applied sciences

MDPI

Review

Machine-Learning Application for a Likelihood Ratio Estimation Problem at LHC

Silvia Auricchio [1,2,*,†], Francesco Cirotto [1,2,*,†] and Antonio Giannini [3,*,†]

[1] Dipartimento di Fisica "Ettore Pancini", Università Degli Studi di Napoli Federico II, Complesso Univ. Monte S. Angelo, Via Cinthia, 21, Edificio 6, 80126 Napoli, Italy
[2] INFN-Sezione di Napoli, Complesso Univ. Monte S. Angelo, Via Cinthia, Edificio 6, 80126 Napoli, Italy
[3] Center of Innovation and Cooperation for Particle and Interaction (CICPI), University of Science and Technology of China (USTC), No. 96, JinZhai Road, Baohe District, Hefei 230026, China
* Correspondences: silvia.auricchio@unina.it (S.A.); francesco.cirotto@unina.it (F.C.); antonio.giannini@cern.ch (A.G.)
† These authors contributed equally to this work.

Abstract: High-energy physics is now entering a new era. Current particle experiments, such as ATLAS at the Large Hadron Collider at CERN, offer the possibility of discovering new and interesting physics' phenomena beyond the Standard Model, the theoretical framework which describes fundamental interactions between particles. In this paper a machine-learning algorithm to deal with the estimate of the expected background processes distribution, a very demanding task for particle physics analyses, is described. A new technique is exploited, inspired by the well-known likelihood ratio estimation problem, called direct importance estimation in statistics. First, its theoretical formulation is discussed, then its performances in two ATLAS analyses are described.

Keywords: experimental particle physics; machine learning; background estimation; likelihood ratio estimation; LHC; ATLAS experiment

check for updates

Citation: Auricchio, S.; Cirotto, F.; Giannini, A. Machine-Learning Application for a Likelihood Ratio Estimation Problem at LHC. *Appl. Sci.* **2023**, *13*, 86. https://doi.org/10.3390/app13010086

Academic Editors: Guijun Bi, Francesco Conventi and Orso Iorio

Received: 14 November 2022
Revised: 13 December 2022
Accepted: 16 December 2022
Published: 21 December 2022

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The Large Hadron Collider (LHC) [1] is a proton–proton particle accelerator placed at CERN whose physics program, after the successful Higgs boson discovery in 2012 [2], aims to search for new phenomena in physics. The LHC long-time schedule involves the collection of a considerable amount of data both to improve the precision of measurements and to find any extremely rare phenomena predicted by Beyond Standard Model theories. This goal can be achieved by increasing the number of proton collisions within the accelerator ring; however, to be able to collect data at a higher frequency, it was necessary to implement a complex upgrade program, on the hardware, electronics and software. The capacity to process and analyse the large amount of data is essential. In this context, machine learning (ML) finds fertile ground for application, as it offers solutions to the need for rapid analysis of inner structures and hidden pieces of information in large datasets.

In this paper, we focus on one of the many different applications of an ML algorithm for particle physics analysis, involving the adoption of a deep neural network to address one of the main challenges, that of background process estimation. Techniques applied have mostly involved Monte Carlo simulations, but, if the complexity of the processes is too high, a data-driven approach is preferred.

The application examples considered are drawn from two publications of the ATLAS Collaboration [3], one of the four major experiments designed to analyze data from particle collisions, involving a broad physics program ranging from precision measurements to direct searches for new particles and new interactions.

In Section 2, the physical problem is introduced and the theoretical solution adopted is described. Two applications to specific analyses are then discussed in Section 3.

2. The Reweighting Problem

In particle physics analyses, we are typically interested in studying a particular process called *signal*. To be sensitive to seeing it, selection of specific data is usually performed, defining regions that contain events of interest called *signal regions* (SR). Non-signal events, referred to as *background*, with the same experimental signature, can pass these selections; this is problematic and requires them to be distinguished from the signal. The estimation of *background* events is typically performed by definition of additional regions, called *control regions* (CRs), which are complementary to the SR where no *signal* events are expected. By studying the data distribution in these CRs, the expected distribution of *background* events in SR can be inferred; however, since these two are not expected to have identical kinematic characteristics, a reweighting function has to be derived.
More technically, if p_0 and p_1 are the probability distribution functions (*pdf*s) of a real variable X describing two different kinematic regions (labelled 0 and 1), the reweighting function between them corresponds to their *likelihood ratio* (or *importance*) $w(X)$:

$$p_1(X) = w(X) \cdot p_0(X) \rightarrow w(X) = \frac{p_1(X)}{p_0(X)}. \tag{1}$$

In the real world, we do not know the *pdf*s and the most we can do is to calculate an approximation of the data sampled from the two different kinematic regions 0 and 1. Unfortunately, even this problem is often not simple and straightforward to solve; however, there is a means of addressing this, which consists of directly estimating the *likelihood ratio* from data without determining the single *pdf*s. In statistics, this is referred to as *direct importance estimation* and can represent a positive solution if only the *pdf* ratio is needed.

