z boson decays into $J/\psi\gamma$, $\psi(2S)\gamma$, and $\Upsilon(nS)\gamma$ at $\sqrt{s} = 13$ TeV with the ATLAS detector*, Phys. Lett. **B786**, 134 (2018), doi:10.1016/j.physletb.2018.09.024, 1807.00802.

[71] C. Delaunay, T. Golling, G. Perez and Y. Soreq, *Enhanced Higgs boson coupling to charm pairs*, Phys. Rev. **D89**(3), 033014 (2014), doi:10.1103/PhysRevD.89.033014, 1310.7029.

[72] G. Perez, Y. Soreq, E. Stamou and K. Tobioka, *Prospects for measuring the Higgs boson coupling to light quarks*, Phys. Rev. **D93**(1), 013001 (2016), doi:10.1103/PhysRevD.93.013001, 1505.06689.

[73] R. Aaij *et al.*, *Identification of beauty and charm quark jets at LHCb*, JINST **10**(06), P06013 (2015), doi:10.1088/1748-0221/10/06/P06013, 1504.07670.

[74] G. Aad *et al.*, *Performance of b-Jet Identification in the ATLAS Experiment*, JINST **11**(04), P04008 (2016), doi:10.1088/1748-0221/11/04/P04008, 1512.01094.

[75] I. Connelly, *Performance and calibration of b-tagging with the ATLAS experiment at LHC Run-2*, EPJ Web Conf. **164**, 07025 (2017), doi:10.1051/epjconf/201716407025.

[76] A. Lenz, M. Spannowsky and G. Tetlalmatzi-Xolocotzi, *Double-charming Higgs boson identification using machine-learning assisted jet shapes*, Phys. Rev. **D97**(1), 016001 (2018), doi:10.1103/PhysRevD.97.016001, 1708.03517.

[77] T. Sjostrand, S. Mrenna and P. Z. Skands, *PYTHIA 6.4 Physics and Manual*, JHEP **05**, 026 (2006), doi:10.1088/1126-6708/2006/05/026, hep-ph/0603175.

[78] M. Cacciari, G. P. Salam and G. Soyez, *FastJet User Manual*, Eur. Phys. J. **C72**, 1896 (2012), doi:10.1140/epjc/s10052-012-1896-2, 1111.6097.

[79] M. Cacciari, G. P. Salam and G. Soyez, *The anti-k_t jet clustering algorithm*, JHEP **04**, 063 (2008), doi:10.1088/1126-6708/2008/04/063, 0802.1189.

[80] D. Krohn, M. D. Schwartz, T. Lin and W. J. Waalewijn, *Jet Charge at the LHC*, Phys. Rev. Lett. **110**(21), 212001 (2013), doi:10.1103/PhysRevLett.110.212001, 1209.2421.

[81] W. J. Waalewijn, *Calculating the Charge of a Jet*, Phys. Rev. **D86**, 094030 (2012), doi:10.1103/PhysRevD.86.094030, 1209.3019.

[82] B. Grinstein and C. W. Murphy, *Bottom-Quark Forward-Backward Asymmetry in the Standard Model and Beyond*, Phys. Rev. Lett. **111**, 062003 (2013), doi:10.1103/PhysRevLett.112.239901, 10.1103/PhysRevLett.111.062003, [Erratum: Phys. Rev. Lett.112,no.23,239901(2014)], 1302.6995.

[83] C. W. Murphy, *Bottom-Quark Forward-Backward and Charge Asymmetries at Hadron Colliders*, Phys. Rev. **D92**(5), 054003 (2015), doi:10.1103/PhysRevD.92.054003, 1504.02493.

[84] R. Gauld, U. Haisch, B. D. Pecjak and E. Re, *Beauty-quark and charm-quark pair production asymmetries at LHCb*, Phys. Rev. **D92**, 034007 (2015), doi:10.1103/PhysRevD.92.034007, 1505.02429.

[85] R. Gauld, U. Haisch and B. D. Pecjak, *Asymmetric heavy-quark hadroproduction at LHCb: Predictions and applications*, JHEP **03**, 166 (2019), doi:10.1007/JHEP03(2019)166, 1901.07573.