In this paper, *background* estimation involves obtaining the *likelihood ratio* between SR and CRs from (1). This has previously been performed using "classical" approaches, involving attempting to approximate the *pdf* of each region with the most discriminant variable distributions; however, obviouszly, this technique cannot take into account the complex multi-dimensionality of data. The use of a deep neural network, which is known to be able to perform well in complex learning and the identification of hidden correlations in multi-dimensional datasets, appeared to offer a natural solution to this problem. Its use is described in the following Section.

2.1. A Least-Squares Approach to Direct Importance Estimation

Let us first formulate the direct importance estimation problem as a least-squares-function-fitting problem, as described in [4], that can be readily generalized.

From this point, the estimate of quantities of data will be indicated with the symbol $\hat{}$.

Let us consider two *pdf*s, p_0 and p_1, and let us suppose that we have N_0 independent and identically distributed (i.i.d.) measurements extracted from p_0 and N_1 i.i.d. measurements extracted from p_1 of B variables: $\{X^0_{i,\,j}\}^{N_0,\,B}_{i=1,\,j=1}$ and $\{X^1_{l,\,j}\}^{N_1,\,B}_{l=1,\,j=1}$.

We neither know the analytical form of $p_0(\underline{X})$ and $p_1(\underline{X})$, nor their empirical realization, but we are interested in the ratio between them and can estimate it from the two samples of the extracted measurements.

Let us call $w(\underline{X}) = \frac{p_1(\underline{X})}{p_0(\underline{X})}$ the analytical ratio and $\hat{w}(\underline{X}) = \sum_f \alpha_f \cdot \phi_f(\underline{X})$ its estimate from data, which can be written as a linear combination of arbitrarily chosen functions of data $\phi_f(\underline{X})$ that must satisfy the same properties of $w(\underline{X})$:

- $w(\underline{X}) = \frac{p_1(\underline{X})}{p_0(\underline{X})} > 0$
- $w(\underline{X}) \rightarrow +\infty$ if $\underline{X} \in p_1(\underline{X})$
- $w(\underline{X}) \rightarrow 0$ if $\underline{X} \in p_0(\underline{X})$

The coefficients $\underline{\alpha}$ are constants that are chosen in such a way to obtain the best estimate of $w(\underline{X})$ by minimizing a loss function which depends on them.

Now that we have fixed the notation, we can formulate our minimization problem by defining a loss function as the sum of the quadratic differences between the true value $w(\underline{X})$ and its estimates on data $\hat{w}(\underline{X})$.

Thus, the loss function J_0 can be written with the following expression:

$$J_0(\underline{\alpha}) = E_0[(\hat{w}(\underline{X}) - w(\underline{X}))^2] = \int d\underline{X}\; p_0(\underline{X})\; (\hat{w}(\underline{X}) - w(\underline{X}))^2 = \int d\underline{X}\; p_0(\underline{X})\; \hat{w}(\underline{X})^2 \; + \int d\underline{X}\; p_0(\underline{X})\; w(\underline{X})^2 \; - \; 2 \int \; d\underline{X}\; p_0(\underline{X})\; \hat{w}(\underline{X}) \cdot w(\underline{X}) = \\ = \int d\underline{X}\; p_0(\underline{X})\; \hat{w}(\underline{X})^2 \; - 2 \cdot \int d\underline{X}\; p_1(\underline{X})\; \hat{w}(\underline{X}) \tag{2}$$

Since, in the centre line of (2), the second term does not depend on data, it can be ignored and the expression can be written as follows:

$$J_0(\underline{\alpha}) = E_0[\hat{w}(\underline{X})^2 - 2 \cdot w(\underline{X}) \cdot \hat{w}(\underline{X})] = E_0[\hat{w}(\underline{X})^2] - 2 \cdot E_1[\hat{w}(\underline{X})] \tag{3}$$

This loss function has to be minimized with respect to the constants $\underline{\alpha}$ on data to obtain the best estimate of $w(\underline{X})$.

If $\hat{w}(\underline{X})$ is intended as the output of a neural network (NN), and the constants $\underline{\alpha}$ as its weights $\underline{\theta}$, the loss function J_0 can effectively be taken to be a loss function to be minimized by an NN.

Since $\hat{w}(\underline{X})$ is positively defined, a constraint term should be added to the loss function, or $\hat{w}(\underline{X})$ should be modified with a monotone function and the loss consequently adapted. For example, the choice of log-likelihood ratio estimation offers a straightforward solution to the problem of a non-negative likelihood ratio. With the latter choice, the variable to be estimated becomes $\ln(w(\underline{X}))$ and, to obtain the desired quantity, the inverse transformation has to be applied to the network output, as described in [5].

A more generic way to write (3) as a function of a monotone function of $\hat{w}(\underline{X})$, $u(\underline{X}, \underline{\theta})$, is the following:

$$J_0(\underline{\theta}) = E_0[\phi(u(\underline{X}, \underline{\theta})) \; + \; w(\underline{X}) \cdot \psi(u(\underline{X}, \underline{\theta}))] = E_0[\phi(u(\underline{X}, \underline{\theta}))] \; + \; E_1[\psi(u(\underline{X}, \underline{\theta}))] \tag{4}$$

Here, ϕ and ψ are two real functions of real variables that should satisfy some mathematical criteria if we want $u(\underline{X}, \underline{\theta})$ to minimize the loss function.

It can be demonstrated that there are no single solutions giving a closed form for ϕ and ψ. A possible choice for $u = \ln(w)$ can be $\phi(u) = \sqrt{e^u}$ and $\psi(u) = \frac{1}{\sqrt{e^u}}$.

With these substitutions, the loss function to minimize becomes:

$$J_0(\underline{\theta}) = E_0[\sqrt{e^{u(\underline{X}, \underline{\theta})}}] \; + \; E_1[\frac{1}{\sqrt{e^{u(\underline{X}, \underline{\theta})}}}] \tag{5}$$

3. Particle Physics Analyses Application

This methodology for data-driven likelihood ratio estimation has been adopted by two ATLAS analyses: one is a search for heavy resonances Y decaying into a Standard Model Higgs boson H and a new boson X in the fully-hadronic final state [6], the other is a search for resonant Higgs boson pair production in the $b\bar{b}b\bar{b}$ final state [7]. The purpose of both analyses is the estimation of the *background* distribution for the final fit discriminant variable, performed in SR.

Since the hadronic final state is common to both the analyses, it is useful here to introduce the concept of jets, which are referred to frequently in the following sections. After high-energy collisions in a particle collider, free quarks or gluons are created. Due to a particular property of QCD (the *colour confinement*), these cannot exist individually, rather, they interact with each other to form composite particles, called *hadrons*. In the *hadronization* process, all the particles created leave traces within the detector, resulting in conic agglomerates of tracks and energies deposits, called *hadronic jets*. Once data are analyzed

offline, these hadronic jets are reconstructed with a dedicated algorithm [8] with different cone opening angles, labeled by the parameter $R(R = \sqrt{(\phi_1 - \phi_2)^2 + (\eta_1 - \eta_2)^2}$, which is a measure of the angular distance between two vectors in cylindrical coordinates. The formula ϕ indicates the angle in the plane x-y in the ATLAS coordinate system, transverse to the beam axis z; η is the pseudo-rapidity, defined as $\eta = -\ln\tan(\theta/2)$, and correlated to the angle θ between the particle trajectory and the z axis. In the definition of a jet, it refers to the angular opening of the cone enclosing all the particles associated with the jet.). Jets with R=0.4 are called *small-R jets*; *large-R jets* are, alternatively, those with R = 1.0.

This section is divided into two parts, with the two applications described separately.

3.1. YXH Fully-Hadronic Analysis

The first analysis described is a search for a Beyond Standard Model (BSM) TeV-scale narrow-width boson Y decaying to a Standard Model Higgs (H) and a new O(100s) GeV-scale boson X [6]. The H is required to decay to $b\bar{b}$, its largest branching ratio decay; explicit requests are not made for X, except that its decay products are jets. The full Large Hadron Collider Run 2 $\sqrt{s} = 13$ TeV pp dataset is used, collected by the ATLAS detector from 2015 to 2018 and corresponding to an integrated luminosity of 139 fb^{-1}. A Feynman diagram for this process is shown in Figure 1.

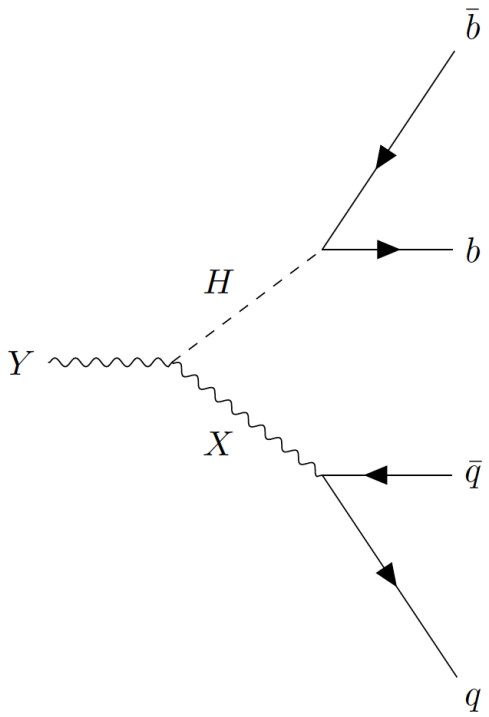


Figure 1. Feynman diagram of the target *signal* process, where the Y is produced in the initial pp collision and decays to a fully hadronic final state via an SM Higgs boson($\rightarrow b\bar{b}$) and a new particle X.

Because H and X masses are of the order of thousands of GeV, they are created with a significant momentum component in the transverse plane with respect to the beam axis, labeled p_T, that increases with the mass of the resonance. This collimates each boson's decay products, which are then reconstructed together as a single jet with a large radius, from now on, identified by J. Therefore, the signature of such a *signal* is a resonant structure in the dijet invariant mass spectrum, where events can be selected based on a final state of two large-radius jets. A supplementary resolved selection is used to reconstruct the X mass with small-R jets (in this paper labeled with j) in the case of insufficient boost from the Y.

The primary *background* comprises SM multi-jet processes, where jets are produced from quantum chromodynamic [9] (QCD) interactions, producing a smoothly falling invariant mass distribution on top of which a signal bump can be isolated.

This search is motivated by several key extensions to the Standard Model, which predict heavy diboson resonances, one of the simplest being a simplified model based on spin-1 heavy vector triplets (HVT) [10], which reproduces a large class of BSM models.

The final fit is performed on the reconstructed invariant mass of the Y in overlapping windows of the X candidate mass to further enrich the signal-to-background ratio. The results are presented as limits on the cross-section times branching ratio of the generic HVT process.

SRs are constructed by selection based on the different properties of the H and X jets. An ambiguity resolution, required to determine which of the two J in the event is more likely to be the Higgs boson, is solved using a neural-net-based classifier, which separates bosons decaying to $b\bar{b}$ from top quark and QCD jets [11]. The outputs of the NN are three classification scores corresponding to the likelihood of the jet originating from a Higgs boson (p_{Higgs}), top quark (p_{Top}), or multijet process ($p_{multijet}$), which are subsequently combined into the jet-level discriminant $D_{X_{bb}}$, as shown in Equation (6).

$$D_{X_{bb}} = \ln \frac{p_{Higgs}}{0.25 \cdot p_{Top} + (1 - 0.25) \cdot p_{multijet}} \tag{6}$$

The jet with the largest value of $D_{X_{bb}}$ is labeled as the Higgs candidate (J_H), and the other J is, by default, the X candidate (J_X), therefore, determining which jet is subject to further H and X tagging.

A novel anomaly detection *signal* region is implemented based on a jet-level score for *signal* model-independent tagging of the boosted X [12], representing the first application of fully unsupervised machine learning to an ATLAS analysis. The primary SR defined through a selection on the jet-level anomaly score (AS) of the X candidate is referred to as the *anomaly SR*. The remaining two SRs target the benchmark $X \to q\bar{q}$ decay and are, thus, referred to as *two-prong SRs*, which differ based on reconstruction of the X as either a single large-R jet (*merged SR*) or two small-R jets (*resolved SR*).

For all *signal* regions, a cut is applied to the Higgs boson candidate $D_{X_{bb}}$, along with a mass window requirement of 75 GeV $< m_H <$ 145 GeV. The further classification of events, where the Higgs boson mass candidate passes the $D_{X_{bb}}$ selection and has mass between 145 and 200 GeV defines the *high-side band* HSB1; HSB0 is in the same mass window, with the cut on $D_{X_{bb}}$ reverted. Validation is performed in the *low-side band* LSB, where the reconstructed Higgs boson mass is required to be between 65 and 75 GeV. LSB0 and LSB1 are similarly defined as having a Higgs boson candidate that fails or passes the $D_{X_{bb}}$ tagging criterion, respectively. CR0 is defined as the set of events in which the J_H is in the SR mass window but fails the $D_{X_{bb}}$ tagging. A scheme for the selection flow and the analysis region used for the *background* estimation is shown in Figure 2.

Background Estimation

The overwhelming *background* to the $Y \to XH$ signal comprises high-p_T multi-jet events. Such processes are known to produce mismodelings in Monte Carlo simulations, making simulation-based *background* estimation very challenging. Therefore, this analysis relies on a fully data-driven estimation of the *background* in the SR. The shape of the expected m_Y distribution in the SR is obtained from data in the CR0 and weights are derived that can be applied to HSB0 to reproduce the shape found in HSB1.

The baseline for this method is the verified assumption of independence of the $D_{X_{bb}}$ cut efficiency from m_H, in such a way that it is possible to define the reweighting function in a certain mass window and then apply it in another one. The reweighting function is defined as the ratio of the multi-dimensional probability distribution functions (PDFs) of the data in HSB1 to data in HSB0. In this analysis, the statistical procedure of direct importance estimation explained in Sections 2 and 2.1 is utilized, where the ratio is estimated directly from data. It is implemented via the training of a DNN, where the loss function (in Equation (5)) is minimized to produce weights that can accurately reproduce the observed ratio in data.

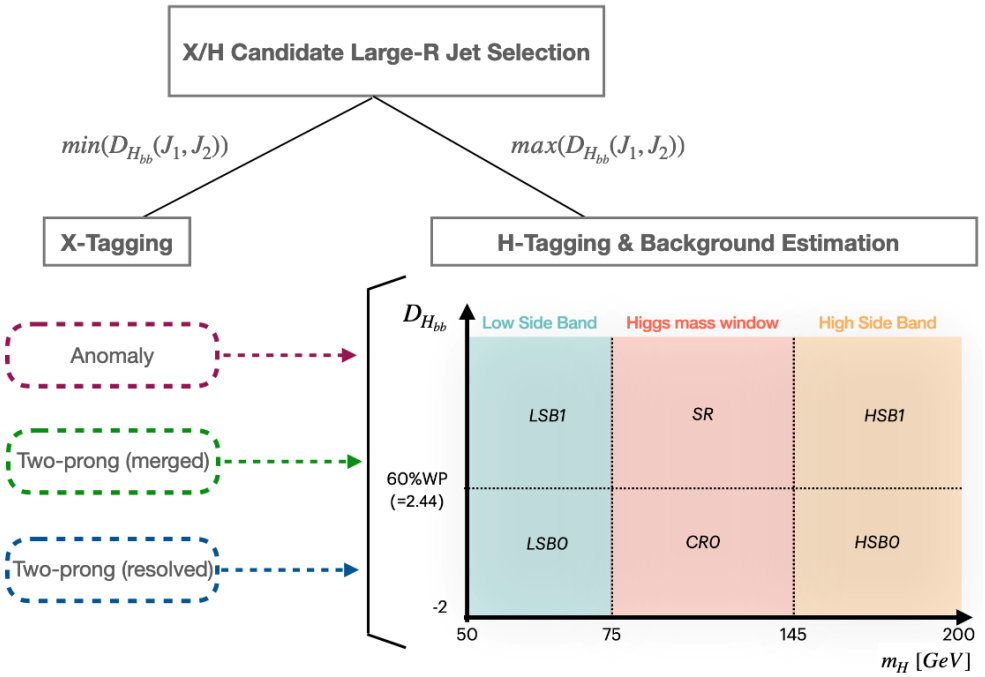


Figure 2. Selection flow after and analysis regions of the $Y \to XH$ search. The CRs (CR0, HSB0, HSB1), VRs (LSB0, LSB1), and SR are defined in the same way for all three analyses using the H candidate mass and $D_{X_{bb}}$ score [6].

The DNN is built using a fully connected sequential model from Keras with three inner layers, each with 20 neurons and a rectified linear unit (ReLU) activation function. To reduce the problem of overfitting during training, 10% of the connections among inner layers are randomly truncated ("dropout"). The last layer has a single output with a simple linear activation function. The model is trained using the Adam optimizer in Keras with Tensorflow as the backend. Training is performed using a batch size equal to the full dataset size for 1600 epochs, with early stopping if the value of the loss calculated on the validation dataset does not decrease for 100 subsequent epochs.

Events are considered for training if they pass the analysis preselection, satisfy 145 GeV $< m_H <$ 175 GeV and, additionally, have at least two track jets associated to the Higgs boson candidate. They are modeled as an unordered set of variables, namely the transverse momentum (p_T), the pseudorapidity (η), the azimuthal angle ϕ and the energy of the Higgs boson candidate (E), the number of tracks associated to the Higgs boson candidate, the transverse momentum, η, ϕ and the mass of the first two track jets associated with the Higgs boson candidate, ordered in p_T.

Each variable x is standardized with the transformation $x = \frac{(x-\mu)}{\sigma}$, where μ and σ are its mean and standard deviation, respectively.

The training was performed in the HSB, using data both in the tagged (HSB1) and untagged regions (HSB0) before applying the SRs categorization. This inclusive training enables use of a single weights set for merged, resolved and anomaly SRs. The dataset was divided into training and test sets using 70% and 30% of the full training dataset, respectively. From the training set, 20% was used for validation, to validate the model and to monitor overfitting during the training phase.

The DNN outputs event-level weights, assumed to be approximately independent of m_H, that can be applied to an untagged region to produce the m_Y shape in the corresponding X_{bb}-tagged region. These weights are validated using data from the LSB.

Figure 3 shows the impact of the reweighting on the distributions of several key analysis variables, using the two-prong merged LSB as an example region.

Three curves are shown in total, comparing the LSB0 data before and after DNN reweighting is applied to the target data distribution in LSB1. These variables are chosen to focus on kinematic variables over which the *background* estimation is extrapolated to generate the SR prediction. Shape differences are observed after the application of the weights, and good agreement is observed for the reweighted shapes to the true tagged data in all distributions, suggesting a robust *background* model. As the training is performed inclusively of the X-tagging, the same conclusion holds for the anomaly and two-prong resolved LSB regions.

This approach enables a normalized distribution to be obtained and the *background* normalization factor in SR is obtained by the fit procedure.

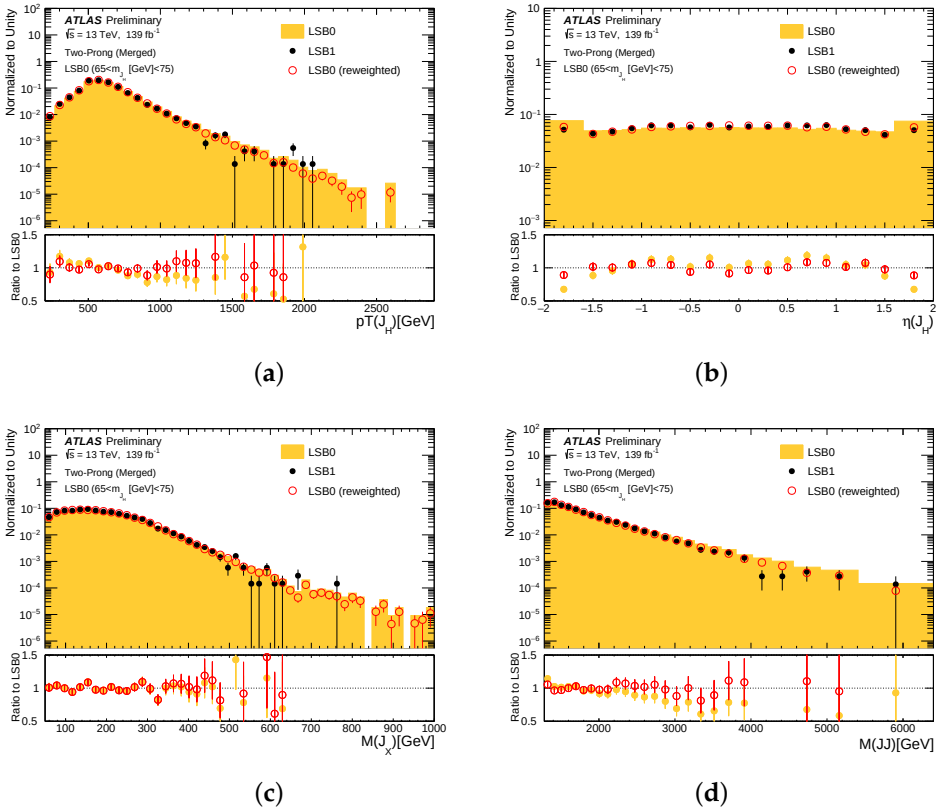


Figure 3. Distribution of the H candidate p_T (**a**) and η (**b**), along with the mass of J_X (**c**) and m_Y (**d**) in the merged LSB validation region, overlaying data from LSB1 with the data in LSB0 shown before (orange) and after (red) reweighting. The ratio of the LSB1 data to both the LSB0 data (orange) and the reweighted LSB0 data (red) is shown in the lower panel. Error bars indicate statistical uncertainties [6].

Several sources of uncertainties are related to the method, i.e., effects that are not considered in the statistical model, but can affect the result of the measurement. They are called *systematic uncertainties*; their impact is taken into account by quantifying the corresponding variation on the *background* shape. Three different kinds of such uncertainties were considered, as explained below.

The first is the potential variation in the obtained weights due to differences in phase space between HSB (where the network is trained) and the Higgs mass window. The related uncertainty is calculated by obtaining an additional *background* model, training the DNN in an alternate region of 165 GeV $< m_H <$ 200 GeV. This region has approximately the same statistics and tagging efficiency as the nominal training region, helping to isolate the effect of the particular training region on the obtained output weights. Up and down variations are defined by symmetrizing the shape difference in m_Y between the two different models, creating an effect of $\mathcal{O}(1-10)\%$ across the distribution.

Another DNN variation is built to account for the finite statistics of the training sample and the random initialization of the weights. It is estimated with a bootstrap procedure [13], where a set of 100 bootstrap networks are trained, each time varying the training dataset by resampling it with a replacement. The correct way to evaluate this systematic is to use the event-level covariance matrix between all bootstrap weights, but, since it is computationally prohibitive, the interquartile range (IQR) for each event's weight distribution is considered as a good approximation of the uncertainty, along with the IQR of the normalization factor for each bootstrap training. Two additional templates are then defined with the median weight for each event, plus or minus half of the IQR, defining the upper and lower symmetric error bands. This corresponds to a $\mathcal{O}(1)\%$ effect across m_Y.

Lastly, a non-closure uncertainty is included to cover modeling discrepancies that may arise from extrapolating weights derived from the NN training in the HSB to the LSB, and subsequently to the SR. It is defined by the symmetrized shape difference between the data and predicted *background* in the LSB. The non-closure is negligible for low m_Y and rises to $\mathcal{O}(10)\%$ in the m_Y tails.

The results of *background*-only fits of the m_Y distribution across all m_X categories in the anomaly SR show good compatibility of data to the expected *background*, after incorporating all statistical and systematic uncertainties. The largest deviation is in the m_X window [75.5, 95.5] GeV, corresponding to a global significance of 1.47σ. The results for the two-prong SRs show no significant deviations of data with respect to the predicted *background* beyond expected statistical fluctuations.

3.2. Resonant HH in the $b\bar{b}b\bar{b}$ Final State

The second ATLAS analysis described in this paper, which adopted the same *background* estimation technique, is a search for resonant (in particle physics, the term *resonant* stands for processes having a peaked invariant mass distribution around the mass of a particle which has decayed into the observed final state) Higgs boson pair production in the $b\bar{b}b\bar{b}$ final state [7]. LHC pp-collisions data collected by ATLAS in 2016–2018 are used, corresponding to 126 fb^{-1}.

To perform the search, a selection of the events used is performed, according to the expected properties of the signal, and *signal* and *control regions* are defined. The considered hadronic final state is explored by selecting only events with at least four small-R jets. Then kinematic cuts are applied to obtain SRs and CRs for the final fit on the 4b invariant mass distribution, summarized as follows: Events are first divided into two categories: *2b*, where exactly two jets are b-tagged (a b-tagged jet is a jet identified as produced by the hadronization of a b quark) and *4b*, where at least four jets are b-tagged. Exactly four jets are selected to construct the two H candidates. For 4b events, the four b-tagged jets with the highest p_T are selected. For *2b* events, the two b-tagged jets and the two untagged jets with the highest transverse momentum are selected. The *2b* events are needed to construct the *background* model for the *4b* category. This selection of untagged jets can introduce a kinematic bias with respect to the *4b* category; however, this is exactly what is accounted for by the reweighting function. After the four jets are chosen, they are paired to form the two Higgs candidates H_1 and H_2 and the pairing is chosen by a boosted decision tree (BDT). Pairs are then ordered in terms of their transverse momentum. The multi-jet *background* is reduced by requiring a certain angular separation between the two H candidates; processes coming from the top decay are suppressed by a proper applied veto. Finally, events are sorted into three kinematic regions based on the invariant masses of the H candidates: a *signal region* (SR), a *validation region* (VR) and a *control region* (CR).

Background Estimation

After the selections described above, the *background* is mainly composed of pure QCD multi-jet processes; therefore, the discussed data-driven technique fits the *background* estimation problem well. Since the *signal* contamination in the *2b* region is found to be negligible compared to the *background* uncertainties, these data are used to predict the

background shape in the *4b* SR. As in the previously discussed analysis, the HSB of the Higgs mass window is used for the training of the neural network; in this case, the CR plays the same role. Here a reweighting function is approximated by data to map the *2b* kinematic region onto the *4b* region. This function is then applied to the corresponding *2b* SR in order to have a *background* model in the *4b* SR.

The neural network used to minimize the loss in Equation (5) is composed of three densely-connected hidden layers, with 50 nodes and a ReLU activation function each, and an output layer with a single neuron with a linear activation function. The variables used in the input to the NN for training performed in CR are chosen to be sensitive to the differences between the two kinematic regions. They include the following: $\log p_T$ of both the p_T-subleading jet and the fourth highest p_T jet; $\log \Delta R$ between the first and the third p_T-ordered jets; the average $|\eta|$ of the four jets; $\log p_T$ from the di-Higgs four-momentum; ΔR between the Higgs' candidates; $\Delta\phi$ between the two jets forming each Higgs' candidate; a constructed variable taking together the difference between the reconstructed and the nominal value of the Higgs mass and the number of jets in the event.

The effect of applying this reweighting to the CR, where it is derived, is shown in Figure 4. The output of this procedure is an estimate of the corrected m_{HH} distribution in the *4b* SR, which is then used as input to the statistical procedure.

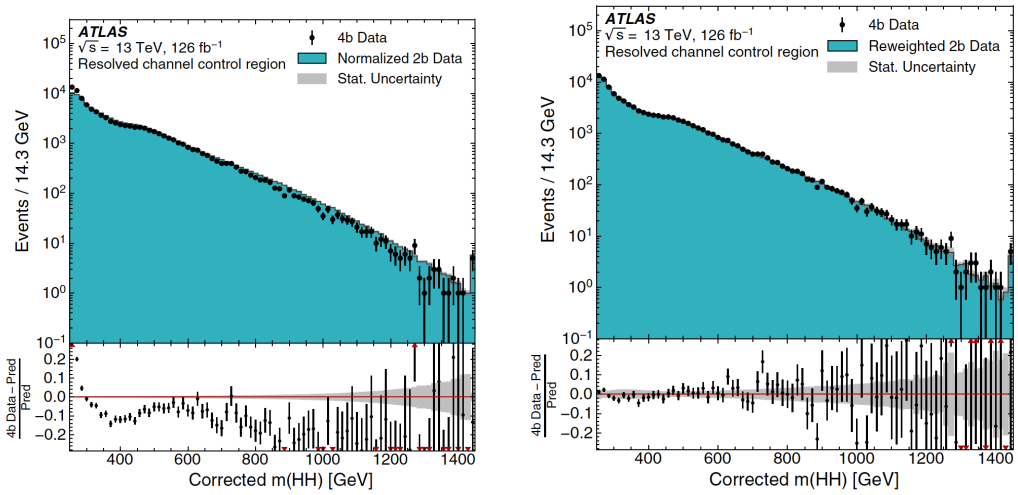
The number of events in the *4b* region is calculated similarly to the YXH analysis. In the VR, a good compatibility between target and reweighted data is found and residual differences are used for the systematic uncertainty estimation.

In the *background* estimation procedure of this analysis, two sources of systematic uncertainties have been considered:

- An uncertainty originating from the limited sample size in CR, which affects the training of the network;
- An uncertainty which takes into account the differences between SR and CR kinematics, which can have an impact in the *background* estimation, as described in Section 3.1.

The training of the NN is subject to fluctuations due to the initial conditions and the limited size of the training sample. For this reason, the bootstrap resampling technique is used [13], resulting in an ensemble of *background* estimates. In this ensemble, distributions are obtained after applying a weight to each event that is varied to the event-level weight IQR and then scaled to the same normalization as the nominal distribution; finally, they are multiplied by the ratio of the upper IQR value of the normalization factors to its nominal value. This new set of distributions creates an envelope centered around the nominal distribution, from which an estimate of the uncertainty in each $m_{(HH)}$ bin can be evaluated.

No significant excess has been observed in the data, which shows good agreement with the *background* prediction.



(**a**) (**b**)

Figure 4. A comparison between the $m_{(HH)}$ distributions in the *2b* (filled histogram) and in the *4b control regions* (dots) is shown, where the *2b* data are before and after the reweighting, respectively in (**a**,**b**). Here, the distributions are all normalized to the *4b* event yields to purely compare the shapes [7].

4. Conclusions

The computation of the likelihood ratio plays a fundamental role in particle physics' analyses. This quantity can be calculated with a data-driven approach involving machine learning by writing a custom loss function derived by setting a least-squared problem. The method itself is not limited to these kinds of problems. It can be applied to estimate likelihood ratios from data without requiring knowledge of single particle distribution functions, which is a non-trivial problem to solve, and can be applied to hypothesis-testing problems in data analysis used for other applications in various fields, such as covariate shift adaptation [14] and outlier detection [15]. The method was applied to two ATLAS analyses in relation to the *background* estimation problem, with fully satisfactory results, confirming the power of the approach.

Author Contributions: The authors equally contributed to the conceptualization, the writing, review and editing of this work. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. LHC Machine. *JINST* **2008**, *3*, S08001. [CrossRef]
2. Aad, G.; Abajyan, T.; Abbott, B.; Abdallah, J.; Khalek, S.A.; Abdelalim, A.A.; Aben, R.; Abi, B.; Abolins, M.; AbouZeid, O.; et al. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett. B* **2012**, *716*, 1–29.
3. Aad, G.; Abat, E.; Abdallah, J.; Abdelalim, A.; Abdesselam, A.; Abi, B.; Abolins, M.; Abramowicz, H.; Acerbi, E.; Acharya, B.; et al. The ATLAS experiment at the CERN large hadron collider. *J. Instrum.* **2008**, *3*, S08003. [CrossRef]
4. Kanamori, T.; Hido, S.; Sugiyama, M. A least-squares approach to direct importance estimation. *J. Mach. Learn. Res.* **2009**, *10*, 1391–1445.
5. Moustakides, G.V.; Basioti, K. Training neural networks for likelihood/density ratio estimation. *arXiv* **2019**, arXiv:1911.00405.

6. *Anomaly Detection Search for New Resonances Decaying into a Higgs Boson and a Generic New Particle X in Hadronic Final States Using $\sqrt{s}$ = 13 TeV pp Collisions with the ATLAS Detector*; Technical Report; CERN: Geneva, Switzerland, 2022. Available online: https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CONFNOTES/ATLAS-CONF-2022-045 (accessed on 6 July 2022).
7. Aad, G.; Abbott, B.; Abbott, D.; Abud, A.A.; Abeling, K.; Abhayasinghe, D.; Abidi, S.; Abramowicz, H.; Abreu, H.; Abulaiti, Y.; et al. Search for resonant pair production of Higgs bosons in the b $\bar{b}$ b $\bar{b}$ final state using p p collisions at s = 13 TeV with the ATLAS detector. *Phys. Rev. D* **2022**, *105*, 092002. [CrossRef]
8. Cacciari, M.; Salam, G.P.; Soyez, G. The anti-kt jet clustering algorithm. *J. High Energy Phys.* **2008**, *2008*, 063. [CrossRef]
9. Marciano, W.; Pagels, H. Quantum chromodynamics. *Phys. Rep.* **1978**, *36*, 137–276. [CrossRef]
10. Pappadopulo, D.; Thamm, A.; Torre, R.; Wulzer, A. Heavy vector triplets: bridging theory and data. *J. High Energy Phys.* **2014**, *2014*, 1–50 [CrossRef]
11. The ATLAS Collaboration. Identification of Boosted Higgs Bosons Decaying Into bb With Neural Networks and Variable Radius Subjets in ATLAS. *arXiv*, **2020**, arXiv:1906.11005.
12. Kahn, A.; Gonski, J.; Ochoa, I.; Williams, D.; Brooijmans, G. Anomalous jet identification via sequence modeling. *J. Instrum.* **2021**, *16*, P08012. [CrossRef]
13. Michelucci, U.; Venturini, F. Estimating neural network's performance with bootstrap: A tutorial. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 357–373. [CrossRef]
14. Sugiyama, M.; Suzuki, T.; Nakajima, S.; Kashima, H.; von Bünau, P.; Kawanabe, M. Direct importance estimation for covariate shift adaptation. *Ann. Inst. Stat. Math.* **2008**, *60*, 699–746. [CrossRef]
15. Ben-Gal, I. Outlier detection. In *Data Mining and Knowledge Discovery Handbook*; Springer: Boston, MA, USA, 2005; pp. 131–146.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.